Nonparametric regression estimation

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

August 27, 2012

e-mail: gyorfi@cs.bme.hu
www.cs.bme.hu/˜gyorfi

Regression function

Y real valued
X observation vector

Regression function

Y real valued
X observation vector
Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

Y real valued
X observation vector
Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

Regression function

$$m(x) = \mathbf{E}\{Y \mid X = x\}$$

Regression function

For each function f, one has

$$\mathbf{E}\{(f(X)-Y)^2\} = \mathbf{E}\{\mathbf{E}\{(f(X)-Y)^2 \mid X\}\}$$

and

$$\mathbf{E}\{(f(X)-Y)^2 \mid X\} \quad = \quad \mathbf{E}\{(f(X)-m(X)+m(X)-Y)^2 \mid X\}$$

Regression function

For each function f, one has

$$\mathbf{E}\{(f(X)-Y)^2\} = \mathbf{E}\{\mathbf{E}\{(f(X)-Y)^2 \mid X\}\}$$

and

$$\begin{aligned}
\mathbf{E}\{(f(X)-Y)^2 \mid X\} &= \mathbf{E}\{(f(X)-m(X)+m(X)-Y)^2 \mid X\} \\
&= \mathbf{E}\{(f(X)-m(X))^2 \mid X\} \\
&\quad +2\mathbf{E}\{(f(X)-m(X))(m(X)-Y) \mid X\} \\
&\quad +\mathbf{E}\{(m(X)-Y)^2 \mid X\}
\end{aligned}$$

Regression function

For each function f, one has

$$\mathbf{E}\{(f(X)-Y)^2\} = \mathbf{E}\{\mathbf{E}\{(f(X)-Y)^2 \mid X\}\}$$

and

$$\begin{aligned}
\mathbf{E}\{(f(X)-Y)^2 \mid X\} &= \mathbf{E}\{(f(X)-m(X)+m(X)-Y)^2 \mid X\} \\
&= \mathbf{E}\{(f(X)-m(X))^2 \mid X\} \\
&\quad +2\mathbf{E}\{(f(X)-m(X))(m(X)-Y) \mid X\} \\
&\quad +\mathbf{E}\{(m(X)-Y)^2 \mid X\} \\
&= (f(X)-m(X))^2 \\
&\quad +2(f(X)-m(X))\mathbf{E}\{m(X)-Y|X\} \\
&\quad +\mathbf{E}\{(m(X)-Y)^2 \mid X\}
\end{aligned}$$

Regression function

For each function f, one has

$$\mathbf{E}\{(f(X)-Y)^2\} = \mathbf{E}\{\mathbf{E}\{(f(X)-Y)^2 \mid X\}\}$$

and

$$\begin{aligned}
\mathbf{E}\{(f(X)-Y)^2 \mid X\} &= \mathbf{E}\{(f(X)-m(X)+m(X)-Y)^2 \mid X\} \\
&= \mathbf{E}\{(f(X)-m(X))^2 \mid X\} \\
&\quad +2\mathbf{E}\{(f(X)-m(X))(m(X)-Y) \mid X\} \\
&\quad +\mathbf{E}\{(m(X)-Y)^2 \mid X\} \\
&= (f(X)-m(X))^2 \\
&\quad +2(f(X)-m(X))\mathbf{E}\{m(X)-Y|X\} \\
&\quad +\mathbf{E}\{(m(X)-Y)^2 \mid X\} \\
&= (f(X)-m(X))^2 + \mathbf{E}\{(m(X)-Y)^2 \mid X\}
\end{aligned}$$

Regression function estimate

Data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$

Regression function estimate

Data: $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$
Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Data: $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$
Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Special case:

$$\mathbf{P}\{X = x\} > 0$$

Regression function estimate

Data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Special case:

$$\mathbf{P}\{X = x\} > 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \frac{\mathbf{E}\{YI_{\{X=x\}}\}}{\mathbf{P}\{X = x\}}$$

Regression function estimate

Data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Special case:

$$\mathbf{P}\{X = x\} > 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \frac{\mathbf{E}\{YI_{\{X=x\}}\}}{\mathbf{P}\{X = x\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\{X_i=x\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{X_i=x\}}}$$

Regression function estimate

Data: $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$
Regression function estimate

$$m_n(x) = m_n(x, D_n)$$

Special case:

$$\mathbf{P}\{X = x\} > 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \frac{\mathbf{E}\{YI_{\{X=x\}}\}}{\mathbf{P}\{X = x\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i I_{\{X_i=x\}}}{\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i=x\}}} \to \frac{\mathbf{E}\{YI_{\{X=x\}}\}}{\mathbf{P}\{X = x\}} = m(x)$$

a.s.

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{\mathbf{E}\{Y I_{\{\|X-x\| \le h\}}\}}{\mathbf{P}\{\|X - x\| \le h\}}$$

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{\mathbf{E}\{Y I_{\{\|X - x\| \leq h\}}\}}{\mathbf{P}\{\|X - x\| \leq h\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\{\|X_i - x\| \leq h\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}$$

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{\mathbf{E}\{Y I_{\{\|X-x\| \le h\}}\}}{\mathbf{P}\{\|X - x\| \le h\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\{\|X_i - x\| \le h\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}} = \frac{\sum_{i=1}^n Y_i I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Bandwidth $h = h_n$

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{\mathbf{E}\{Y I_{\{\|X-x\| \le h\}}\}}{\mathbf{P}\{\|X - x\| \le h\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i I_{\{\|X_i - x\| \le h\}}}{\frac{1}{n}\sum_{i=1}^{n} I_{\{\|X_i - x\| \le h\}}} = \frac{\sum_{i=1}^{n} Y_i I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^{n} I_{\{\|X_i - x\| \le h\}}}$$

Bandwidth $h = h_n$
h_n should be "small"

Regression function estimate

General case:

$$\mathbf{P}\{X = x\} = 0$$

Then

$$m(x) = \mathbf{E}\{Y \mid X = x\} \stackrel{\text{def}}{=} \lim_{h \to 0} \frac{\mathbf{E}\{Y I_{\{\|X-x\| \le h\}}\}}{\mathbf{P}\{\|X - x\| \le h\}}$$

Local averaging estimate

$$m_n(x) = \frac{\frac{1}{n}\sum_{i=1}^n Y_i I_{\{\|X_i - x\| \le h\}}}{\frac{1}{n}\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}} = \frac{\sum_{i=1}^n Y_i I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Bandwidth $h = h_n$
h_n should be "small"
nh_n^d should be "large"

Error decomposition

$$\bar{m}_n(x) = \frac{\sum_{i=1}^{n} m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^{n} I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$m_n(x) - m(x)$$

Error decomposition

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$m_n(x) - m(x) = m_n(x) - \bar{m}_n(x) + \bar{m}_n(x) - m(x)$$

Error decomposition

$$\bar{m}_n(x) = \frac{\sum_{i=1}^{n} m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^{n} I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$\begin{aligned} m_n(x) - m(x) &= m_n(x) - \bar{m}_n(x) + \bar{m}_n(x) - m(x) \\ &= \text{variation} + \text{bias} \end{aligned}$$

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$\begin{aligned} m_n(x) - m(x) &= m_n(x) - \bar{m}_n(x) + \bar{m}_n(x) - m(x) \\ &= \text{variation} + \text{bias} \end{aligned}$$

Variation

$$m_n(x) - \bar{m}_n(x) = \frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n m(X_i) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Error decomposition

$$\begin{aligned} m_n(x) - m(x) &= m_n(x) - \bar{m}_n(x) + \bar{m}_n(x) - m(x) \\ &= \text{variation} + \text{bias} \end{aligned}$$

Variation

$$m_n(x) - \bar{m}_n(x) = \frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Bias

$$\bar{m}_n(x) - m(x) \approx \frac{\sum_{i=1}^n (m(X_i) - m(x)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}$$

Error analysis: variation

$$\mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}\right)^2\right\}$$

Error analysis: variation

$$
\begin{aligned}
& \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}\right)^2\right\} \\
= \quad & \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\}
\end{aligned}
$$

Error analysis: variation

$$
\begin{aligned}
& \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}\right)^2\right\} \\
= \quad & \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\} \\
& + \mathbf{E}\left\{\frac{\sum_{i \ne j} (Y_i - m(X_i))(Y_j - m(X_j)) I_{\{\|X_i - x\| \le h\}} I_{\{\|X_j - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\}
\end{aligned}
$$

Error analysis: variation

$$\mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}\right)^2\right\}$$

$$= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\}$$

$$+\mathbf{E}\left\{\frac{\sum_{i \ne j} (Y_i - m(X_i))(Y_j - m(X_j)) I_{\{\|X_i - x\| \le h\}} I_{\{\|X_j - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\}$$

