

## Comparación del desempeño

Clasificación correcta 85 %

Clasificación incorrecta 11 %

No clasificación 4 %

TOTAL 100 %

## Cap. 2. Reconocimiento estadístico de patrones.

Los patrones de la misma clase están sujetos a variaciones las cuales no necesariamente son causadas por un proceso aleatorio, sin embargo se recomienda tratarlos como si así fuera.

- La estadística y procesos estocásticos es una herramienta útil para describir variaciones.

Objetivo: Encantar un mapeo óptimo entre el espacio de medida y el espacio de decisión

- Problema de optimización estadística —

25/02/2019

Patrón :  $X \rightarrow$  vector aleatorio,  $W \rightarrow$  estados, clase

$W = [W_1, W_2, \dots, W_c] \rightarrow$  conjunto finito de estados

→ Maximización de la verosimilitud (ML).  
 → Estimación Bayesiana.

→ Aproximación de una curva polinomial, corresponde a un problema que puede ser visto como aproximación estadística ya sea mediante regresión ML o de Bayes.

## 2.1. Algunas nociones de probabilidad

Un concepto clave en el reconocimiento de patrones corresponde a la noción de incertidumbre. Dicha incertidumbre se presenta tanto en las mediciones, como ruido, y en el hecho de que los conjuntos de datos son de longitud finita.

La teoría de la probabilidad provee de un marco de trabajo consistente para la cuantificación y manipulación de la incertidumbre y forma parte central de las bases del reconocimiento de patrones

Sean dos variables aleatorias, una podría denotar la identidad o color de una caja que contiene frutas, y la otra podría denotar la identidad de las frutas contenidas. Sean entonces  $X = x_i$  e  $Y = y_j$ , en donde  $i = 1, 2, \dots, M$  y  $j = 1, 2, \dots, L$ .

- + La probabilidad de un evento es la fracción de veces que ocurre un evento con respecto del número total de intentos, pensando en que el total de intentos tiende a infinito.

Consideremos por el momento un total de  $N$  intentos para los cuales muestriaremos las dos variables  $X$  y  $Y$ , y sean el número total de intentos  $n_{ij}$ . También, sean el número de intentos en los que  $X$  toma los valores  $x_i$  dados por  $c_i$ , y de manera similar el número de intentos en los que  $X$  toma  $y_j$  denotadas por  $r_j$ .

Entonces, la probabilidad de que  $X$  tome los valores  $x_i$  se escribe como  $p(X = x_i)$  y la probabilidad de que  $Y$  tome los valores  $y_j$ ,  $p(Y = y_j)$ , y la probabilidad conjunta se escribe  $p(X = x_i, Y = y_j)$ .

Dicha probabilidad conjunta está dada por el número de puntos que caen en la celda  $i, j$  como una fracción del número total de puntos, entonces

$$P(X=x_i, Y=y_j) = \frac{n_{ij}}{N}.$$

Aquí se considera implicitamente que  $N \rightarrow \infty$ .

De manera similar, la probabilidad de que  $X$  tome los valores  $x_i$  de manera independiente a  $Y$  es

$$P(X=x_i) = \frac{c_i}{N}.$$

De lo anterior, tenemos que

$$c_i = \sum_j n_{ij},$$

y entonces podemos escribir

$$P(X=x_i) = \sum_{j=1}^L P(X=x_i, Y=y_j)$$

que corresponde a la regla de suma de probabilidades.

Nótese que  $p(X=x_i)$  también es conocida como probabilidad marginal, debido a que esta se obtiene marginalizando o sumando sobre las variables  $Y$ .

(\*) Si consideramos sólo los valores para los cuales

$X=x_i$ , están en dependencia de los  $Y=y_i$ , entonces la probabilidad para este caso es

$p(Y=y_j | X=x_i)$  la cual es denominada probabilidad

condicional de  $Y=y_j$  dado  $X=x_i$ . Esta se obtiene

determinando la función de esos puntos en la

columna  $i$  que caen en la celda  $ij$  y está dada

por

$$p(Y=y_j | X=x_i) = \frac{n_{ij}}{c_i}.$$

Y de las ecuaciones anteriores, se puede deducir que

$$p(X=x_i, Y=y_i) = \frac{n_{ii}}{N} = \frac{n_{ii}}{c_i} \cdot \frac{c_i}{N}$$

$$p(X=x_i, Y=y_i) = p(Y|X)p(X)$$

que corresponde a la regla de productos de probabilidades.

