Neural Network Interest Group

Título/*Title*:
MLE, MSE *et alia*

Autor(es)/*Author(s)*:
J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 1 /2006

FEUP/INEB, Rua Dr. Roberto Frias, 4200-465, Porto, Portugal

Título/*Title*:

MLE, MSE *et alia*

Autor(es)/*Author(s)*:

J.P. Marques de Sá

Relatório Técnico/*Technical Report* No. 1 /2006

Publicado por/*Published by*: NNIG. http://paginas.fe.up.pt/~nnig/

Janeiro de 2006

Índice

1 Estimação Pontual

1.1 Estimação pontual de um parâmetro de uma distribuição

É dado $X_1,...,X_n$ i.i.d. de $f(x;\theta)$ (ou $P(x;\theta)$ no caso discreto)[†], com $\theta \in \Theta$. Como estimar θ?

Definição 1-1

$t_n(X_1,...,X_n)$ é uma v.a. designada por *estatística* ou *estimador* de θ. Um valor particular de t_n, $t_n(x_1,\ldots,x_n)$, é uma *estimativa* de θ.

■

Definição 1-2

Estimador não-enviezado: $\mathrm{E}[t_n(X_1,...,X_n)] = \theta \quad (\forall\,\theta \in \Theta)$.

Viés: $b_n(\theta) \stackrel{\Delta}{=} \mathrm{E}[t_n(X_1,...,X_n)] - \theta$

(A esperança é avaliada com $F_{X_1,...,X_n}(x_1,...,x_n \mid \theta)$.)

■

Definição 1-3

Estimador assimptoticamente não-enviezado: $\lim_{n\to\infty} b_n(\theta) = 0, \quad \forall\,\theta \in \Theta$.

■

Definição 1-4

A sequência t_1, t_2,…, $t_n(X_1,...,X_n)$ é uma *sequência de estimadores consistente* de θ, sse:

$t_n(X_1,\ldots,X_n) \xrightarrow{p} \theta, \quad \forall\,\theta \in \Theta$

■

Exemplo 1-1

$\overline{X}_n = \sum_{i=1}^{n} X_i / n$ é um estimador não enviezado e consistente de μ. (ver caso particular para seq. de Bernoulli). (Ver prova em [3].)

$s_n^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 / (n-1)$ é um estimador não enviezado e consistente de σ. (Ver prova em [3].)

[†] Por $f(x;\theta)$ (ou $P(x;\theta)$) entenda-se $f_X(x;\theta)$ (ou $P_X(x;\theta)$). Em geral referir-nos-emos a v.a. contínuas. A transposição de resultados para v.a. discretas não oferece dificuldades. As v.a. podem ser uni- ou multi-dimensionais.

Exemplo 1-2

Note-se que se tivermos mais do que um estimador não-enviezado, então teremos uma infinidade de estimadores não enviezados.

De facto, se $\mathrm{E}[t_i(X_1,...,X_n)]=\theta$, então $\mathrm{E}\left[\sum_i w_i t_i(X_1,...,X_n)\right]=\theta$, desde que $\sum_i w_i = 1$.

(Ver prova em [3].)

Por vezes uma estimativa não-enviezada pode não ser a "melhor". Pode, até, ser absurda.

Exemplo 1-3

Suponhamos que $X_1,..,X_n$ são v.a. i.i.d. de Bernoulli com parâmetro p (e $q = 1-p$). Seja n o número de experiências necessárias até obter o primeiro sucesso. Suponhamos que pretendemos obter estimadores não-enviezados de p que sejam funções apenas de n, $J(n)$:

$$\mathrm{E}[J(n)] = p$$

Mas $P(n=k)=q^{k-1}p$, k = 1, 2, ...(distribuição geométrica). Logo, a esperança calcula-se como se segue:

$$\sum_{k=1}^{\infty} J(n)q^{k-1}p = p$$

Se q = 1 (e p = 0) a equação anterior corresponde a 0 = 0. Se $q \neq 1$ (e $p \neq 0$), temos

$$\sum_{k=1}^{\infty} J(n)q^{k-1} = 1 \qquad \text{ou seja} \qquad \sum_{k=0}^{\infty} J(n)q^{k} = 1 + \sum_{k=1}^{\infty} 0.q^{k}$$

Ora, duas séries infinitas de potências só produzem o mesmo valor para a variável independente se os coeficientes forem iguais, o que implica:

$$\hat{p} = J(n) = \begin{cases} 1 & n = 1 \\ 0 & n > 1 \end{cases}$$

e obtemos o estimador não-enviezado, mas absurdo: observe-se X_1; se sucesso, $\hat{p}=1$; caso contrário $\hat{p}=0$. De facto, num número muito grande, N, de repetições da experiência obtemos um sucesso em X_1 cerca de Np das vezes; logo, uma estimativa empírica $\approx(1.Np + 0.Nq)/N = p$.

Definição 1-5

$t_n(X_1,...,X_n)$ é *consistente em média quadrática* sse:

$$\mathrm{E}[(t_n(X_1,\ldots,X_n)-\theta)^2] \underset{n\to\infty}{\to} 0, \quad \forall\theta\in\Theta$$

A quantidade $\mathrm{E}[(t_n-\theta)^2]$ é designada por *erro quadrático médio* de t_n.

■

Exemplo 1-4

$t_n(X_1,...,X_n)$ é *consistente em média quadrática* sse:

- $Var[t_n] \underset{n\to\infty}{\to} 0$
- t_n é assimptoticamente não-enviezado.

(Ver prova em [3].)

Exemplo 1-5

Se $\mathrm{E}\left[(t_n-\theta)^2\right] \underset{n\to\infty}{\to} 0 \quad \Rightarrow \quad t_n \overset{p}{\to} \theta \quad (\forall\theta)$ (Ver prova em [3].)

1.2 Estimação pontual de uma v.a.

Existe uma v.a. *X*, desconhecida, e certos dados *Z* de alguma forma relacionados com *X*. São conhecidas as estatísticas conjuntas de *X* e *Z* (*f*(*x*, *z*) ou *P*(*x*, *z*)). Como estimar *X* ?

Definição 1-6

Uma estimativa $\hat{X}$ de *X* diz-se *não-enviezada* se $\mathrm{E}[\hat{X}] = X$.

■

Seja $\widetilde{X} = X - \hat{X}$ o desvio entre *X* e a sua estimativa. Decorre da linearidade de E[] que:

$$\mathrm{E}[\widetilde{X}] = 0$$

sse a estimativa for não-enviezada.

2 Estimação de Máxima Verosimilhança

Maximum-likelihood estimation (MLE)

A estimação de máxima verosimilhança aplica-se a problemas de estimação em que existe uma quantidade *determinística x* sobre a qual nada sabemos e um processo estocástico (de medida) *Z* de alguma forma relacionado com *x*. A MLE proporciona uma estimativa pontual de *x* (nem sempre não-enviezada, como veremos).

A ideia é: Escolher a estimativa que maximiza o valor da probabilidade de *Z* dado *x*.

Definição 2-1

A *função de verosimilhança* de *Z* dado *x* é:

$$L(x) \stackrel{\Delta}{=} \begin{cases} f_{Z|x}(z \mid x) & (\textit{contínua}) \\ P_{Z|x}(z \mid x) & (\textit{discreta}) \end{cases}$$

■

Atenção: Como função de *x*, *L*(*x*) **não é** uma fdp (ou função de probabilidade). Por exemplo, o integral de *L*(*x*) pode não ser 1.