For $i \ne j$,

$$\mathbf{E}\left\{(Y_i - m(X_i))(Y_j - m(X_j)) I_{\{\|X_i - x\| \le h, \|X_j - x\| \le h\}} \mid X_1, \ldots, X_n, Y_i\right\}$$

$$= (Y_i - m(X_i)) \mathbf{E}\left\{Y_j - m(X_j) \mid X_1, \ldots, X_n, Y_i\right\} I_{\{\|X_i - x\| \le h, \|X_j - x\| \le h\}}$$

$$= 0$$

$$\mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i-x\|\le h\}}}{\sum_{i=1}^n I_{\{\|X_i-x\|\le h\}}}\right)^2\right\}$$

$$= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i-x\|\le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\le h\}}\right)^2}\right\}$$

$$+\mathbf{E}\left\{\frac{\sum_{i\neq j} (Y_i - m(X_i))(Y_j - m(X_j)) I_{\{\|X_i-x\|\le h\}} I_{\{\|X_j-x\|\le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\le h\}}\right)^2}\right\}$$

For $i \neq j$,

$$\mathbf{E}\left\{(Y_i - m(X_i))(Y_j - m(X_j)) I_{\{\|X_i-x\|\le h, \|X_j-x\|\le h\}} \mid X_1, \dots, X_n, Y_i\right\}$$

$$= (Y_i - m(X_i))\mathbf{E}\left\{Y_j - m(X_j) \mid X_1, \dots, X_n, Y_i\right\} I_{\{\|X_i-x\|\le h, \|X_j-x\|\le h\}}$$

$$= 0$$

Error analysis: variation

Assume that $|Y| \leq L$

Assume that $|Y| \leq L$

$$\mathbf{E}\{(m_n(x) - \bar{m}_n(x))^2\}$$

Error analysis: variation

Assume that $|Y| \leq L$

$$\mathbf{E}\{(m_n(x) - \bar{m}_n(x))^2\} \quad = \quad \mathbf{E}\left\{ \left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}} \right)^2 \right\}$$

Error analysis: variation

Assume that $|Y| \le L$

$$\mathbf{E}\{(m_n(x) - \bar{m}_n(x))^2\} = \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \le h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}}\right)^2\right\}$$

$$= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \le h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \le h\}}\right)^2}\right\}$$

Error analysis: variation

Assume that $|Y| \leq L$

$$\begin{aligned}
\mathbf{E}\{(m_n(x) - \bar{m}_n(x))^2\} &= \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}\right)^2\right\} \\
&= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}\right)^2}\right\} \\
&\leq \mathbf{E}\left\{\frac{\sum_{i=1}^n 4L^2 I_{\{\|X_i - x\| \leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}\right)^2}\right\}
\end{aligned}$$

Error analysis: variation

Assume that $|Y| \leq L$

$$
\begin{aligned}
\mathbf{E}\{(m_n(x)-\bar{m}_n(x))^2\} &= \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i-x\|\leq h\}}}{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}}\right)^2\right\} \\
&= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i-x\|\leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}\right)^2}\right\} \\
&\leq \mathbf{E}\left\{\frac{\sum_{i=1}^n 4L^2 I_{\{\|X_i-x\|\leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}\right)^2}\right\} \\
&= \mathbf{E}\left\{\frac{4L^2}{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}} I_{\{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}>0\}}\right\}
\end{aligned}
$$

Error analysis: variation

Assume that $|Y| \leq L$

$$
\begin{aligned}
\mathbf{E}\{(m_n(x)-\bar{m}_n(x))^2\} &= \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i-x\|\leq h\}}}{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}}\right)^2\right\} \\
&= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i-x\|\leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}\right)^2}\right\} \\
&\leq \mathbf{E}\left\{\frac{\sum_{i=1}^n 4L^2 I_{\{\|X_i-x\|\leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}\right)^2}\right\} \\
&= \mathbf{E}\left\{\frac{4L^2}{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}} I_{\{\sum_{i=1}^n I_{\{\|X_i-x\|\leq h\}}>0\}}\right\} \\
&\leq \frac{c}{nh_n^d}
\end{aligned}
$$

Error analysis: variation

Assume that $|Y| \leq L$

$$
\begin{aligned}
\mathbf{E}\{(m_n(x) - \bar{m}_n(x))^2\} &= \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (Y_i - m(X_i)) I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}\right)^2\right\} \\
&= \mathbf{E}\left\{\frac{\sum_{i=1}^n (Y_i - m(X_i))^2 I_{\{\|X_i - x\| \leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}\right)^2}\right\} \\
&\leq \mathbf{E}\left\{\frac{\sum_{i=1}^n 4L^2 I_{\{\|X_i - x\| \leq h\}}}{\left(\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}\right)^2}\right\} \\
&= \mathbf{E}\left\{\frac{4L^2}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}} I_{\{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}} > 0\}}\right\} \\
&\leq \frac{c}{nh_n^d} \\
&\to 0
\end{aligned}
$$

if $nh_n^d \to \infty$.

Error analysis: bias

Assume the Lipschitz condition:

$$|m(x) - m(z)| \leq C\|x - z\|$$

Error analysis: bias

Assume the Lipschitz condition:

$$|m(x) - m(z)| \leq C\|x - z\|$$

$\mathbf{E}\{(\bar{m}_n(x) - m(x))^2\}$

Error analysis: bias

Assume the Lipschitz condition:

$$|m(x) - m(z)| \leq C\|x - z\|$$

$$\mathbf{E}\{(\bar{m}_n(x) - m(x))^2\} \quad \approx \quad \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (m(X_i) - m(x)) I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}\right)^2\right\}$$

Error analysis: bias

Assume the Lipschitz condition:

$$|m(x) - m(z)| \leq C\|x - z\|$$

$$\begin{aligned}
\mathbf{E}\{(\bar{m}_n(x) - m(x))^2\} &\approx \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n (m(X_i) - m(x))I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}\right)^2\right\} \\
&\leq \mathbf{E}\left\{\left(\frac{\sum_{i=1}^n C\|X_i - x\| I_{\{\|X_i - x\| \leq h\}}}{\sum_{i=1}^n I_{\{\|X_i - x\| \leq h\}}}\right)^2\right\} \\
&\leq C^2 h^2 \\
&\to 0
\end{aligned}$$

if $h = h_n \to 0$.

Regression function estimate

Usual consistency conditions:
- $m(x)$ is smooth
- X has a density
- Y is bounded

Regression function estimate

Usual consistency conditions:

- $m(x)$ is smooth
- X has a density
- Y is bounded

Nonparametric features:

- construction of the estimate
- consistency

Universal consistency

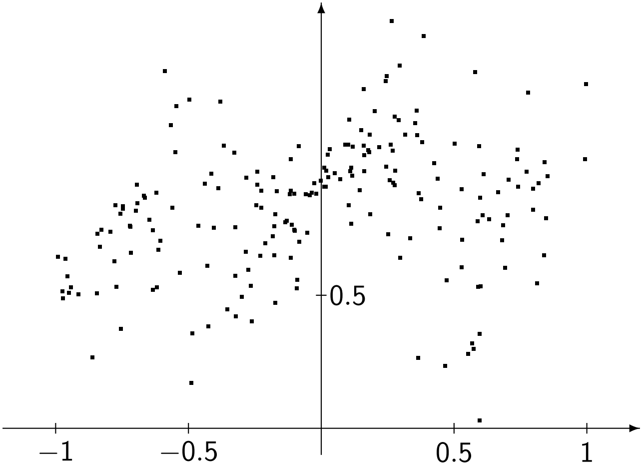
Definition

The estimator m_n is called **weakly universally consistent** if

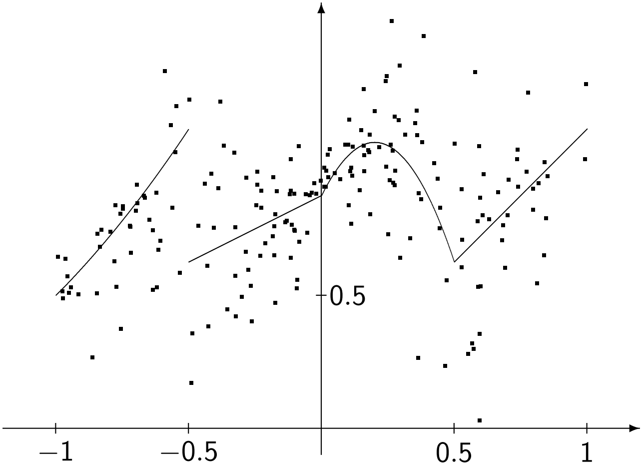
$$\mathbf{E}\left\{(m(X) - m_n(X))^2\right\} \to 0$$

for all distributions of (X, Y) with $\mathbf{E}Y^2 < \infty$.

Simulated data points.



Data points and regression function.