Entonces las dos reglas fundamentales de la teoría de la probabilidad son:

i) Regla de suma:  $p(x) = \sum_y p(x, y)$

ii) Regla del producto:  $p(x, y) = p(y|x) p(x)$ .

En donde  $p(x, y)$  es la probabilidad conjunta y se dice "la probabilidad de  $X$  y  $Y$ ", mientras que  $p(y|x)$  es la probabilidad condicional y se dice "la probabilidad de  $Y$  dado  $X$ ", mientras que  $p(x)$  es la probabilidad marginal o simplemente "la probabilidad de  $X$ ".

Estas dos reglas son parte fundamental del tratamiento estadístico de datos y se utilizan para la definición de la regla o teorema de Bayes, la cual se establece a partir de la regla del producto, y aprovechando su simetría, entonces

$$p(x, y) = p(y, x),$$

08

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}. \quad 29$$

Y utilizando ahora la regla de suma, podemos escribir

$$P(X) = \sum_Y P(X|Y)p(Y)$$

que corresponde al término del denominador de la ecuación de Bayes y puede ser visto como una constante de normalización

requerida para asegurar que la suma de probabilidades condicionales sobre los valores de  $Y$  tienden a un valor 1.

Densidades de probabilidad (v.a. continuas).

Al igual que cuando se consideran probabilidades definidas para conjuntos discretos de eventos,

también es posible considerar probabilidades con respecto de variables aleatorias continuas

Si la probabilidad de una variable  $X$  de valores reales cae dentro del intervalo  $(x, x + \delta x)$  y está dada por  $p(x)dx$  para  $\delta x \rightarrow 0$ , entonces  $p(x)$  es llamada

(29)

(30)

densidad de probabilidad sobre  $x$ . La probabilidad de que  $X$  caiga en el intervalo  $(a, b)$  está dada por la ecuación

$$P(X \in (a, b)) = \int_a^b p(x) dx.$$

Debido a que las probabilidades son no negativas, y dado que el valor de  $X$  debe caer en algún lugar del eje de los reales, la densidad de probabilidad  $p(x)$  debe satisfacer dos

condiciones:

$$a) \quad p(x) \geq 0, \text{ y}$$

$$b) \quad \int_{-\infty}^{\infty} p(x) dx = 1.$$

Bajo un cambio de variable no lineal, una densidad de probabilidad transforma de manera diferente a una simple función, debido al factor Jacobiano.

NOTA: Si  $X$  es una variable discreta, entonces  $p(x)$  es llamada algunas veces función masa de probabilidad dado que puede ser vista como un "conjunto de masas de probabilidad" concentradas alrededor de  $X$ .

También, para el caso de v.a. continuas aplican las reglas de probabilidad. Si  $x$  y son v.a. reales, entonces

I) Regla de suma:  $p(x) = \int p(x,y)dy$

II) Regla del producto:  $p(x,y) = p(y|x)p(x)$

Una justificación formal de ambas reglas para variables continuas requiere de fundamentos matemáticos de teoría de la probabilidad conocidos como teoría de medidas, en donde las medidas de Lebesgue son utilizadas.

Esperanzas y covarianzas.

Una de las operaciones más importantes que involucra a las probabilidades es aquella en

(31)

(32)

en la que se determinan ~~estadísticos~~ promedios de funciones ponderadas. El valor promedio de alguna función  $f(x)$  bajo una distribución de probabilidad  $p(x)$  es denominado esperanza de  $f(x)$  y será denotado como

$\mathbb{E}[f]$ . Para una distribución discreta, se escribe como:

$$\mathbb{E}[f] = \sum_x p(x)f(x),$$

y dado que el promedio está ponderado por probabilidades relativas de diferentes valores de  $x$ . En el caso de variables continuas, se escribe como

$$\mathbb{E}[f] = \int p(x)f(x) dx.$$

27/02/2019

En cualquier caso, si brindamos un número finito  $N$  de puntos muestreados de una distribución de probabilidad o de una densidad de probabilidad, entonces la esperanza puede

puede ser aproximada como una suma finita sobre dichos puntos, de modo que

$$\mathbb{E}[f] \approx \frac{1}{N} \sum_{n=1}^N f(x_n).$$

La aproximación se hace cada vez más exacta cuando  $N \rightarrow \infty$ .