A MLE aplica-se frequentemente na determinação um parâmetro θ de uma distribuição com base em *N* medições $X_1,...,X_n$. Isto é, sejam $X_1,...,X_n$ *n* v.a. (i.i.d. ou não) com função de distribuição $F(X_1,...,X_n \mid \theta)$ e θ desconhecido. A *função de verosimilhança* de $X_1,...,X_n$ é:

$$L(\theta) \stackrel{\Delta}{=} \begin{cases} f(X_1,...,X_n \mid \theta) & (\textit{contínua}) \\ P(X_1,...,X_n \mid \theta) & (\textit{discreta}) \end{cases}$$

Se as variáveis são independentes: $L(\theta) = \prod_{i=1}^{n} f_i(X_i \mid \theta)$ (P_i, para v.a. discreta)

Se são i.i.d.: $L(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$ (P, para v.a. discreta)

Considere-se, p. ex., a situação discreta. Seja $P(X_1,...,X_n,\theta)$ a função de probabilidade conjunta dos X_i e θ. Temos:

$$P(X_1,...,X_n,\theta) = P(X_1,...,X_n \mid \theta)P(\theta)$$

Sejam dois valores alternativos, θ_1 e θ_2, com igual probabilidade: $P(\theta_1) = P(\theta_2)$. Então, faz sentido escolher θ_1 se $P(X_1,...,X_n \mid \theta_1) > P(X_1,...,X_n \mid \theta_2)$. O parâmetro θ_1 é o mais verosímil.

Em condições muito gerais (lei fraca dos grandes números, continuidade das fdp's) a estimativa MLE é consistente. (Ver prova em [3].)

A ideia da MLE foi já usada por Karl Gauss. Suponhamos os erros de *n* medições i.i.d.: $\Delta_i = X_i - \mu$. Seja a "lei combinada dos erros":

$$L(\mu) = \prod_{i=1}^{n} f(X_i - \mu)$$

O valor mais provável para μ é aquele que maximiza $L(\mu)$, ou seja, aplicando o logaritmo, corresponde a:

$$\sum_{i=1}^{n} \frac{d \ln f(X_i - \mu)}{d\mu} = 0$$

Por outro lado, considerando que a média aritmética das medições é o valor mais provável, temos:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{ou seja} \qquad \sum_{i=1}^{n}(X_i - \mu) = 0$$

Tomemos, então:

$$\frac{d \ln f(X_i - \mu)}{d\mu} = k(X_i - \mu)$$

Resulta que:

$$f(\Delta) \propto e^{-\frac{k}{2}\Delta^2} \quad \Rightarrow \quad f(\Delta) = \frac{h}{\sqrt{\pi}} e^{-h^2\Delta^2}$$

Exemplo 2-1

- $\overline{X}_n = \sum_{i=1}^{n} X_i / n$ é a estimativa (não-enviezada) MLE de μ. para $X_i \sim N(\mu,1)$
- Suponhamos que $X_1,...,X_n$ i.i.d. $N(\mu,\sigma)$ com μ conhecido e σ desconhecido. A MLE de σ^2 é $u_n^2 = \sum_{i=1}^{n}(X_i - \overline{X}_n)^2 / n$; trata-se de um estimador assimptoticamente não enviezado e consistente de σ.

(Ver prova em [3].)

Exemplo 2-2

Seja a situação do Exemplo 1-3. Para uma experiência em que o sucesso ocorre para X_n, a verosimilhança calcula-se como:

$$L(p) = P(X_1 = 0 \mid p).P(X_2 = 0 \mid p)...P(X_n = 1 \mid p) = q^{n-1}p \quad \dagger$$

Ora:

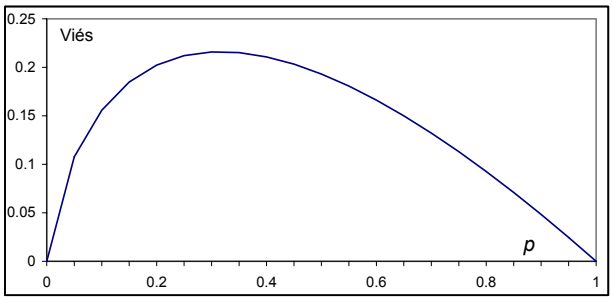
$$\frac{dL(p)}{dp} = q^{n-1} - (n-1)q^{n-2}p$$

O máximo corresponde a $q = (n-1)p$ ou seja $\hat{p} = 1/n$. A estimação MLE significa, portanto: se foram necessárias n tentativas para obter o 1º sucesso a estimativa de probabilidade é $1/n$. Esta estimativa é enviezada:

$$\mathrm{E}[\hat{p}] = \sum_{k=1}^{\infty} \frac{1}{k} q^{k-1} p = p\sum_{k=1}^{\infty} \frac{1}{k} q^{k-1} = -\frac{p}{q}\ln(1-q) = \frac{-p\ln p}{q}$$

† Notar que P(sucesso em $k \mid p$), k = 0,1,…, é a distribuição geométrica de probabilidade. Mas, como já vimos, $L(p)$ não é uma fdp.

O viés é, portanto: $b(p) = \dfrac{-p \ln p}{q} - p$ (figura abaixo)



Exemplo 2-3

Seja x uma quantidade desconhecida e suponhamos que medimos a quantidade Z, relacionada com x da seguinte forma:

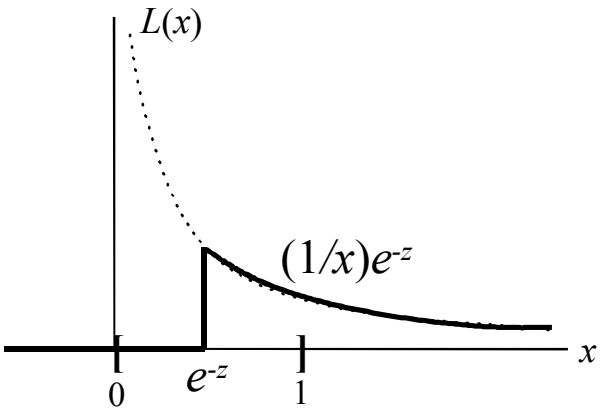
$$Z = \ln\left(\frac{1}{x}\right) + V$$

onde V é ruído de medição com distribuição exponencial:

$$f_V(v) = \begin{cases} e^{-v} & v \geq 0 \\ 0 & v < 0 \end{cases}$$

A função de verosimilhança é:

$$L(x) = f_{Z|x}(z \mid x) = f_V\left(z - \ln\left(\frac{1}{x}\right)\right) = \begin{cases} e^{-\left(z - \ln(1/x)\right)} & z - \ln\dfrac{1}{x} \geq 0 \\ 0 & z - \ln\dfrac{1}{x} < 0 \end{cases} = \begin{cases} \dfrac{1}{x} e^{-z} & x \geq e^{-z} \\ 0 & x < e^{-z} \end{cases}$$



O máximo de $L(x)$ é atingido para $x = e^{-z}$. Isto é, $\hat{x}_{MLE} = e^{-z}$.

3 Estimação de Mínimo Quadrático

Mean-Square Estimation/Error (MSE)
Least-Square Estimation (LSE)
Minimum-Mean-Squared-Error Estimation (MMSE)

A técnica dos mínimos quadráticos remonta a Abraham de Moivre e foi posteriormente masis desenvolvida e usada por Gauss.

Enquanto a MLE parte de uma ideia inteiramente probabilística a MSE parte de uma ideia métrica: minimizar uma distância.

A estimação MSE aplica-se a problemas de estimação em que existe uma variável desconhecida X e uma outra, conhecida, Z de alguma forma relacionada com X. Se as variáveis são aleatórias supõe-se conhecida a distribuição conjunta de (X, Z).

3.1 Fundamentação Geométrica

O espaço euclidiano, $\Re^n$, constituído por elementos $X = [x_1, x_2,\ldots, x_n]'$ designados por *vectores* e dotado da soma de vectores, do produto de um vector por um escalar e da norma de um vector, é um espaço linear normado no qual podemos definir a operação de *produto interno*: $X|Y = x_1y_1 + x_2 y_2 + \ldots + x_n y_n$.

Distância entre dois vectores: $\|X - Y\|$

Ângulo entre dois vectores: $\theta_{xy} = \arccos\left(\dfrac{X \mid Y}{\|X\|\|Y\|}\right)$

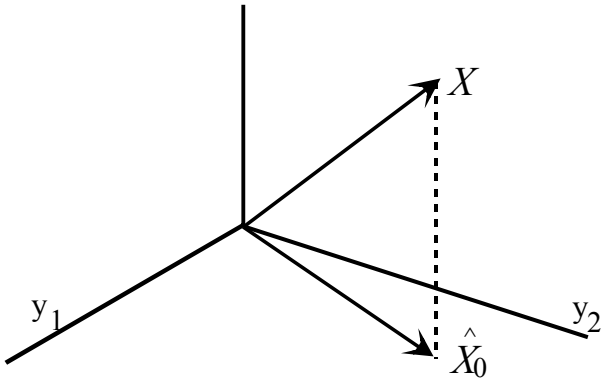
Desigualdade de Cauchy-Schwartz: $|X \mid Y| \le \|X\|\|Y\|$ (consequência óbvia de $|\cos\theta| \le 1$).