Local averaging estimates

Stone (1977)

$$m_n(x) = \sum_{i=1}^{n} W_{ni}(x; X_1, \dots, X_n) Y_i .$$

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \dots, X_n$, and W_{ni} is 0 otherwise.

k-nearest neighbor estimate

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and W_{ni} is 0 otherwise. Formally

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

k-nearest neighbor estimate

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \dots, X_n$, and W_{ni} is 0 otherwise. Formally

$$(X_1, Y_1), \dots, (X_n, Y_n)$$

nearest neighbor permutation: fix x

$$(X_1^{(x)}, Y_1^{(x)}), \dots, (X_n^{(x)}, Y_n^{(x)})$$

k-nearest neighbor estimate

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and W_{ni} is 0 otherwise. Formally

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

nearest neighbor permutation: fix x

$$(X_1^{(x)}, Y_1^{(x)}), \ldots, (X_n^{(x)}, Y_n^{(x)})$$

such that $\|X_1^{(x)} - x\| \leq \|X_2^{(x)} - x\| \leq \cdots \leq \|X_n^{(x)} - x\|$

k-nearest neighbor estimate

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and W_{ni} is 0 otherwise. Formally

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

nearest neighbor permutation: fix x

$$(X_1^{(x)}, Y_1^{(x)}), \ldots, (X_n^{(x)}, Y_n^{(x)})$$

such that $\|X_1^{(x)} - x\| \leq \|X_2^{(x)} - x\| \leq \cdots \leq \|X_n^{(x)} - x\|$
Nearest neighbor estimate

$$m_n(x) = \frac{1}{k} \sum_{j=1}^{k} Y_j^{(x)}$$

W_{ni} is $1/k$ if X_i is one of the k nearest neighbors of x among $X_1, \ldots, X_n$, and W_{ni} is 0 otherwise. Formally

$$(X_1, Y_1), \ldots, (X_n, Y_n)$$

nearest neighbor permutation: fix x

$$(X_1^{(x)}, Y_1^{(x)}), \ldots, (X_n^{(x)}, Y_n^{(x)})$$

such that $\|X_1^{(x)} - x\| \leq \|X_2^{(x)} - x\| \leq \cdots \leq \|X_n^{(x)} - x\|$
Nearest neighbor estimate

$$m_n(x) = \frac{1}{k} \sum_{j=1}^{k} Y_j^{(x)}$$

Theorem

If $k_n \to \infty, \quad k_n/n \to 0$ then the k-nearest neighbor estimate is weakly universally consistent.

Nearest neighbor estimate

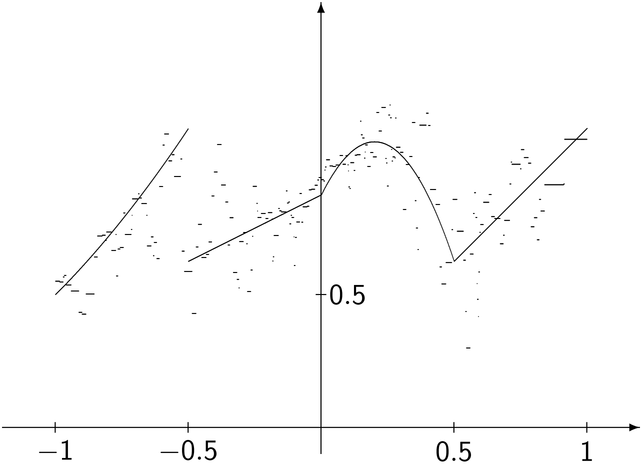


Figure: Undersmoothing: $k_n = 3$, L_2 error =0.011703.

Nearest neighbor estimate

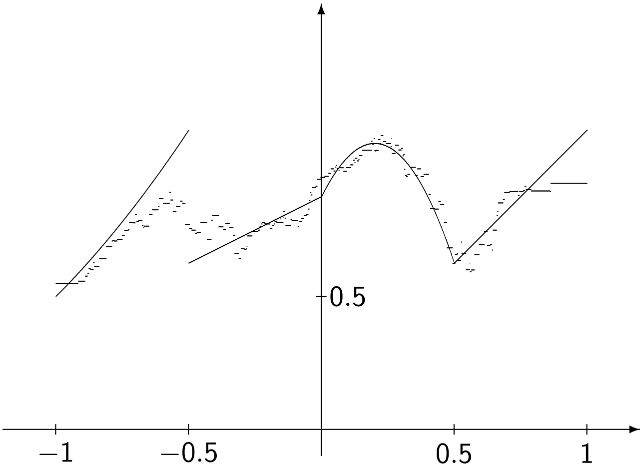


Figure: Good choice: $k_n = 12$, L_2 error $=0.004247$.

Nearest neighbor estimate

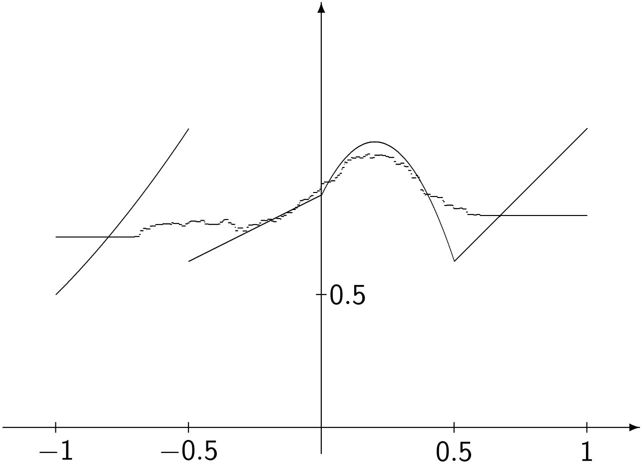


Figure: Oversmoothing: $k_n = 50$, L_2 error =0.009931.

Partitioning estimate

Partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \ldots\}$

Partitioning estimate

Partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$
$A_n(x) = A_{n,j}$ if $x \in A_{n,j}$

Partitioning estimate

Partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$
$A_n(x) = A_{n,j}$ if $x \in A_{n,j}$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}}$$

Partitioning estimate

Partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$
$A_n(x) = A_{n,j}$ if $x \in A_{n,j}$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}}$$

Example: $A_{n,j}$ are cubes with volume h_n^d

Partitioning estimate

Partition $\mathcal{P}_n = \{A_{n,1}, A_{n,2} \dots\}$
$A_n(x) = A_{n,j}$ if $x \in A_{n,j}$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i I_{\{X_i \in A_n(x)\}}}{\sum_{i=1}^n I_{\{X_i \in A_n(x)\}}}$$

Example: $A_{n,j}$ are cubes with volume h_n^d

Theorem

For cubic partition, if

$$\lim_{n\to\infty} h_n = 0$$

and

$$\lim_{n\to\infty} nh_n^d \to \infty$$

then the partitioning estimate is weakly universally consistent.

Partitioning estimate

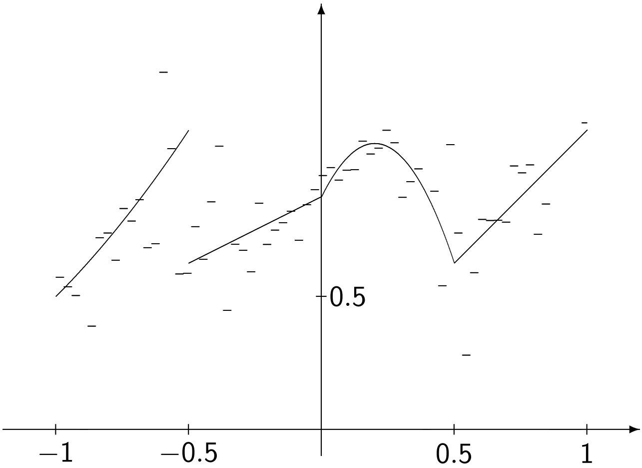


Figure: Undersmoothing: $h = 0.03$, L_2 error $= 0.062433$.

Partitioning estimate

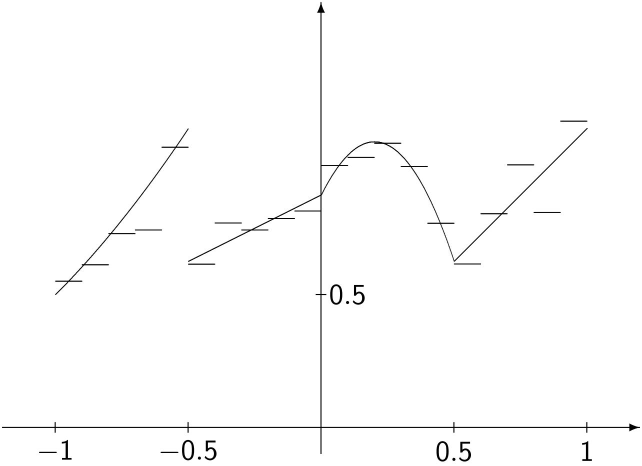


Figure: Good choice: $h = 0.1$, L_2 error $= 0.003642$.

Partitioning estimate

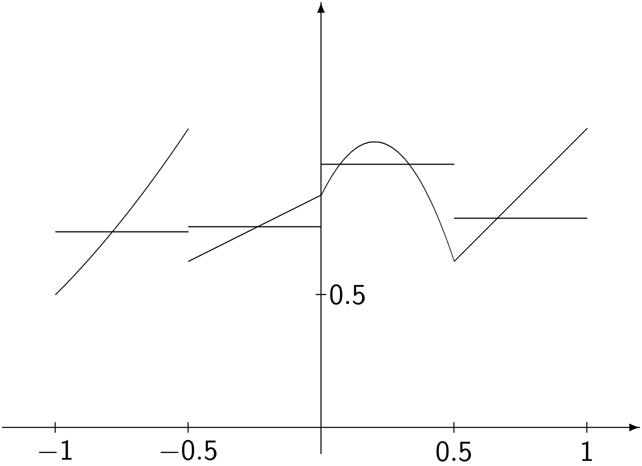


Figure: Oversmoothing: $h = 0.5$, L_2 error $= 0.013208$.