Algunas veces, también es posible considerar esperanzas de funciones de varias variables, en tal caso se suele utilizar un sub-índice para indicar sobre cual variable se está promediando, entonces

$$\mathbb{E}_x [f(x, y)],$$

denota el promedio de la función  $f(x, y)$  con respecto a la distribución de  $x$ . Nótese que  $\mathbb{E}_x [f(x, y)]$  será una función de  $y$ .

También, se pueden considerar esperanzas condicionales con respecto a la distribución condicional, entonces

$$\mathbb{E}_x [f|y] = \sum_x p(x|y) f(x)$$

con una definición análoga para variables continuas.

Por otro lado, la varianza de  $f(x)$  está definida por

$$\text{var}[f] = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])^2]$$

la cual brinda una medida de que tan variable es la función  $f(x)$  al rededor de su valor promedio  $\mathbb{E}[f(x)]$ , también se puede escribir como

$$\text{var}[f] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

Para el caso de dos vectores de v.a.  $x$  e  $y$ , la covarianza se define como una matriz

$$\begin{aligned} \text{cov}[x, y] &= \mathbb{E}_{x,y} [(x - \mathbb{E}[x])(y^T - \mathbb{E}[y])] \\ &= \mathbb{E}_{x,y} [xy^T] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned}$$

Si se considera la covarianza de las componentes del vector  $x$  con ellas mismas, entonces se puede sólo escribir

$$\text{cov}[x] \equiv \text{cov}[x, x].$$

La desviación estandar es:  $\sigma = \sqrt{\text{var}[f]}$ ,

si la varianza es homocedástica.

- Proceso estocástico o aleatorio: Es aquel

proceso en el cual las componentes de un vector aleatorio  $X_t$  varían con el tiempo, es decir

$$X_t = [x_{1t}, x_{2t}, \dots, x_{nt}]^T.$$

- Proceso estacionario: Es aquel en el cual

la media y la varianza no son funciones del tiempo, y las fdp multidimensionales no dependen del tiempo origen, la correlación

sólo depende de la diferencia  $\tau = t_2 - t_1, t_3 - t_1, \dots, t_n - t_1$ . En este caso, la autocorrelación se define como

$$R_{xx}(t_1 - t_2) = R_{xx}(\tau) = \mathbb{E}[x_{t_1} \cdot x_{t_1 + \tau}],$$

que corresponde a un proceso estacionario en sentido amplio.

Un proceso estrictamente estacionario es

es aquel en el cual todos los momentos son constantes y no dependen del tiempo.

- Proceso ergódico: Es aquel en el cual los promedios estadísticos de un conjunto coinciden con los promedios temporales usando sólo una función muestra. Por ejemplo,

$$\bar{X} = \mathbb{E}[X_t] = \int_{-\infty}^{\infty} x p(x) dx,$$

$$= \lim_{P \rightarrow \infty} \frac{1}{P} \int_0^P X_t dt.$$

Puesto que los promedios temporales son independientes del tiempo, un proceso ergódico implica un proceso estacionario pero no a la inversa. Ahora podemos escribir la covarianza en términos de la correlación:

$$\text{cov}[xx(\tau)] = R_{xx}(\tau) - \bar{x}^2,$$

$$\text{cov}[xy(\tau)] = R_{xy}(\tau) - \bar{x}\bar{y}.$$

## • Teoría de la información

Además de los conceptos de teoría de la probabilidad y de la decisión, es necesario introducir otros conceptos relacionados que vienen del campo de teoría de la información.