Quadrado da distância:

$$\|X - Y\|^2 = \|X\|^2 + \|Y\|^2 - 2\|X\|\,\|Y\|\cos\theta = \|X\|^2 + \|Y\|^2 - 2X \mid Y$$

O máximo do quadrado da distância (e, portanto, da distância) ocorre quando $X|Y = 0$: vectores ortogonais; o mínimo, quando os vectores são coincidentes.

Uma *variedade linear M* de $\Re^n$ é um sub-espaço linearmente fechado de $\Re^n$. Exemplo: sub-espaço que consiste em colocar a zero uma ou mais coordenadas.



Teorema 3-1

Seja $\hat{X}$ uma aproximação de um vector X de $\Re^n$ numa sua variedade linear M. Então, a *minimização da norma do desvio* $X - \hat{X}$, corresponde a:

$$\arg\left(\min_{\hat{X}\in M}\left\|X-\hat{X}\right\|\right)\stackrel{\Delta}{=}\hat{X}_0 \quad \Leftrightarrow \quad (X-\hat{X}_0)\,|\,\hat{X}=0,\ \forall\hat{X}\in M$$

■

Vejamos a solução para $\Re^2 \subset \Re^3$:

$\hat{X} = h_1 Y_1 + h_2 Y_2$; logo, $(X - h_{10}Y_1 - h_{20}Y_2)\,|\,(h_1 Y_1 - h_2 Y_2) = 0$

ou seja (dada a linearidade do produto interno):

$$h_1\left[X\,|\,Y_1 - h_{10}Y_1\,|\,Y_1 - h_{20}Y_2\,|\,Y_1\right] + h_2\left[X\,|\,Y_2 - h_{10}Y_1\,|\,Y_2 - h_{20}Y_2\,|\,Y_2\right] = 0$$

o que obriga ($\forall h_1, h_2$) a:

$$\begin{bmatrix} Y_1\,|\,Y_1 & Y_2\,|\,Y_1 \\ Y_1\,|\,Y_2 & Y_2\,|\,Y_2 \end{bmatrix}\begin{bmatrix} h_{10} \\ h_{20} \end{bmatrix} = \begin{bmatrix} X\,|\,Y_1 \\ X\,|\,Y_2 \end{bmatrix}$$

Seja $X = (a_1, a_2, a_3)$ e os vectores base de M: $Y_1 = (1,0,0)$ e $Y_2 = (0,1,0)$. Obtém-se:

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} h_{10} \\ h_{20} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$$ ou seja $\hat{X}_0 = (a_1, a_2, 0)$.

Para $\Re^m \subset \Re^n$ o sistema de equações é:

$$\sum_{j=1}^{m} h_{j0}(Y_j\,|\,Y_i) = X\,|\,Y_i \quad i = 1,2,...,m$$

ou, em notação matricial: $R_{YY}H_0 = R_{XY}$

onde:

- H_0 é o vector das incógnitas ($m\times1$).
- R_{YY} é matriz $m\times m$ dos produtos internos $Y_j \mid Y_i$. (Em geral, autocorrelações.)
- R_{XY} é o vector $m\times1$ dos produtos internos $X \mid Y_i$. (Em geral, correlações cruzadas.)

Designam-se estas equações por *equações normais* (normais, no sentido de envolver um princípio de ortogonalidade).

Existem várias situações particulares de espaços vectoriais onde são aplicáveis as equações normais do MSE (ver Apêndice). Em particular, os espaços vectoriais de funções de quadrado integrável designados por *espaços de Hilbert*, com:

norma: $\|X\| = \left[\int_T \alpha^2(t)dt\right]^{1/2}$ (corresponde à autocorrelação)

produto interno: $X \mid Y = \int_T \alpha(t)\beta(t)dt$ (corresponde à correlação cruzada)

3.2 MSE para funções de uma variável aleatória

Definição 3-1

Seja M um subespaço de L que consiste em todas as funções de uma v.a. X de média quadrática finita: $L = \left\{g(X);\quad \mathrm{E}[g^2(X)] < \infty\right\}$. A estimativa MMSE de $Y\in M$ é definida como a função $\hat{Y}$ tal que:

$$\min_{\hat{Y}=g(X)} \mathrm{E}\left[\left(Y-\hat{Y}\right)^2\right] \overset{\Delta}{=} \min_{\hat{Y}\in M} \left\|Y-\hat{Y}\right\|$$

com E[] = E_{XY}[].

■

Ora, L é um espaço de Hilbert (ver Apêndice); logo, é possível a interpretação geométrica habitual de um espaço vectorial. Assim, a melhor estimativa $\hat{Y}_0$ corresponde à projecção ortogonal de Y em M:

$$\mathrm{E}[(Y-\hat{Y}_0)\hat{Y}] = 0 \quad \forall \hat{Y} \in M$$

que se pode exprimir por $\mathrm{E}\left[(Y-g_0(X))g(X)\right] = 0 \quad \forall g(.)$,
onde $g_0(.)$ é a função de estimação óptima que buscamos e $g(.)$ é qualquer função de estimação admissível ($\mathrm{E}[g^2(X)] < \infty$).

Tentemos a solução:

$$\hat{Y}_0 = \mathrm{E}\left[Y \mid X\right], \qquad \text{ou seja,} \qquad g_0(.) = \mathrm{E}\left[Y \mid (.)\right]$$

Substituindo esta solução na condição (necessária e suficiente) anterior, obtemos:

$$\mathrm{E}\left[\left(Y - \mathrm{E}[Y \mid X]\right)g(X)\right] = 0 \quad \Rightarrow \quad \mathrm{E}\left[Yg(X)\right] = \mathrm{E}\left[\mathrm{E}[Y \mid X]g(X)\right] \quad \forall g(.)$$

Mas $\mathrm{E}[Y \mid X]g(X) = \mathrm{E}[Yg(X) \mid X], \forall X, Y, g(.)$. Logo:

$$\mathrm{E}\left[Yg(X)\right] = \mathrm{E}\left[\mathrm{E}[Yg(X) \mid X]\right]$$

o que é sempre verdadeiro (o valor esperado da média condicional é a média não condicional).

Logo a v.a. $\mathrm{E}[Y \mid X]$ é a MMSE (e projecção ortogonal em *M*): $\hat{Y}_{MSE} = \mathrm{E}[Y \mid X]$

Embora aparentemente simples, na prática pode ser intratável o cálculo de $\mathrm{E}[Y \mid X]$.

Notas:

1 - O princípio da ortogonalidade para v.a. pode exprimir-se assim:

$$\mathrm{E}\left[\left(Y - \mathrm{E}[\mathrm{Y} \mid \mathrm{X}]\right)g(X)\right] = 0 \quad \forall g(.)$$

Isto é, qualquer função de *X* é ortogonal a *Y* desde que subtraiamos a esperança condicional de *Y* dado *X*.

2 - Note-se que (ver Apêndice):

$$\mathrm{E}[\hat{Y}_{MSE}] = \mathrm{E}[\mathrm{E}[Y \mid X]] = \mathrm{E}[Y];$$

logo, a estimativa MSE é não enviezada.

3 - No caso de *g*(.) ser uma função linear - $\hat{Y} = hX$ -, obtém-se (ver Apêndice):

$$h_{MSE} = \frac{\mathrm{E}[XY]}{\mathrm{E}[X^2]}$$

Isto é, tal como no modelo determinístico, a solução depende da autocorrelação e da correlação cruzada.