Kernel estimate

Kernel function $K(x) \geq 0$
Bandwidth $h_n > 0$

Kernel estimate

Kernel function $K(x) \geq 0$
Bandwidth $h_n > 0$

$$m_n(x) = \frac{\sum_{i=1}^n Y_i K\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)}$$

Kernel estimate

Kernel function $K(x) \geq 0$
Bandwidth $h_n > 0$

$$m_n(x) = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{x-X_i}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)}$$

Theorem

If $h_n \to 0$, $nh_n^d \to \infty$ then under some conditions on K the kernel estimate is weakly universally consistent.

Kernel estimate

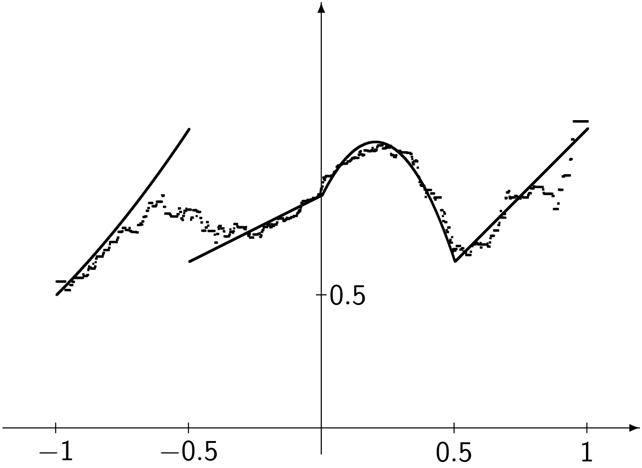


Figure: Kernel estimate for the naive kernel: $h = 0.1$, L_2 error $=$ 0.004066.

Least squares estimates

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

Least squares estimates

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

empirical L_2 error

$$\frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2$$

Least squares estimates

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

empirical L_2 error

$$\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_j|^2$$

class of functions $\mathcal{F}_n$

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

empirical L_2 error

$$\frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2$$

class of functions $\mathcal{F}_n$
select a function from $\mathcal{F}_n$ which minimizes the empirical error:

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

empirical L_2 error

$$\frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2$$

class of functions $\mathcal{F}_n$
select a function from $\mathcal{F}_n$ which minimizes the empirical error:
$m_n \in \mathcal{F}_n$ and

$$\frac{1}{n}\sum_{j=1}^{n} |m_n(X_j) - Y_j|^2 = \min_{f \in \mathcal{F}_n} \frac{1}{n}\sum_{j=1}^{n} |f(X_j) - Y_j|^2.$$

Regression problem

$$\min_f \mathbf{E}\{(Y - f(X))^2\}$$

empirical L_2 error

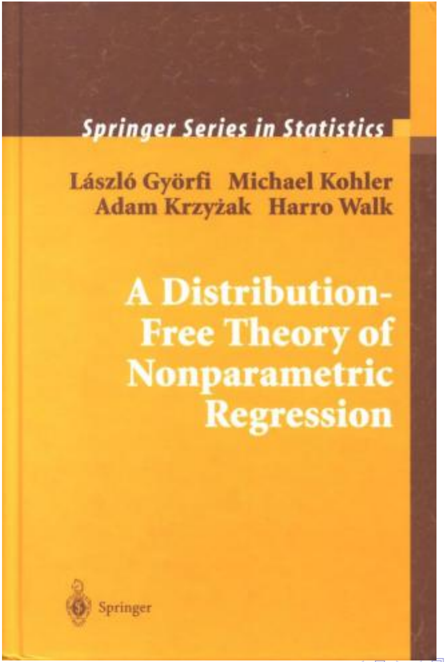
$$\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_j|^2$$

class of functions $\mathcal{F}_n$
select a function from $\mathcal{F}_n$ which minimizes the empirical error:
$m_n \in \mathcal{F}_n$ and

$$\frac{1}{n}\sum_{j=1}^{n}|m_n(X_j) - Y_j|^2 = \min_{f\in\mathcal{F}_n}\frac{1}{n}\sum_{j=1}^{n}|f(X_j) - Y_j|^2.$$

the class $\mathcal{F}_n$ grows slowly as n grows

Examples for $\mathcal{F}_n$:
- polynomials
- splines
- neural networks
- radial basis functions


Springer Series in Statistics
László Györfi Michael Kohler
Adam Krzyżak Harro Walk
A Distribution-Free Theory of Nonparametric Regression
Springer

Prediction of time series: squared loss

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

August 27, 2012

e-mail: gyorfi@cs.bme.hu
www.cs.bme.hu/∼gyorfi

Prediction for squared loss

Y_i real valued
X_i vector valued

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$
and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$
and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \ldots, X_i) = X_1^i$
and $(Y_1, \ldots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$
$g_i(X_1^i, Y_1^{i-1})$ is the estimate of Y_i

Prediction for squared loss

Y_i real valued
X_i vector valued
At time instant i the predictor is asked to guess Y_i
with knowledge of the past $(X_1, \dots, X_i) = X_1^i$
and $(Y_1, \dots Y_{i-1}) = Y_1^{i-1}$
The predictor is a sequence of functions $g = \{g_i\}_{i=1}^{\infty}$
$g_i(X_1^i, Y_1^{i-1})$ is the estimate of Y_i
After n time instant the empirical squared error

$$L_n(g) = \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2.$$

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process

Dependent data: time series

The data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

The data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

The fundamental limit: for any predictor g

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{almost surely,}$$

The data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ are **dependent**
long-range dependent
form a stationary and ergodic process
The best predictor is the conditional expectation

$$g_i^*(X_1^i, Y_1^{i-1}) = \mathbf{E}\{Y_i \mid X_1^i, Y_1^{i-1}\}.$$

The fundamental limit: for any predictor g

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{almost surely,}$$

where

$$L^* = \mathbf{E}\{(Y_0 - \mathbf{E}\{Y_0 \big| X_{-\infty}^0, Y_{-\infty}^{-1}\})^2\}$$

is the minimal mean squared error of any prediction for the value of Y_0 based on the infinite past $X_{-\infty}^0, Y_{-\infty}^{-1}$.

Universal consistency

there are universally consistent prediction sequence g_n:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for a class of stationary and ergodic sequences.

there are universally consistent prediction sequence g_n:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for a class of stationary and ergodic sequences.
Such prediction sequence is called **universally consistent**.

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell \to \infty} r'_{k,\ell} = 0.$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k,\ell = 1,2,\dots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell\to\infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell\to\infty} r'_{k,\ell} = 0.$$

location of the matches

$$J_n^{(k,\ell)} = \Big\{ k < i < n : \|x_{i-k}^{i} - x_{n-k}^{n}\| \le r_{k,\ell},$$

Elementary experts via local averaging

window kernel based elementary predictor (expert): $h^{(k,\ell)}$, $k, \ell = 1, 2, \ldots$.
two radii $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$
for any fixed k

$$\lim_{\ell \to \infty} r_{k,\ell} = 0,$$

and

$$\lim_{\ell \to \infty} r'_{k,\ell} = 0.$$

location of the matches

$$J_n^{(k,\ell)} = \left\{ k < i < n : \|x_{i-k}^i - x_{n-k}^n\| \le r_{k,\ell}, \ \|y_{i-k}^{i-1} - y_{n-k}^{n-1}\| \le r'_{k,\ell} \right\}$$

Then the local averaging prediction of the expert (k,ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i\in J_n^{(k,\ell)}} y_i}{\left|J_n^{(k,\ell)}\right|}.$$

Then the local averaging prediction of the expert (k,ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

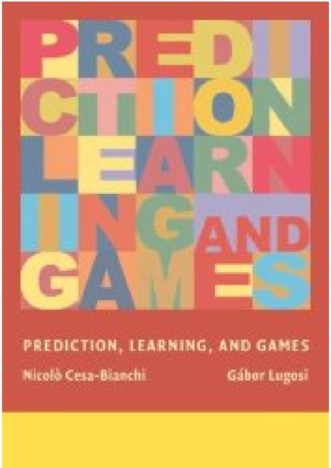
These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.
for large radius, the bias is large and, for small radius, the variance is large because of the few matchings.

Then the local averaging prediction of the expert (k, ℓ) is the average of y_i's if $i \in J_n^{(k,\ell)}$:

$$h_n^{(k,\ell)}(x_1^n, y_1^{n-1}) = \frac{\sum_{i \in J_n^{(k,\ell)}} y_i}{|J_n^{(k,\ell)}|}.$$

These predictors are not universally consistent
for small k, the bias is large and, for large k, the variance is large because of the few matchings.
for large radius, the bias is large and, for small radius, the variance is large because of the few matchings.

The problem is how to choose k, $r_{k,\ell} > 0$ and $r'_{k,\ell} > 0$ in a data dependent way.


PREDICTION LEARNING AND GAMES
PREDICTION, LEARNING, AND GAMES
Nicolò Cesa-Bianchi
Gábor Lugosi

Machine learning: exponential weighting

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k, ℓ),

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k,ℓ), and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)}$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k,ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \le B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k,ℓ), and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

and

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} .$$

Combination of experts: bounded Y

Machine learning: exponential weighting
$|Y| \leq B$, and B is known.
Let $\{q_{k,\ell}\}$ be a probability distribution over (k,ℓ),
and put

$$w_{t,k,\ell} = q_{k,\ell} e^{-(t-1)L_{t-1}(h^{(k,\ell)})/(8B^2)} = q_{k,\ell} e^{-\sum_{i=1}^{t-1}(h_i^{(k,\ell)}(X_1^i, Y_1^{i-1}) - Y_i)^2/(8B^2)}$$

and

$$p_{t,k,\ell} = \frac{w_{t,k,\ell}}{\sum_{i,j=1}^{\infty} w_{t,i,j}} .$$

Then the combined prediction

$$g_t(x_1^t, y_1^{t-1}) = \sum_{k,\ell=1}^{\infty} p_{t,k,\ell} h^{(k,\ell)}(x_1^t, y_1^{t-1}) .$$

Consistency theorem for bounded Y

Theorem. (Györfi, Lugosi (2001)) If $|Y_0| \leq B$, then the combined predictor g is universally consistent.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \ldots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_n(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$.

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_n(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$,

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_n(x_1^n, y_1^{n-1})| \le B$ and $|y_n| \le B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \ge 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_n(x_1^n, y_1^{n-1})| \le B$ and $|y_n| \le B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \ge 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(x_1^t, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(x_1^t, y_1^{t-1})$$

Expert Lemma

Let $\tilde{h}_1, \tilde{h}_2, \dots$ be a sequence of prediction strategies (experts), and let $\{q_k\}$ be a probability distribution on the set of positive integers. Assume that $|\tilde{h}_n(x_1^n, y_1^{n-1})| \leq B$ and $|y_n| \leq B$. Define

$$w_{t,k} = q_k e^{-(t-1)L_{t-1}(\tilde{h}_k)/c}$$

with $c \geq 8B^2$, and

$$v_{t,k} = \frac{w_{t,k}}{\sum_{i=1}^{\infty} w_{t,i}}.$$

If the prediction strategy $\tilde{g}$ is defined by

$$\tilde{g}_t(x_1^t, y_1^{t-1}) = \sum_{k=1}^{\infty} v_{t,k} \tilde{h}_k(x_1^t, y_1^{t-1})$$

then for every $n \geq 1$,

$$L_n(\tilde{g}) \leq \inf_k \left(L_n(\tilde{h}_k) - \frac{c \ln q_k}{n} \right).$$

Proof

Introduce

$$W_1 = 1$$

and

$$W_t = \sum_{k=1}^{\infty} w_{t,k}$$

for $t > 1$.

Proof

Introduce

$$W_1 = 1$$

and

$$W_t = \sum_{k=1}^{\infty} w_{t,k}$$

for $t > 1$. Note that

$$W_{t+1} = \sum_{k=1}^{\infty} w_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)^2/c} = W_t \sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)^2/c},$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)^2/c} \right).$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)^2 / c} \right).$$

Introduce the function

$$F_t(z) = e^{-(y_t - z)^2 / c}$$

so that

$$-c \ln \frac{W_{t+1}}{W_t} = -c \ln \left(\sum_{k=1}^{\infty} v_{t,k} e^{-\left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)^2 / c} \right).$$

Introduce the function

$$F_t(z) = e^{-(y_t - z)^2/c}$$

Because of $c \geq 8B^2$, the function F_t is concave on $[-B, B]$, therefore Jensen's inequality implies that

$$\left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1}) \right) \right]^2 \leq -c \ln \frac{W_{t+1}}{W_t}$$

Thus,

$$nL_n(\tilde{g}) = \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2$$

Thus,

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)\right]^2 \end{aligned}$$

Thus,

$$
\begin{aligned}
nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\
&= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1}) \right) \right]^2 \\
&\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t}
\end{aligned}
$$

Thus,

$$\begin{aligned}
nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\
&= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)\right]^2 \\
&\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\
&= -c \ln W_{n+1}
\end{aligned}$$

Thus,

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)\right]^2 \\ &\le -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\ &= -c \ln W_{n+1} \end{aligned}$$

and therefore

$$nL_n(\tilde{g}) \le -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right)$$

Thus,

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)\right]^2 \\ &\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\ &= -c \ln W_{n+1} \end{aligned}$$

and therefore

$$\begin{aligned} nL_n(\tilde{g}) &\leq -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\ &= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c}\right) \end{aligned}$$

Thus,

$$
\begin{aligned}
nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\
&= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1})\right)\right]^2 \\
&\le -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\
&= -c \ln W_{n+1}
\end{aligned}
$$

and therefore

$$
\begin{aligned}
nL_n(\tilde{g}) &\le -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k}\right) \\
&= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c}\right) \\
&\le -c \ln \left(\sup q_k e^{-nL_n(\tilde{h}_k)/c}\right)
\end{aligned}
$$

Thus,

$$\begin{aligned} nL_n(\tilde{g}) &= \sum_{t=1}^{n} \left(y_t - \tilde{g}(x_1^t, y_1^{t-1})\right)^2 \\ &= \sum_{t=1}^{n} \left[\sum_{k=1}^{\infty} v_{t,k} \left(y_t - \tilde{h}_k(x_1^t, y_1^{t-1}) \right) \right]^2 \\ &\leq -c \sum_{t=1}^{n} \ln \frac{W_{t+1}}{W_t} \\ &= -c \ln W_{n+1} \end{aligned}$$

and therefore

$$\begin{aligned} nL_n(\tilde{g}) &\leq -c \ln \left(\sum_{k=1}^{\infty} w_{n+1,k} \right) \\ &= -c \ln \left(\sum_{k=1}^{\infty} q_k e^{-nL_n(\tilde{h}_k)/c} \right) \\ &\leq -c \ln \left(\sup q_k e^{-nL_n(\tilde{h}_k)/c} \right) \end{aligned}$$

Sketch of the proof of the Theorem

Because of the fundamental limit

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{a.s.}$$

Sketch of the proof of the Theorem

Because of the fundamental limit

$$\liminf_{n\to\infty} L_n(g) \geq \lim_{n\to\infty} L_n(g^*) = L^* \quad \text{a.s.}$$

it is enough to show that

$$\limsup_{n\to\infty} L_n(g) \leq L^* \qquad \text{a.s.}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

By the Expert Lemma

$$L_n(g) \leq \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\limsup_{n\to\infty} L_n(g) \quad \leq \quad \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right)$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right)
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)})
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell}
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty} \epsilon_{k,\ell}
\end{aligned}$$

By the Expert Lemma

$$L_n(g) \le \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right),$$

and therefore, almost surely,

$$\begin{aligned}
\limsup_{n\to\infty} L_n(g) &\le \limsup_{n\to\infty} \inf_{k,\ell} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} \left(L_n(h^{(k,\ell)}) - \frac{c \ln q_{k,\ell}}{n} \right) \\
&\le \inf_{k,\ell} \limsup_{n\to\infty} L_n(h^{(k,\ell)}) \\
&=: \inf_{k,\ell} \epsilon_{k,\ell} \\
&= \lim_{k,\ell\to\infty} \epsilon_{k,\ell} \\
&= L^*
\end{aligned}$$

Pattern recognition

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

August 27, 2012

e-mail: gyorfi@cs.bme.hu
www.cs.bme.hu/˜gyorfi

Decision problem

Y $\{1, 2, \dots M\}$ valued
X feature vector

Decision problem

Y $\{1, 2, \dots M\}$ valued
X feature vector
Classifier

$$g : \mathbb{R}^d \to \{1, 2, \dots M\}.$$

Decision problem

Y $\{1, 2, \ldots M\}$ valued
X feature vector
Classifier

$$g : \mathbb{R}^d \to \{1, 2, \ldots M\}.$$

Probability of error:

$$P(g(X) \neq Y).$$

Decision problem

Y $\{1, 2, \ldots M\}$ valued
X feature vector
Classifier

$$g : \mathbb{R}^d \to \{1, 2, \ldots M\}.$$

Probability of error:

$$P(g(X) \neq Y).$$

Problem: find

$$\min_g P(g(X) \neq Y).$$

Bayes decision

Bayes decision

a posteriori probability

$$P_i(x) = \mathbf{P}\{Y = i | X = x\}.$$

Bayes decision

a posteriori probability

$$P_i(x) = \mathbf{P}\{Y = i | X = x\}.$$

Bayes decision

$$g^*(x) = \arg\max_i P_i(x).$$

Bayes decision

a posteriori probability

$$P_i(x) = \mathbf{P}\{Y = i | X = x\}.$$

Bayes decision

$$g^*(x) = \arg\max_i P_i(x).$$

L^* Bayes error

Lemma

For any decision g,

$$L^* := \mathbf{P}\{g^*(X) \leq Y\} \leq \mathbf{P}\{g(X) \neq Y\}.$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\mathbf{P}\{g(X) \neq Y \mid X\}$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\mathbf{P}\{g(X) \neq Y \mid X\} \quad = \quad 1 - \mathbf{P}\{g(X) = Y \mid X\}$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\begin{aligned} \mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \mathbf{P}\{g(X) = Y \mid X\} \\ &= \end{aligned}$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\begin{aligned} \mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \mathbf{P}\{g(X) = Y \mid X\} \\ &= 1 - \sum_{j=1}^{M} \mathbf{P}\{g(X) = j, Y = j \mid X\} \end{aligned}$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \mathbf{P}\{g(X) = Y \mid X\} \\
&= 1 - \sum_{j=1}^{M} \mathbf{P}\{g(X) = j, Y = j \mid X\} \\
&= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \mathbf{P}\{Y = j \mid X\}
\end{aligned}$$