4 - No caso de *g*(.) ser uma função linear com termo independente - $\hat{Y} = aX + b$ -, obtém-se (ver [4]):

$$a_{MSE} = \frac{\mathrm{V}[YX]}{\mathrm{V}[X]} \qquad \text{e} \qquad b_{MSE} = \mathrm{E}[Y] - \mathrm{E}[X]\,a_{MSE}$$

com $\mathrm{V}[YX] = \mathrm{E}[(YX - \mathrm{E}[YX])^2]$ e $\mathrm{V}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$.

Exemplo 3-1

Consideremos a situação do Exemplo 2.3 mas, agora, X é uma v.a. uniformemente distribuída em [0, 1]. Tal como anteriormente, temos:

$$f_{Z|X}(z \mid x) = \begin{cases} \frac{1}{x} e^{-z} & x \geq e^{-z} \\ 0 & x < e^{-z} \end{cases}$$

(A única alteração é que agora, sendo X uma v.a., temos $f_{Z|X}$ e não $f_{Z|x}$.)

Temos de calcular $\mathrm{E}[X \mid Z] = \int x f_{X|Z}(x \mid z) dx$ e só dispomos de $f_{Z|X}$.

Mas $f_{X|Z} = \frac{f_{XZ}}{f_Z}$ e $f_{Z|X} = \frac{f_{XZ}}{f_X}$; logo:

$$\mathrm{E}[X \mid Z] = \int x f_{X|Z}(x \mid z) dx = \int x \frac{f_{XZ}(x,z)}{f_Z(z)} dx = \frac{\int x f_{XZ}(x,z) dx}{\int f_{XZ}(x,z) dx} = \frac{\int x f_{Z|X}(z \mid x) f_X(x) dx}{\int f_{Z|X}(z \mid x) f_X(x) dx}$$

Portanto:

$$\hat{X}_{MSE} = \frac{\int_{e^{-z}}^{1} e^{-z} dx}{\int_{e^{-z}}^{1} \frac{1}{x} e^{-z} dx} = \frac{1 - e^{-Z}}{Z}$$

Exemplo 3-2

A mesma situação do exemplo anterior mas, agora, queremos uma MMSE linear $\hat{X} = aZ + b$. Temos:

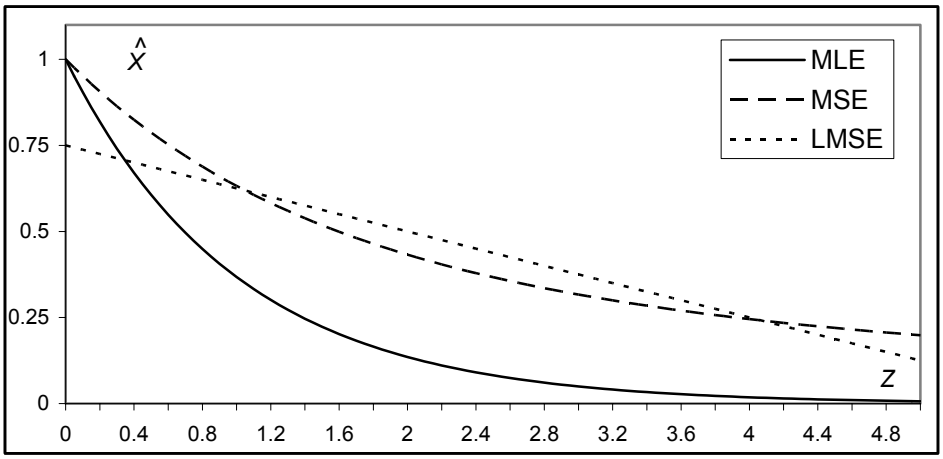
$$f_{XZ}(x,z) = f_{Z|X}(z \mid x) f_X(x) = \begin{cases} \frac{1}{x} e^{-z} & e^{-z} \leq x \leq 1; \quad z \geq 0 \\ 0 & \text{outros casos} \end{cases}$$

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{XZ}(x,z) dx = \begin{cases} z e^{-z} & z \geq 0 \\ 0 & z < 0 \end{cases}$$

Logo:

E[X] = ½
E[Z] = 2
E[XZ] = ¾
V[XZ] = ¼
E[Z^2] = 6
V[Z] = 2

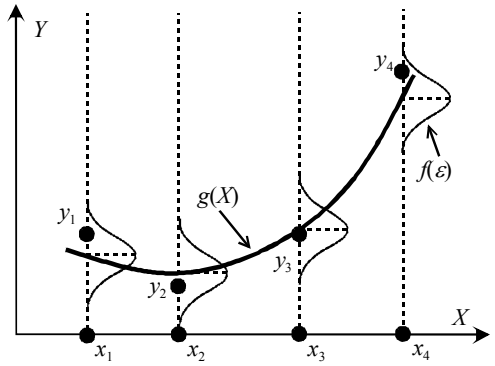
Portanto: $\hat{X}_{LMSE} = (3/4) - (1/8)Z$



3.3 MSE no modelo de regressão

Temos uma v.a. Y que depende duma variável preditora X com ruído ε (ver figura abaixo):

$$Y = g(X) + \varepsilon.$$



A MMSE é:

$$\mathrm{E}[Y \mid X] = \mathrm{E}[(g(X) + \varepsilon) \mid X] = \mathrm{E}[g(X) \mid X] + \mathrm{E}[\varepsilon \mid X]$$

No caso habitual de X ser não aleatória e ε ser de média nula, obtemos:

$$\mathrm{E}[Y \mid X] = g(X)$$

3.4 Situação de equivalência entre MSE e MLE

Consideremos, como nas redes neuronais (NN), n v.a. $Y_1,...,Y_n.$, que representam valores desejados obtidos por uma aproximação funcional $g(.)$ a partir de variáveis preditoras X_i. Concretamente, seja:

$$Y_i = g(X_i; w) + \varepsilon_i \,,$$

onde w representa o vector dos pesos e os ε_i os erros cometidos pela aproximação (rede). O objectivo é estimar w que "explica" Y_i. Temos a função de verosimilhança:

$$L(w) = \prod_{i=1}^{n} f(Y_i \mid w) = \prod_{i=1}^{n} f(g(X_i; w) + \varepsilon_i \mid w)$$

Para dados valores de X_i, os $g(X_i; w)$ são termos determinísticos, logo:

$$L(w) = \prod_{i=1}^{n} f(\varepsilon_i \mid w)$$

Suponhamos que:

1. Os erros ε_i são i.i.d. e normais, $N(\mu, \sigma)$.
2. Os parâmetros μ e σ são independentes de X_i

Então, podemos escrever:

$$L(w) = \prod_{i=1}^{n} f(\varepsilon_i \mid w) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left[-(Y_i - g(X_i; w))^2 / \left(2\sigma^2\right)\right]$$

Logo;

$$\ln L(w) = \ln\left[\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\right] - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - g(X_i; w))^2$$

Nestas condições maximizar $L(w)$ corresponde a minimizar a soma dos erros quadráticos, i.e., a MLE é a MMSE.

4 Estimação Bayesiana

A ideia-chave da estimação Bayesiana é a minimização da esperança matemática de uma função de custo.