Proof

$$\mathbf{P}\{g(X) \neq Y\} = \mathbf{E}\{\mathbf{P}\{g(X) \neq Y \mid X\}\}$$

and

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \mathbf{P}\{g(X) = Y \mid X\} \\
&= 1 - \sum_{j=1}^{M} \mathbf{P}\{g(X) = j, Y = j \mid X\} \\
&= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \mathbf{P}\{Y = j \mid X\} \\
&= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X)
\end{aligned}$$

Thus

$$\mathbf{P}\{g(X) \neq Y \mid X\}$$

Thus

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X)$$

Thus

$$\begin{aligned}\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) \\ &\geq 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \max_k P_k(X)\end{aligned}$$

Thus

$$\begin{aligned}\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) \\ &\geq 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \max_k P_k(X) \\ &= 1 - \max_k P_k(X)\end{aligned}$$

Thus

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) \\
&\geq 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \max_k P_k(X) \\
&= 1 - \max_k P_k(X) \\
&= 1 - \sum_{j=1}^{M} I_{\{g^*(X)=j\}} \max_k P_k(X)
\end{aligned}$$

Thus

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) \\
&\geq 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \max_k P_k(X) \\
&= 1 - \max_k P_k(X) \\
&= 1 - \sum_{j=1}^{M} I_{\{g^*(X)=j\}} \max_k P_k(X) \\
&= 1 - \sum_{j=1}^{M} I_{\{g^*(X)=j\}} P_j(X)
\end{aligned}$$

Thus

$$
\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} &= 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) \\
&\geq 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} \max_k P_k(X) \\
&= 1 - \max_k P_k(X) \\
&= 1 - \sum_{j=1}^{M} I_{\{g^*(X)=j\}} \max_k P_k(X) \\
&= 1 - \sum_{j=1}^{M} I_{\{g^*(X)=j\}} P_j(X) \\
&= \mathbf{P}\{g^*(X) \neq Y \mid X\}
\end{aligned}
$$

Plug-in rule

$\tilde{P}_i(x)$ approximations of $P_i(x)$

Plug-in rule

$\tilde{P}_i(x)$ approximations of $P_i(x)$
plug-in decision g

$$g(x) = \arg\max_i \tilde{P}_i(x).$$

Plug-in rule

$\tilde{P}_i(x)$ approximations of $P_i(x)$
plug-in decision g

$$g(x) = \arg\max_i \tilde{P}_i(x).$$

Lemma

$$\mathbf{P}\{g(X) \neq Y\} - L^* \leq \sum_{j=1}^{M} \mathbf{E}\{|P_j(X) - \tilde{P}_j(X)|\}.$$

Proof

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X)$$

Proof

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) = 1 - P_{g(X)}(X)$$

Proof

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) = 1 - P_{g(X)}(X)$$

Thus

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\}$$

Proof

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) = 1 - P_{g(X)}(X)$$

Thus

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} = P_{g^*(X)}(X) - P_{g(X)}(X)$$

Proof

$$\mathbf{P}\{g(X) \neq Y \mid X\} = 1 - \sum_{j=1}^{M} I_{\{g(X)=j\}} P_j(X) = 1 - P_{g(X)}(X)$$

Thus

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} = P_{g^*(X)}(X) - P_{g(X)}(X)$$

If $g^*(X) = g(X)$ then

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} = 0$$

If $g^*(X) \neq g(X)$ then

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\}$$

If $g^*(X) \neq g(X)$ then

$$\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} = P_{g^*(X)}(X) - P_{g(X)}(X)$$

If $g^*(X) \neq g(X)$ then

$$
\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} &= P_{g^*(X)}(X) - P_{g(X)}(X) \\
&= P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X)
\end{aligned}
$$

If $g^*(X) \neq g(X)$ then

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} &= P_{g^*(X)}(X) - P_{g(X)}(X) \\
&= P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X) \\
&\quad + \tilde{P}_{g^*(X)}(X) - \tilde{P}_{g(X)}(X)
\end{aligned}$$

If $g^*(X) \neq g(X)$ then

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} &= P_{g^*(X)}(X) - P_{g(X)}(X) \\
&= P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X) \\
&\quad + \tilde{P}_{g^*(X)}(X) - \tilde{P}_{g(X)}(X) \\
&\quad + \tilde{P}_{g(X)}(X) - P_{g(X)}(X)
\end{aligned}$$

If $g^*(X) \neq g(X)$ then

$$
\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} &= P_{g^*(X)}(X) - P_{g(X)}(X) \\
&= P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X) \\
&\quad + \tilde{P}_{g^*(X)}(X) - \tilde{P}_{g(X)}(X) \\
&\quad + \tilde{P}_{g(X)}(X) - P_{g(X)}(X) \\
&\leq P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X)
\end{aligned}
$$

If $g^*(X) \neq g(X)$ then

$$\begin{aligned}
\mathbf{P}\{g(X) \neq Y \mid X\} - \mathbf{P}\{g^*(X) \neq Y \mid X\} &= P_{g^*(X)}(X) - P_{g(X)}(X) \\
&= P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X) \\
&\quad +\tilde{P}_{g^*(X)}(X) - \tilde{P}_{g(X)}(X) \\
&\quad +\tilde{P}_{g(X)}(X) - P_{g(X)}(X) \\
&\leq P_{g^*(X)}(X) - \tilde{P}_{g^*(X)}(X) \\
&\quad +\tilde{P}_{g(X)}(X) - P_{g(X)}(X) \\
&\leq \sum_{j=1}^{M} |P_j(X) - \tilde{P}_j(X)|.
\end{aligned}$$

Universal consistency

Data: $(X_1, Y_1), \ldots, (X_n, Y_n)$

Data: $(X_1, Y_1), \dots, (X_n, Y_n)$

$$g_n(x) = g_n((X_1, Y_1), \dots, (X_n, Y_n), x).$$

Definition

The classifier g_n is called **weakly universally consistent** if

$$P(g_n(X) \neq Y) \to L^*$$

for all distributions of (X, Y).

Local majority voting

the a posteriori probabilities

$$P_i(x) = \mathbf{P}\{Y = i | X = x\} = \mathbf{E}\{I_{\{Y=i\}} | X = x\}$$

are regression functions,

Local majority voting

the a posteriori probabilities

$$P_i(x) = \mathbf{P}\{Y = i | X = x\} = \mathbf{E}\{I_{\{Y=i\}} | X = x\}$$

are regression functions, their local averaging estimates

$$\tilde{P}_{n,j}(x) = \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i = j\}}$$

Local majority voting

the a posteriori probabilities

$$P_i(x) = \mathbf{P}\{Y = i | X = x\} = \mathbf{E}\{I_{\{Y=i\}} | X = x\}$$

are regression functions, their local averaging estimates

$$\tilde{P}_{n,j}(x) = \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}}$$

plug-in rule: local majority voting

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} \tilde{P}_{n,j}(x) = \arg\max_j \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}},$$

Local majority voting

the a posteriori probabilities

$$P_i(x) = \mathbf{P}\{Y = i | X = x\} = \mathbf{E}\{I_{\{Y=i\}} | X = x\}$$

are regression functions, their local averaging estimates

$$\tilde{P}_{n,j}(x) = \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}}$$

plug-in rule: local majority voting

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} \tilde{P}_{n,j}(x) = \arg\max_j \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}},$$

$$\begin{aligned} \mathbf{P}\{g(X) \neq Y\} - L^* &\leq \sum_{j=1}^{M} \mathbf{E}\{|P_j(X) - \tilde{P}_{n,j}(X)|\} \\ &\leq \sum_{j=1}^{M} \sqrt{\mathbf{E}\{|P_j(X) - \tilde{P}_{n,j}(X)|^2\}}. \end{aligned}$$

Local majority voting

***k*-nearest neighbor rule**

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}},$$

Local majority voting

k-**nearest neighbor rule**

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}},$$

Partitioning rule:

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} I_{\{X_i \in A_n(x)\}} I_{\{Y_i=j\}}$$

Local majority voting

k-**nearest neighbor rule**

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} W_{n,i}(x) I_{\{Y_i=j\}},$$

Partitioning rule:

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} I_{\{X_i \in A_n(x)\}} I_{\{Y_i=j\}}$$

Kernel rule rule:

$$g_n(x) = \arg\max_j \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) I_{\{Y_i=j\}}.$$

Local majority voting

k-**nearest neighbor rule**

$$g_n(x) = \arg\max_j \sum_{i=1}^n W_{n,i}(x) I_{\{Y_i=j\}},$$

Partitioning rule:

$$g_n(x) = \arg\max_j \sum_{i=1}^n I_{\{X_i \in A_n(x)\}} I_{\{Y_i=j\}}$$

Kernel rule rule:

$$g_n(x) = \arg\max_j \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) I_{\{Y_i=j\}}.$$

Theorem

The k-NN rule and the partitioning rule and the kernel rule are strongly universally consistent.

Empirical error minimization

empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

Empirical error minimization

empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

class of classifiers $\mathcal{G}_n$

Empirical error minimization

empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

class of classifiers $\mathcal{G}_n$

select a classifier from $\mathcal{G}_n$ which minimizes the empirical error:

Empirical error minimization

empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

class of classifiers $\mathcal{G}_n$
select a classifier from $\mathcal{G}_n$ which minimizes the empirical error:
$g_n \in \mathcal{G}_n$ and

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)\neq Y_j\}} = \min_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

Empirical error minimization

empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

class of classifiers $\mathcal{G}_n$
select a classifier from $\mathcal{G}_n$ which minimizes the empirical error:
$g_n \in \mathcal{G}_n$ and

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)\neq Y_j\}} = \min_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

the VC dimension of $\mathcal{G}_n$ grows slowly as n grows

Empirical error minimization

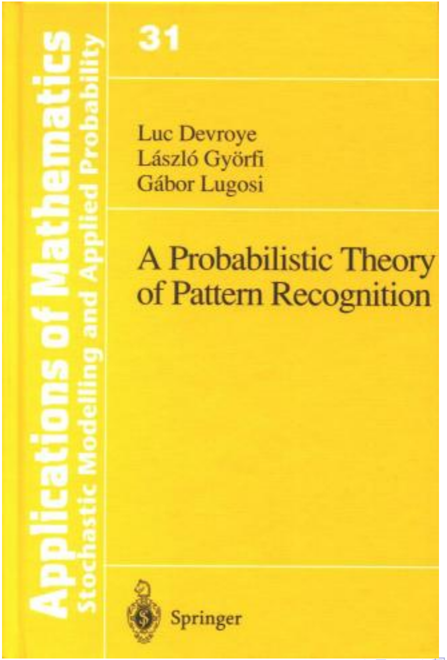
empirical error

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

class of classifiers $\mathcal{G}_n$
select a classifier from $\mathcal{G}_n$ which minimizes the empirical error:
$g_n \in \mathcal{G}_n$ and

$$\frac{1}{n}\sum_{j=1}^{n} I_{\{g_n(X_j)\neq Y_j\}} = \min_{g\in\mathcal{G}_n} \frac{1}{n}\sum_{j=1}^{n} I_{\{g(X_j)\neq Y_j\}}$$

the VC dimension of $\mathcal{G}_n$ grows slowly as n grows
Examples for $\mathcal{G}_n$:
- polynomial classifiers
- tree classifiers
- neural networks classifiers
- radial basis functions classifiers


Applications of Mathematics
Stochastic Modelling and Applied Probability
31
Luc Devroye
László Györfi
Gábor Lugosi
A Probabilistic Theory of Pattern Recognition
Springer

Prediction of time series: $0-1$ loss

László (Laci) Györfi[1]

[1]Department of Computer Science and Information Theory
Budapest University of Technology and Economics
Budapest, Hungary

August 27, 2012

e-mail: gyorfi@cs.bme.hu
www.cs.bme.hu/∼gyorfi

$0 - 1$ loss

Y_i takes values in the finite set $\{0, 1\}$.

$0-1$ loss

Y_i takes values in the finite set $\{0,1\}$.
At time instant i the classifier decides on Y_i based on the past observation (X_1^i, Y_1^{i-1}).

$0-1$ loss

Y_i takes values in the finite set $\{0,1\}$.
At time instant i the classifier decides on Y_i based on the past observation (X_1^i, Y_1^{i-1}).
After n round the empirical $0-1$ error for X_1^n, Y_1^n is

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} I_{\{f(X_1^i, Y_1^{i-1}) \neq Y_i\}},$$

i.e., the loss is the $0-1$ loss, and $R_n(f)$ is the relative frequency of errors.

Dependent data: time series

data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ form a stationary and ergodic process

Dependent data: time series

data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ form a stationary and ergodic process

Optimal classification scheme:

$$f_t^*(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } \mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Dependent data: time series

data $D_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ form a stationary and ergodic process

Optimal classification scheme:

$$f_t^*(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } \mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Fundamental limit: for any classification strategy f and stationary ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n\to\infty} R_n(f) \geq \lim_{n\to\infty} R_n(f^*) = R^* \quad \text{a.s.},$$

Dependent data: time series

data $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ form a stationary and ergodic process

Optimal classification scheme:

$$f_t^*(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } \mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Fundamental limit: for any classification strategy f and stationary ergodic process $\{(X_n, Y_n)\}_{n=-\infty}^{\infty}$,

$$\liminf_{n\to\infty} R_n(f) \ge \lim_{n\to\infty} R_n(f^*) = R^* \quad \text{a.s.,}$$

where

$$R^* = \mathbf{E}\left\{\min\left(\mathbf{P}\{Y_0 = 1 | X_{-\infty}^0, Y_{-\infty}^{-1}\}, \mathbf{P}\{Y_0 = 0 | X_{-\infty}^0, Y_{-\infty}^{-1}\}\right)\right\}.$$

there are universally consistent classifier sequence f_n:

$$\lim_{n\to\infty} R_n(f) = R^*$$

a.s. for all stationary and ergodic sequence there are universally consistent classifier sequence f_n:

$$\lim_{n \to \infty} R_n(f) = R^*$$

a.s. for all stationary and ergodic sequence

Such classifier sequence is called **universally consistent**.

Theorem

Let $g_t(X_1^t, Y_1^{t-1})$ be a universally consistent prediction scheme, for bounded Y, estimating the conditional probability

$$\mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} = \mathbf{E}\{Y_t \mid X_1^t, Y_1^{t-1}\}.$$

Theorem

Let $g_t(X_1^t, Y_1^{t-1})$ be a universally consistent prediction scheme, for bounded Y, estimating the conditional probability

$$\mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} = \mathbf{E}\{Y_t \mid X_1^t, Y_1^{t-1}\}.$$

The corresponding classification scheme:

$$f_t(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } g_t(X_1^t, Y_1^{t-1}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Theorem

Let $g_t(X_1^t, Y_1^{t-1})$ be a universally consistent prediction scheme, for bounded Y, estimating the conditional probability

$$\mathbf{P}\{Y_t = 1 \mid X_1^t, Y_1^{t-1}\} = \mathbf{E}\{Y_t \mid X_1^t, Y_1^{t-1}\}.$$

The corresponding classification scheme:

$$f_t(X_1^t, Y_1^{t-1}) = \begin{cases} 1 & \text{if } g_t(X_1^t, Y_1^{t-1}) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

Then f_t is a universally consistent classifier (Györfi, Lugosi (2001)).

Martingale difference sequences

for the proof of the Theorem we use the concept of martingale differences:

Martingale difference sequences

for the proof of the Theorem we use the concept of martingale differences:

Definition

there are two sequences of random variables:

$$\{Z_n\} \qquad \{X_n\}$$

- Z_n is a function of $X_1, \ldots, X_n$,
- $\mathbf{E}\{Z_n \mid X_1, \ldots, X_{n-1}\} = 0$ almost surely.

Then $\{Z_n\}$ is called martingale difference sequence with respect to $\{X_n\}$.

A strong law of large numbers

Chow Theorem:

A strong law of large numbers

Chow Theorem: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ and

$$\sum_{n=1}^{\infty} \frac{\mathbf{E}\{Z_n^2\}}{n^2} < \infty$$

then

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} Z_i = 0 \text{ a.s.}$$

A weak law of large numbers

Lemma:

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\mathbf{E}\{Z_i Z_j\}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\mathbf{E}\{Z_i Z_j\} = \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \ldots, X_{j-1}\}\}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\begin{aligned}\mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \ldots, X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \ldots, X_{j-1}\}\}\end{aligned}$$

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$
\begin{aligned}
\mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \dots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \dots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \cdot 0\}
\end{aligned}
$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.

Proof. Put $i < j$.

$$\begin{aligned}
\mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \dots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \dots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \cdot 0\} = 0
\end{aligned}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\begin{aligned}\mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \dots, X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \dots, X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \cdot 0\} = 0\end{aligned}$$

Lemma:

$$\mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)^2\right\}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.