4.1 Risco e função de custo

Suponhamos que, no modelo $Y_i = g(X_i; w) + \varepsilon_i$ queríamos estimar w. Podemos fazê-lo usando uma *função de custo* que avalie o ajuste de $g(X; w)$ a Y:

$$Q(y, g(x; w))$$

e procurando minimizar a média da função de custo, designada por *risco* (de Bayes):

$$R(w) = \int Q(y, g(x;w)) dF(y,x), \quad w \in W$$

Normalmente designamos $R(w)$ por erro, $E(w)$.
A MSE é um caso particular de estimação Bayesiana. Corresponde a usar a função de custo:

$$Q(y, g(x;w)) = (y - g(x;w))^2$$

4.2 Erro de Minkowski

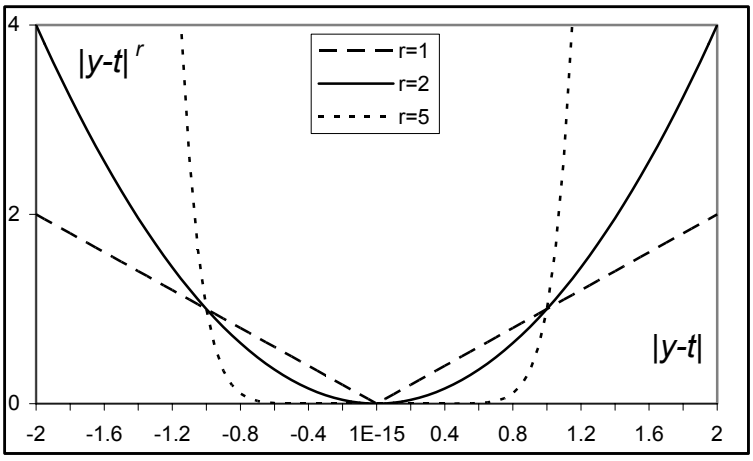
A distância de Minkowski entre duas funções $x(t)$ e $y(t)$ é definida por:

$$\rho(x,y) = \left\{ \int_a^b | x(t) - y(t) |^p \, dt \right\}^{1/p}$$

$p = 2$: Distância euclidiana
$p = 1$: Distância city-block
$p = \infty$: Distância de Tchebichef

Por analogia, o erro de Minkowski corresponde a:

$$E(w) = \int \left| y - g(x;w) \right|^r dF(y,x)$$



A MSE corresponde a tomar $r = 2$.†

† O erro de Minkowski também pode ser usado na retropropagação actualizando os pesos com:

$$\frac{\partial E}{\partial w_{ji}} = \sum_n \sum_k \left| y_k(x_n;w) - t_{nk} \right|^{r-1} sign(y_k(x_n;w) - t_{nk}) \frac{\partial y_k}{\partial w_{ji}}$$

O erro de Minkowski corresponde a considerar na estimativa MLE a seguinte pdf dos erros (generalização da gaussiana):

$$p(\varepsilon) = \frac{r\beta^{1/r}}{2\Gamma(1/r)} e^{-\beta|\varepsilon|^r}$$

É instrutivo determinar as soluções para $r = 2$ e $r = 1$, supondo a existência de fdp.

1) $r = 2$:

$$R(w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x;w))^2\, dF(y,x) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x;w))^2 f(y,x)dydx\,;$$

Mas:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(y - g(x;w))^2 f(y,x)dydx = \int_{-\infty}^{\infty} f(x)\int_{-\infty}^{\infty}(y - g(x;w))^2 f(y \mid x)dydx$$

O integrando em y é não-negativo; logo, minimizar $R(w)$ corresponde a minimizar:

$$R^*(w) = \int_{-\infty}^{\infty}(y - g(x;w))^2 f(y \mid x)dy = \mathrm{E}\left[y^2 \mid x\right] - 2g(x;w)\mathrm{E}\left[y \mid x\right] + g(x;w)^2$$

Logo:

$$\frac{\partial R^*(w)}{\partial w} = 0 \quad \Rightarrow \quad -2\mathrm{E}\left[y \mid x\right] + 2g(x;w) = 0 \quad \Rightarrow$$

$$g(x, w_0) = \mathrm{E}\left[y \mid x\right] = \int_{-\infty}^{\infty} y f(y \mid x)dydx$$

$\mathrm{E}\left[y \mid x\right]$ é a média condicional de y dado x.

2) $r = 1$:

$$Q(y, g(x;w)) = |\, y - g(x;w)\,|$$

Portanto:

$$R(w) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}|\, y - g(x;w)\,| f(y,x)dydx = \int_{-\infty}^{\infty} f(x)\int_{-\infty}^{\infty}|\, y - g(x;w)\,| f(y \mid x)dydx$$

Temos que minimizar:

$$R^*(w) = \int_{-\infty}^{\infty} | y - g(x;w) | f(y \mid x) dy =$$

$$\int_{-\infty}^{g(x;w)} (g(x;w) - y) f(y \mid x) dy + \int_{g(x;w)}^{+\infty} (y - g(x;w)) f(y \mid x) dy$$

Designemos $g(x; w)$ por a. A expressão anterior desenvolve-se como:

$$R^*(w) = a\left[\int_{-\infty}^{a} f(y \mid x) dy - \int_{a}^{+\infty} f(y \mid x) dy\right] + \left[\int_{a}^{+\infty} y f(y \mid x) dy - \int_{-\infty}^{a} y f(y \mid x) dy\right]$$

Usando o teorema fundamental do cálculo:

$$\frac{\partial R^*(w)}{\partial w} = 0 \quad \Rightarrow$$

$$\left[\int_{-\infty}^{a} f(y \mid x) dy - \int_{a}^{+\infty} f(y \mid x) dy\right] + a\left[f(a \mid x) - f(a \mid x)\right] + \left[a f(a \mid x) - a f(a \mid x)\right] = 0$$

Logo:

$$\frac{\partial R^*(w)}{\partial w} = 0 \Rightarrow \left[\int_{-\infty}^{a} f(y \mid x) dy - \int_{a}^{+\infty} f(y \mid x) dy\right] = 0 \Rightarrow$$

$$w_0 = g(x; w_0) = \text{mediana}\left[f(y \mid x)\right]$$

Efeitos das métricas da função de custo:

Seja o conjunto de dados:

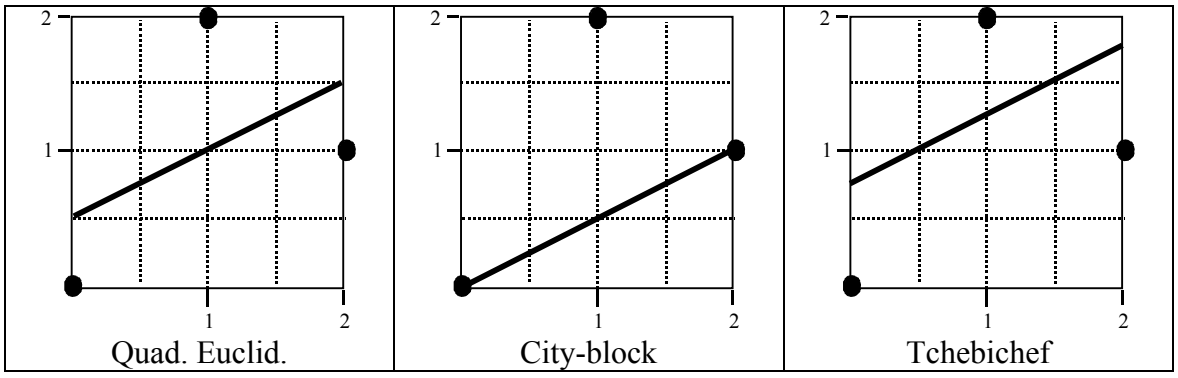
X_i	0	1	2
Y_i	0	2	1

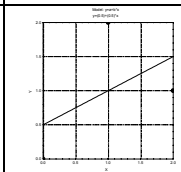
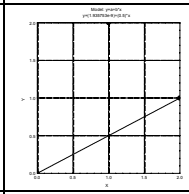
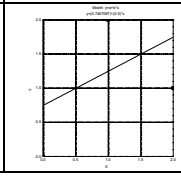
Queremos a regressão linear de Y_i dado X_i. Temos:

Quadrado dist. euclidiana (MSE): $\hat{Y} = (1 + X)/2$

City-block: $\hat{Y} = X/2$

Tchebichef: $\hat{Y} = (3 + 2X)/4$



Métrica		Erro Quad. Euclidiana	Erro City-block	Erro Tchebichef	Soma dos desvios
Quad. Euclid.		1.5	2.25	1.6875	0 (sempre)
City-block		2.0	1.5	2.25	1.5
Tchebichef		1.0	1.5	0.75	0.75

Vantagens da função de custo "soma dos quadrados dos desvios":

- Diferenciável
- Quadrado da distância Euclidiana
- Relacionada com grandezas físicas, e.g. energia
- Permite soluções (equações normais) com resultados interpretáveis e modelizáveis por distribuições estatísticas conhecidas.
- Equivalente, para fins de estimação, ao método da máxima verosimilhança

5 Interpretação de saídas de NN como probabilidades

Assunções:

1. NN com c saídas $y_k \in [0, 1]$ que pretendem aproximar $t_k \in \{0, 1\}$.

2. Distribuições dos erros para as *c* saídas são independentes:

$$E(x) = \sum_{k=1}^{c} f(y_k(x), t_k(x)) \qquad \text{para o padrão } x.$$

3. $f(y_k(x), t_k(x)) = f\left(\left|y_k(x) - t_k(x)\right|\right)$. Notar que:

$$t_k = \begin{cases} 0 & \Rightarrow \quad |\, y_k - t_k \,| = y_k \\ 1 & \Rightarrow \quad |\, y_k - t_k \,| = 1 - y_k \end{cases}$$

Temos, então, um erro por padrão dado por: $E = \sum_{k=1}^{c} f(|\, y_k - t_k \,|)$ (omitindo "(*x*)").

O erro médio é: $E = \iint E_n \, dP(\omega, x) = \int \{ \sum_{l=1}^{c} \sum_{k=1}^{c} f(|\, y_k - t_k \,|) P(\omega_l \mid x) \} p(x) dx$

Consideremos o integrando em *x* trocando a ordem dos somatórios:

$$\sum_{k=1}^{c} \sum_{l=1}^{c} f(|\, y_k - t_k \,|) P(\omega_l \mid x)$$

Para um dado *k* ($t_k = 1$) vejamos as possibilidades para *l*:

$$l \begin{cases} = k & \text{verifica – se a classe } \omega_k & \Rightarrow t_k = 1 \\ \neq k & \text{verifica – se uma classe} \neq \omega_k & \Rightarrow t_k = 0 \end{cases}$$

ou seja

$$l \begin{cases} = k & \Rightarrow & f(1 - y_k) P(\omega_k \mid x) \\ \neq k & \Rightarrow & \sum_{l \neq k} f(y_k) P(\omega_l \mid x) = f(y_k) \sum_{l \neq k} P(\omega_l \mid x) = f(y_k)(1 - P(\omega_k \mid x)) \end{cases}$$

Logo:

$$E = \sum_{k=1}^{c} \int \left\{ f(1 - y_k) P(\omega_k \mid x) + f(y_k)(1 - P(\omega_k \mid x)) \right\} p(x) dx$$

(Na referência [2] chega-se ao mesmo resultado por outra via (ver Apêndice).)

A condição a impor às saídas y_k por forma a obter o mínimo do erro médio por padrão corresponde a:

$$\frac{\partial E}{\partial y_k(x)} = f'(1 - y_k) P(\omega_k \mid x) + f'(y_k)(1 - P(\omega_k \mid x)) = 0$$

ou seja:

$$\frac{f'(1-y_k)}{f'(y_k)} = \frac{1-P(\omega_k \mid x)}{P(\omega_k \mid x)}$$

Para que as saídas possam representar probabilidades "a posteriori", deverá ser:

$$\frac{f'(1-y)}{f'(y)} = \frac{1-y}{y}$$

A classe de funções que satisfaz a esta condição é dada por:

$$f(y) = \int y^r (1-y)^{r-1} dy$$

$$r = \begin{cases} 0 & f(y) = -\ln(1-y) \quad cross-entropy \\ 1 & f(y) = y^2/2 \quad MSE \end{cases}$$

A função $f(y) = y^r$, correspondente ao erro de Minkowski, não satisfaz a condição acima para $r \neq 2$. Contudo, as fronteiras de decisão correspondem ao mínimo de $P(\omega_k \mid x)$.

6 Apêndice

6.1 Espaço métrico

Definição 6-1

Um *espaço métrico* é um par (X,ρ) de um conjunto X, cujos elementos se designam por pontos, e de uma função real e não negativa $\rho(x,y)$ definida $\forall x, y \in X$, designada por distância que satisfaz as seguintes condições:

1. $\rho(x,y) = 0$ sse $x = y$.
2. $\rho(x,y) = \rho(y,x)$ (simetria)
3. $\rho(x,y) + \rho(y,z) \geq \rho(x,z)$ (desigualdade triangular)

■

Exemplo 6-1

Seja X o conjunto das sequências de números reais $(x_1, x_2, \ldots, x_n, \ldots)$ que satisfazem a condição $\sum_{i=1}^{\infty} x_i^2 < \infty$ e seja:

$$\rho(x,y) = \left\{ \sum_{i=1}^{\infty} (y_i - x_i)^2 \right\}^{1/2}$$

Então (X,ρ) é um espaço métrico – métrica euclidiana - designado por l_2. (Ver prova em [1].)

■

Exemplo 6-2

Seja X o conjunto das funções contínuas no intervalo $[a, b]$ que satisfazem a condição $\int_a^b x^2(t)dt < \infty$ e seja:

$$\rho(x,y) = \left\{\int_a^b \left(x(t) - y(t)\right)^2 dt\right\}^{1/2}$$

Então (X,ρ) é um espaço métrico designado por $C^2[a, b]$. (Ver prova, baseada na desigualdade de Schwarz, em [1].)

■

Exemplo 6-3

Seja X o conjunto das funções contínuas no intervalo $[a, b]$ que satisfazem a condição $\int_a^b |x(t)|^p \, dt < \infty$ e seja:

$$\rho(x,y) = \left\{\int_a^b |\, x(t) - y(t)\,|^p \; dt\right\}^{1/p}$$

Então (X,ρ) é um espaço métrico designado por $C^p[a, b]$ (ver prova em [1]). A métrica é designada por métrica de Minkowski. Casos particulares da métrica de Minkowski:

$p = 1$: métrica quarteirão
$p = 2$: métrica euclidiana
$p = \infty$: métrica de Tchebichef ($\max(x(t), y(t))$)

■

6.2 Espaço linear

Definição 6-2

Um conjunto R de elementos x, y, z, ... é designado por *espaço linear* (ou *espaço vectorial*) se existirem duas operações, soma e multiplicação por um escalar α, β, ..., univocamente determinadas, com as seguintes propriedades: soma comutativa, associativa, com simétrico e elemento neutro; multiplicação associativa e com elemento neutro; soma distributiva relativamente ao produto e vice-versa.

■

Definição 6-3

Um espaço linear R diz-se *normado* se $\forall x \in R$ existe um número não negativo $||x||$, designado por norma de x, tal que:

1. $||x|| = 0$ sse $x = 0$
2. $|| \alpha x || = | \alpha | \, ||x||$
3. $||x + y|| \leq ||x|| + ||y||$

■

Todo o espaço normado é um espaço métrico, bastando definir $\rho(x,y) = ||x - y||$.

Exemplo 6-4

O espaço l_2 é um espaço linear normado se definirmos para $x = (x_1, x_2,\ldots, x_n,\ldots)$ e $y = (y_1, y_2,\ldots, y_n,\ldots)$:

1. $x + y = (x_1 + y_1, x_2 + y_2,\ldots, x_n + y_n,\ldots)$
2. $x = (\alpha x_1, \alpha x_2,\ldots, \alpha x_n,\ldots)$
3. $\|x\| = \left(\sum_{i=1}^{\infty} |x_i|^2\right)^{1/2}$

■

Caso particular de l_2: Espaço euclidiano de n-tuplos reais.

Exemplo 6-5

O espaço $C^2[a, b]$ é um espaço linear normado se definirmos:

$$\|f(t)\| = \left(\int_a^b f^2(t)dt\right)^{1/2}$$

■

Definição 6-4

Uma *variedade linear* ("linear manifold") L num espaço linear normado R é qualquer conjunto de elementos de R que satisfazem a condição: se $x, y \in L$ então $\alpha x + \beta y \in L$, para quaisquer números α, β.

■

6.3 Espaço de Hilbert de funções de variável real

Definição 6-5

Um espaço L de funções de variável real num certo intervalo T, de elementos $X = \{\alpha(t);\quad t \in T\}$, contendo todas as combinações lineares dos seus elementos, diz-se de *Hilbert* se as funções forem de quadrado integrável e estiverem definidas uma norma e um produto interno:

$$\|X\| = \left[\int_T \alpha^2(t)dt\right]^{1/2}$$
$$X \mid Y = \int_T \alpha(t)\beta(t)dt$$

■

Com esta definição as funções podem ser objecto de interpretação geométrica.