Proof. Put $i < j$.

$$\begin{aligned}
\mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \ldots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \ldots, X_{j-1}\}\} \\
&= \mathbf{E}\{Z_i \cdot 0\} = 0
\end{aligned}$$

Lemma:

$$\mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)^2\right\} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\{Z_i Z_j\}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\begin{aligned} \mathbf{E}\{Z_i Z_j\} &= \mathbf{E}\{\mathbf{E}\{Z_i Z_j \mid X_1, \dots, X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \mathbf{E}\{Z_j \mid X_1, \dots, X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \cdot 0\} = 0 \end{aligned}$$

Lemma:

$$\begin{aligned} \mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)^2\right\} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{E}\{Z_i Z_j\} \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbf{E}\{Z_i^2\} \end{aligned}$$

A weak law of large numbers

Lemma: If $\{Z_n\}$ is a martingale difference sequence with respect to $\{X_n\}$ then $\{Z_n\}$ are uncorrelated.
Proof. Put $i < j$.

$$\begin{aligned}\mathbf{E}\{Z_iZ_j\} &= \mathbf{E}\{\mathbf{E}\{Z_iZ_j \mid X_1,\ldots,X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i\mathbf{E}\{Z_j \mid X_1,\ldots,X_{j-1}\}\} \\ &= \mathbf{E}\{Z_i \cdot 0\} = 0\end{aligned}$$

Lemma:

$$\begin{aligned}\mathbf{E}\left\{\left(\frac{1}{n}\sum_{i=1}^n Z_i\right)^2\right\} &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n \mathbf{E}\{Z_iZ_j\} \\ &= \frac{1}{n^2}\sum_{i=1}^n \mathbf{E}\{Z_i^2\} \\ &\to 0\end{aligned}$$

if, for example, $\mathbf{E}\{Z_i^2\}$ is a bounded sequence.

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$
Put

$$Z_n = Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\}$$

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$
Put

$$Z_n = Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\}$$

Then $\{Z_n\}$ is a martingale difference sequence:

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$
Put

$$Z_n = Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\}$$

Then $\{Z_n\}$ is a martingale difference sequence:

- Z_n is a function of $X_1, \ldots, X_n$,

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$
Put

$$Z_n = Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\}$$

Then $\{Z_n\}$ is a martingale difference sequence:

- Z_n is a function of $X_1, \ldots, X_n$,
-

$$\mathbf{E}\{Z_n \mid X_1, \ldots, X_{n-1}\}$$

Constructing martingale difference sequence

$\{Y_n\}$ is an arbitrary sequence such that Y_n is a function of $X_1, \ldots, X_n$

Put

$$Z_n = Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\}$$

Then $\{Z_n\}$ is a martingale difference sequence:

- Z_n is a function of $X_1, \ldots, X_n$,
- $$\begin{aligned} & \mathbf{E}\{Z_n \mid X_1, \ldots, X_{n-1}\} \\ = & \mathbf{E}\{Y_n - \mathbf{E}\{Y_n \mid X_1, \ldots, X_{n-1}\} \mid X_1, \ldots, X_{n-1}\} \\ = & 0 \end{aligned}$$

  almost surely.

Sketch of the proof of the Theorem

Because of the fundamental limit, we have to show that

$$\lim_{n\to\infty} (R_n(f) - R_n(f^*)) = 0 \quad \text{a.s.}$$

Sketch of the proof of the Theorem

Because of the fundamental limit, we have to show that

$$\lim_{n\to\infty}(R_n(f) - R_n(f^*)) = 0 \quad \text{a.s.}$$

Put

$$\bar{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\{I_{\{f(X_1^i, Y_1^{i-1}) \neq Y_i\}} \mid X_1^i, Y_1^{i-1}\},$$

Sketch of the proof of the Theorem

Because of the fundamental limit, we have to show that

$$\lim_{n\to\infty}(R_n(f) - R_n(f^*)) = 0 \quad \text{a.s.}$$

Put

$$\bar{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\{I_{\{f(X_1^i,Y_1^{i-1})\neq Y_i\}} \mid X_1^i, Y_1^{i-1}\},$$

and

$$\bar{R}_n(f^*) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\{I_{\{f^*(X_1^i,Y_1^{i-1})\neq Y_i\}} \mid X_1^i, Y_1^{i-1}\}.$$

Sketch of the proof of the Theorem

Because of the fundamental limit, we have to show that

$$\lim_{n\to\infty}(R_n(f) - R_n(f^*)) = 0 \quad \text{a.s.}$$

Put

$$\bar{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\{I_{\{f(X_1^i, Y_1^{i-1}) \neq Y_i\}} \mid X_1^i, Y_1^{i-1}\},$$

and

$$\bar{R}_n(f^*) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}\{I_{\{f^*(X_1^i, Y_1^{i-1}) \neq Y_i\}} \mid X_1^i, Y_1^{i-1}\}.$$

Then $R_n(f) - \bar{R}_n(f)$ and $R_n(f^*) - \bar{R}_n(f^*)$ are the averages of bounded martingale differences, therefore

Sketch of the proof of the Theorem

Because of the fundamental limit, we have to show that

$$\lim_{n\to\infty}(R_n(f) - R_n(f^*)) = 0 \quad \text{a.s.}$$

Put

$$\bar{R}_n(f) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\{I_{\{f(X_1^i,Y_1^{i-1})\neq Y_i\}} \mid X_1^i, Y_1^{i-1}\},$$

and

$$\bar{R}_n(f^*) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\{I_{\{f^*(X_1^i,Y_1^{i-1})\neq Y_i\}} \mid X_1^i, Y_1^{i-1}\}.$$

Then $R_n(f) - \bar{R}_n(f)$ and $R_n(f^*) - \bar{R}_n(f^*)$ are the averages of bounded martingale differences, therefore

$$R_n(f) - \bar{R}_n(f) \to 0 \quad \text{a.s.}$$

and

$$R_n(f^*) - \bar{R}_n(f^*) \to 0 \quad \text{a.s.}$$

Task left:

$$\bar{R}_n(f^*) - \bar{R}_n(f) \to 0 \quad \text{a.s.}$$

Task left:

$$\bar{R}_n(f^*) - \bar{R}_n(f) \to 0 \quad \text{a.s.}$$

By the Plug-in Lemma,

$$\begin{aligned} |\bar{R}_n(f^*) - \bar{R}_n(f)| &\leq \frac{1}{n}\sum_{i=1}^{n} |g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\}| \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} |g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\}|^2} \end{aligned}$$

Task left:

$$\bar{R}_n(f^*) - \bar{R}_n(f) \to 0 \quad \text{a.s.}$$

By the Plug-in Lemma,

$$\begin{aligned} |\bar{R}_n(f^*) - \bar{R}_n(f)| &\leq \frac{1}{n}\sum_{i=1}^{n} |g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\}| \\ &\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n} |g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\}|^2} \end{aligned}$$

Need

$$\frac{1}{n}\sum_{i=1}^{n} |g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\}|^2 \to 0 \quad \text{a.s.}$$

Corollary

Let $\{g_n\}$ be a sequence of universally consistent predictors for the class of stationary, ergodic processes with $|Y| < B$,

Corollary

Let $\{g_n\}$ be a sequence of universally consistent predictors for the class of stationary, ergodic processes with $|Y| < B$, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for the class of stationary and ergodic sequences with $|Y| < B$.

Corollary

Let $\{g_n\}$ be a sequence of universally consistent predictors for the class of stationary, ergodic processes with $|Y| < B$, i.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 = L^*$$

a.s. for the class of stationary and ergodic sequences with $|Y| < B$. Then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})^2 = 0$$

a.s.

Proof

Decomposition

$$
\begin{aligned}
& (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 \\
= \; & (g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})^2 \\
& +2(g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i) \\
& +(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i)^2
\end{aligned}
$$

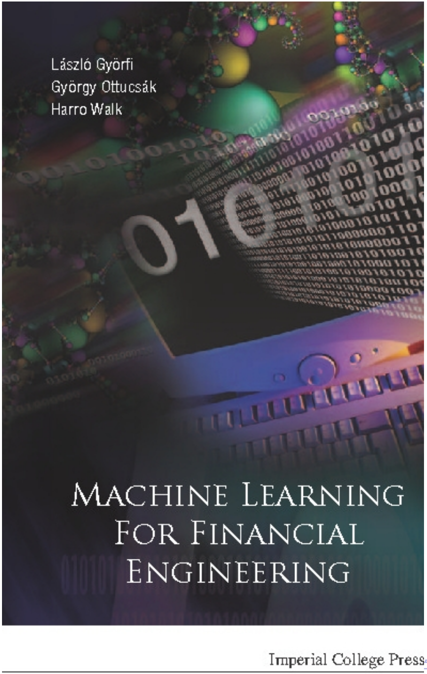
Proof

Decomposition

$$
\begin{aligned}
& (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 \\
= \; & (g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})^2 \\
& +2(g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i) \\
& +(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i)^2
\end{aligned}
$$

Thus

$$
\begin{aligned}
& (g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})^2 \\
= \; & (g_i(X_1^i, Y_1^{i-1}) - Y_i)^2 \\
& -2(g_i(X_1^i, Y_1^{i-1}) - \mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\})(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i) \\
& -(\mathbf{E}\{Y_i \mid X_{-\infty}^i, Y_{-\infty}^{i-1}\} - Y_i)^2
\end{aligned}
$$


László Györfi
György Ottucsák
Harro Walk
MACHINE LEARNING
FOR FINANCIAL
ENGINEERING
Imperial College Press