6.4 Espaço de Hilbert de variáveis aleatórias

Seja um *espaço de probabilidade* (S,B,P), onde S, conjunto de eventos é o *espaço amostral*; B é uma σ-álgebra de S ((S,B) designa-se por espaço mensurável); P é uma medida definida em B que satisfaz os axiomas de Kolmogorov.
Uma v.a. X é uma função de elementos de S:

$$X = \{\alpha(s);\quad s \in S\}$$

Definição 6-6

Uma v.a. diz-se de *média quadrática finita* se:

$$\mathrm{E}\left[X^2\right] = \int_S \alpha^2(s) dP(s) < \infty$$

■

Definição 6-7

A *correlação cruzada* de duas v.a. X e Y é o valor esperado do seu produto:

$$\mathrm{E}[XY] = \begin{cases} \sum_S \alpha(s)\beta(s)P(s) & \text{se discretas} \\ \int_S \alpha(s)\beta(s)dP(s) & \text{se contínuas} \end{cases}$$

■

E[XY] pode também ser expressa usando a função de distribuição F_{XY}. No caso contínuo, temos:

$$\mathrm{E}[XY] = \iint xy dF_{XY}(x,y)$$

Existindo fdp, podemos escrever: $\mathrm{E}[XY] = \iint xy f_{XY}(x,y) dxdy$

Definição 6-8

Um espaço L de v.a. num espaço amostral S, $X = \{\alpha(t);\ \ t \in T\}$, contendo todas as combinações lineares dos seus elementos, diz-se de *Hilbert* se as v.a. forem de média quadrática finita e estiverem definidas uma norma e um produto interno:

$$\|X\| = \left\{\mathrm{E}\left[X^2\right]\right\}^{1/2}$$
$$X \perp Y = \mathrm{E}[XY]$$

■

Com esta definição as v.a. podem ser objecto de interpretação geométrica. Por exemplo, duas v.a. são ortogonais se $X \perp Y = \mathrm{E}[XY] = 0$.

6.5 Esperança condicional

Dadas duas v.a. X e Y a esperança condicional de Y dado X é uma v.a. (e não um valor) definida como:

$$\mathrm{E}[Y \mid X] \stackrel{\Delta}{=} \int y f_{Y|X}(y) dy$$

Um valor da v.a. E[$Y \mid X$] é: $\mathrm{E}[Y \mid x] = \mathrm{E}[Y \mid X = x] = \int y f_{Y|X}(y \mid X = x) dy$

Propriedades da esperança condicional:

1 – A esperança da esperança condicional é a esperança incondicional: $\mathrm{E}[\mathrm{E}[Y \mid X]] = \mathrm{E}[Y]$.

De facto, dado que E[$Y \mid X$] é uma v.a. que depende de X, temos:

$$\mathrm{E}\big[\mathrm{E}[Y \mid X]\big] = \int \Big\{ \int y f_{Y|X}(y)dy \Big\} f_X(x)dx = \iint y f_{Y,X}(y,x)dydx = \mathrm{E}[Y]$$

Note-se, também, que: $\int f_{Y,X}(y,x)dx = f_Y(y)$

2 – Seja $g(.)$ uma função determinística da v.a. X. Então:

$$\mathrm{E}[Yg(X) \mid X = x] = \mathrm{E}[Y \mid X = x]\, g(x)$$

Esta propriedade de factorização exprime o facto de que $g(X)$ quando condicionada em $X=x$ é determinística.

3 – Seja $Z = g(X)$. Então:

$$\mathrm{E}\big[\mathrm{E}[Y \mid X] \mid Z\big] = \mathrm{E}[Y \mid Z]$$
$$\mathrm{E}\big[\mathrm{E}[Y \mid Z] \mid X\big] = \mathrm{E}[Y \mid Z]$$

Se g(.) tem inverso, verifica-se que: E[$Y \mid X$] = E[$Y \mid Z$]

6.6 Estimação linear MSE de séries temporais

Seja o problema de estimar a série temporal $X = [x_1, x_2, \ldots, x_n]'$ a partir de uma sua versão corrompida, com erros, $Y = [y_1, y_2, \ldots, y_n]'$ através de filtragem (convolução) com m pesos:

$$H_0 = [h_0, h_1, ..., h_{m-1}]'$$

A estimativa é: $\hat{x}_k = \sum_{i=0}^{m-1} h_i y_{k-i}, \quad k = 1,2,...,n$ ou $\hat{X} = \sum_{i=0}^{m-1} h_i Y_i$

Suponhamos que o critério é minimizar $\left\| X - \hat{X} \right\|^2$, ou seja, determinar $\min_H \left\| X - \hat{X} \right\|$. Aplicamos as equações normais com:

$[R_{XY}]_i = \sum_{k=1}^{n} x_k y_{k-i}$ correlação cruzada de X e Y

$[R_{YY}]_{ij} = \sum_{k=1}^{n} y_{k-j} y_{k-i}$ autocorrelação de Y

6.7 Estimação linear MSE de v.a.

Seja M o subespaço das funções lineares da v.a. X; i.e. M consiste em todos os $\hat{Y}$ da forma:

$$\hat{Y} = g(X) = hX$$

A condição de ortogonalidade corresponde agora a:

$$\mathrm{E}\left[(Y - h_0 X)hX\right] = 0 \quad \forall h \quad \Rightarrow \quad h\left\{\mathrm{E}\left[XY\right] - h_0 \mathrm{E}\left[X^2\right]\right\} = 0 \quad \Rightarrow \quad h_0 = \frac{\mathrm{E}\left[XY\right]}{\mathrm{E}\left[X^2\right]}$$

Por outro lado, da condição de ortogonalidade resulta que o erro médio quadrático é:

$$\mathrm{MSE}_0 = \mathrm{E}\left[\left(Y - \hat{Y}_0\right)^2\right] = \mathrm{E}\left[Y^2\right] - \mathrm{E}\left[\hat{Y}_0^{\,2}\right] = \mathrm{E}\left[Y^2\right] - \mathrm{E}\left[Y\hat{Y}_0\right] = \mathrm{E}\left[Y^2\right] - \frac{\left(\mathrm{E}\left[XY\right]\right)^2}{\mathrm{E}\left[Y^2\right]}$$

ou seja

$$\mathrm{MSE}_0 = \mathrm{E}[Y^2](1 - \rho^2)$$

com $\rho = \dfrac{\mathrm{E}\left[XY\right]}{\sqrt{\mathrm{E}\left[X^2\right]\mathrm{E}\left[Y^2\right]}}$, coeficiente de correlação (se X e Y têm média nula).

6.7.1 Situação n-dimensional

M é agora o subespaço de todas as funções lineares de n v.a. $\mathbf{X} = [X_1, X_2, ..., X_n]'$; i.e. M consiste em todas as v.a. da forma:

$$\hat{Y} = g(\mathbf{X}) = \sum_{i=1}^{n} h_i X_i = \mathbf{h}'\,\mathbf{X}$$

Aplicando a condição de ortogonalidade obtém-se:

$$\mathbf{h_0} = \mathbf{R}_\mathbf{X}^{-1}\mathbf{R}_{\mathbf{X}Y} \text{ com } \mathbf{R}_\mathbf{X} = \mathrm{E}\left[\mathbf{X}\mathbf{X}'\right] \text{ e } \mathbf{R}_{\mathbf{X}Y} = \mathrm{E}\left[\mathbf{X}Y'\right]$$

$$\mathrm{MSE}_0 = R_X(1 - \rho^2) \text{ e } \rho^2 = \mathbf{R}_{\mathbf{X}Y}'\,\mathbf{R}_\mathbf{X}^{-1}\mathbf{R}_{\mathbf{X}Y} / R_Y,\ R_Y = \mathrm{E}[Y^2]$$

6.7.2 Caso de dimensão infinita

Seja M o subespaço de todas as combinações lineares das v.a. $\{X(t); t \in V\}$ de um processo aleatório X:

$$\hat{Y} = \int_V h(t)X(t)dt$$

A solução para $\min_{\hat{Y} \in M} \mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ é dada pela condição de ortogonalidade:

$$\mathrm{E}[\left(X - \int_V h_0(u)X(u)du\right)X(t)] = 0 \quad \forall t \in V$$

Dada a linearidade de E[] podemos re-exprimir como:

$$\int_V R_X(t,u)h_o(u)du = R_{XY}(t) \quad \forall t \in V$$

que é uma expressão semelhante à obtida para a situação *n*-dimensional. Da mesma forma: $\mathrm{MSE}_0 = R_Y(1-\rho^2)$ e $\rho^2 = \frac{1}{R_Y}\int_V R_{XY}(t)h_0(t)dt$.

6.7.3 Estimação Linear de Ondas

Dada uma amostra $\{y(u);\ u \in V\}$ de um segmento de um processo aleatório $Y(u)$, desejamos estimar uma v.a. X relacionada com Y mas não observável. Usamos uma regra de estimação linear:

$$\hat{X} = \int_V h(u)Y(u)du \text{ e queremos } \min_{h(.)} E\left[\left(X - \hat{X}\right)^2\right]$$

Aplicando a condição de ortogonalidade, obtemos:

$$\int_V h_0(u)\mathrm{E}[Y(v)Y(u)]du = \mathrm{E}[XY(v)] \quad \forall v \in V \text{ (design equation) com}$$

$$\mathrm{MSE}_0 = \mathrm{E}[X^2] - \mathrm{E}[\hat{X}_0^2] = \mathrm{E}[X^2] - \int_V h_0(u)\mathrm{E}[XY(u)]du \qquad \text{(performance formula)}$$

Caso mais geral: estimar o valor de um processo aleatório $X(t)$ para $t \in T$, usando $\hat{X}(t) = \int_V h(t,u)Y(u)du$. Obtém-se:

$$\int_V h_0(t,u)\mathrm{E}[Y(v)Y(u)]du = \mathrm{E}[X(t)Y(v)] \quad \forall v \in V, t \in T \text{ (design equation) com}$$

$$\mathrm{MSE}_0 = \mathrm{E}[X^2(t)] - \int_V h_0(t,u)\mathrm{E}[X(t)Y(u)]du \quad \forall t \in T \qquad \text{(performance formula)}$$

Exemplos:

Seja a v.a. X uniformemente distribuída em [0,1]. Se escolhermos uma constante como estimativa a MMSE é:

$$\hat{X} = \int_{-\infty}^{\infty} x f_X(x)dx = 1/2$$

6.7.4 Teorema de Gauss-Markov

Dada uma estimativa linear $\hat{X} = \int_V h(u)Y(u)du$ a estimativa MMSE é a de menor variância.

6.8 Erro de Classificação

Na referência [2] a dedução do erro de classificação baseia-se na utilização de funções de Dirac. A vantagem é que tratamos todas as variáveis, incluindo a de classificação (etiqueta), como se fossem contínuas. O domínio de uma etiqueta t_k pasa a ser um domínio contínuo (p.ex. [0,1]) em vez de discreto ({0,1}). Desta forma, $dP(t_k,x)$ é representado como $p(t_k \mid x)p(x)dt_k\,dx$; i.e., como se estivesse definida uma fdp $p(t_k \mid x)$. (x é o vector das entradas.)

Vejamos a forma de operar com esta abordagem. Vimos que a estimativa MMSE de uma NN corresponde às saídas a aproximarem as médias condicionais. Supondo c saídas, y_k, k=1,…,c, temos:

$$y_k = \mathrm{E}[t_k \mid x] = \int t_k \, p(t_k \mid x) dt_k$$

Como definir $p(t_k \mid x)$? Usando o esquema de codificação dos valores desejados t_k com 1 se se tratar da classe ω_k e 0 no caso contrário ("1-out of-c"), buscamos uma função definida para $t \in \Re$ que produza $P(\omega_k \mid x)$ para $t_k = 1$ e $1-P(\omega_k \mid x)$ para $t_k = 0$, isto é:

$$\int p(t_k \mid x) dt = \begin{cases} P(\omega_k \mid x) & t_k = 1 \\ 1 - P(\omega_k \mid x) & t_k = 0 \end{cases}$$

Usando funções de Dirac, corresponde a:

$$p(t_k \mid x) = \begin{cases} \delta(t_k - 1) P(\omega_k \mid x) & t_k = 1 \\ \delta(t_k)(1 - P(\omega_k \mid x)) & t_k = 0 \end{cases}$$

ou seja,

$$p(t_k \mid x) = \delta(t_k - 1) P(\omega_k \mid x) + \sum_{\substack{l=1 \\ l \neq k}}^{c} \delta(t_k) P(\omega_l \mid x)$$

Usando o símbolo de Kronecker δ_{kl} (δ_{kl} =1 se $k = l$; δ_{kl} =0, $k \neq l$), reescrevemos como:

$$p(t_k \mid x) = \sum_{l=1}^{c} \delta(t_k - \delta_{kl}) P(\omega_l \mid x)$$

Temos, portanto:

$$y_k = \mathrm{E}[t_k \mid x] = \int t_k \{\delta(t_k - 1) P(\omega_k \mid x) + \sum_{\substack{l=1 \\ l \neq k}}^{c} \delta(t_k) P(\omega_l \mid x)\} dt_k =$$

$$\int t_k \delta(t_k - 1) P(\omega_k \mid x) dt_k + \sum_{\substack{l=1 \\ l \neq k}}^{c} \int t_k \delta(t_k) P(\omega_l \mid x) dt_k = P(\omega_k \mid x)$$

porque $\int \delta(x-a) f(x) dx = f(a)$.

Vejamos, agora, o cálculo do valor médio do erro como na secção 4.1:

$$\mathrm{E}[E] = \sum_{k=1}^{c} \int\int f(|y_k - t_k|) p(t \mid x)\, p(x) dt dx$$

Agora, *t* corresponde a *c* classes ω_k com distribuições independentes. Logo:

$$p(t \mid x) = \prod_{k=1}^{c} \left\{ \sum_{l=1}^{c} \delta(t_k - \delta_{kl}) P(\omega_l \mid x) \right\}$$

Substituindo, obtém-se:

$$\mathrm{E}[E] = \sum_{k=1}^{c} \int\int f(|y_k - t_k|) \prod_{k=1}^{c} \left\{ \sum_{l=1}^{c} \delta(t_k - \delta_{kl}) P(\omega_l \mid x) \right\} p(x) dt dx$$

Donde se deduz na referência [2] o resultado da secção 5. (Esta última expressão, além de uma notação complexa, não parece fácil de trabalhar porque mistura a expressão contínua dos $p(t_k \mid x)$ com a sua descrição discreta no 1º somatório. É fácil perder o sentido do que se está a fazer.)

6.9 Intervalos de confiança de estimativas

Os intervalos de confiança baseiam-se na utilização de estimativas de variâncias.
Para a MSE podemos usar estimativas de variâncias dos erros.
Para a MLE podemos usar o seguinte resultado:

Teorema 6-1 (Minorante de Crámer-Rao)

Se $\hat{X}$ é uma estimativa não-enviezada de uma variável *determinística* X baseada em medidas *Z*, a covariância de $\tilde{X} = X - \hat{X}$ é limitada por

$$V[\tilde{X}] \geq J_F^{-1},$$

onde J_F é a matriz de informação de Fisher dada por:

$$J_F = -\mathrm{E}\left[\frac{\partial^2 \ln f(z \mid X)}{\partial X^2} \right]$$

■

Para mais detalhes ver [4].

Referências

[1] A. Kolmogorov, S. Fomin (1957) Elements of the Theory of Functions and Functional Analysis. Dover Publications, Inc.
[2] Christopher Bishop (1995) Neural Networks for Pattern Recognition. Clarendon Press.
[3] Dudewicz EJ, Mishra SN (1988) Modern Mathematical Statistics. John Wiley & Sons, Inc.
[4] Frank Lewis (1986) Optimal Estimation. With an Introduction to Stochastic Control Theory. John Wiley & Sons.
[5] M. Priestley (1981) Spectral Analysis and Time Series. Academic Press.
[6] William Gardner (1990) Introduction to Random Processes. With Applications to Signals & Systems. McGraw Hill.