Sigamos considerando que  $X$  es una v.a. discreta y ahora nos preguntamos ¿cuánta información es recibida? cuando observamos un valor específico para dicha variable. La cantidad de información puede ser vista como "el grado de sorpresa" cuando se aprende a partir del valor de  $X$ . Si decimos que es altamente improbable que ocurra un evento, y ocurre dicho evento, entonces hemos recibido más información que si dijéramos que un evento muy probable ha ocurrido, y si supieramos que el evento sucederá no recibimos nada nuevo de información. Nuestra medida de información ~~en función de~~ dependrá de la distribución de probabilidad  $p(x)$ , y entonces buscaremos la cantidad  $h(x)$  que es una función monótona de la probabilidad  $p(x)$  y que expresa el contenido de información en  $x$ .

La forma de  $h(x)$  puede ser determinada dependiendo de la relación con los eventos. Si tenemos dos eventos  $x$  e  $y$  no relacionados, entonces la ganancia de información al observar ambos eventos será la suma de la información ganada por cada uno de los eventos por separado, esto es

$$h(x, y) = h(x) + h(y).$$

Dos eventos no relacionados serán estadísticamente independientes y entonces

$$p(x, y) = p(x)p(y).$$

A partir de estas dos relaciones, es fácil demostrar que  $h(x)$  debe ser obtenida a partir del logaritmo de  $p(x)$  y entonces tendremos que

$$h(x) = -\log_2 p(x) \quad (1.42)$$

en donde el signo negativo asegura que la información es positiva o cero. Nótese que eventos de  $x$  de baja probabilidad corresponden

(39)

(4)

a un contenido alto de información. La elección de la base para el logaritmo es arbitraria, aunque por el momento se adoptará la convención que prebalece en teoría de la información usando logaritmos de base 2. Dado que  $h(x)$  se expresa en unidades binarias o bits (0 y 1).

Ahora, suponga que un envíador desea transmitir el valor de una v.a. a un receptor. La cantidad promedio de información que este transmitirá en el proceso se obtiene tomando la esperanza sobre  $h(x)$  con respecto de la distribución  $p(x)$ , y entonces tendremos que

$$H(x) = - \sum_x p(x) \log_2 p(x). \quad (1.43)$$

Esta cantidad importante es conocida como "Entropía" de la v.a.  $x$ . Nótese que

$$\lim_{p \rightarrow 0} p \ln p = 0 \quad y$$

entonces podemos tomar

$$= \sum_x p(x) \ln p(x) = 0$$

siempre que encontramos un valor para  $x$  tal que  $p(x) = 0$ .

(40)

(P8)

La relación entre entropía y el teorema de codificación sin ruido de Shannon (1948)

establecen que la entropía corresponde al límite inferior del número de bits necesarios para transmitir el estado de una v.a.

En adelante, podemos entonces substituir el  $\log_2$  por  $\ln$ , es decir, podemos escribir la entropía de una v.a.  $X$ , en donde

$$p(X=x_i) = p_i, \text{ como}$$

$$H[p] = - \sum_x p(x_i) \ln p(x_i).$$

La definición de entropía para v.a. discretas, puede ser extendida para v.a. continuas, bajo la definición de que

$$p(x) = q \int_{i\Delta}^{(i+1)\Delta} p(x) dx = p(x_i) \Delta,$$

en donde  $\Delta$  es el ancho de los bins de  $p(x_i)$ , entonces

$$H[p] = \lim_{\Delta \rightarrow 0} \left\{ \sum_x p(x_i) \ln p(x_i) \right\} =$$

(41)

(54)

$$x \text{ de los sistemas} \quad - \int p(x) \ln p(x) dx, \quad (1.103)$$

en donde la parte derecha de la ecuación, basada en la integral es conocida como entropía diferencial. Entonces podemos escribir  $H + [x]H = [p, x]H$

$$H[x] = - \int p(x) \ln(p(x)) dx,$$

en donde  $x$  es un vector de valores alrededor de  $x$ .

### • Entropía condicional.

Suponga que tenemos la distribución conjunta

$p(x, y)$ , a partir de la cual "tiramos" un

par de valores  $x$  e  $y$ . Si un valor de  $x$  ya

se conoce, entonces la información adicional

necesaria para especificar el valor correspondiente

de  $y$  estará dada por  $-\ln p(y|x)$ .

Entonces, el promedio de información adicional

necesaria para especificar  $y$  puede escribirse como

$$H[y|x] = - \iint p(x, y) \ln p(y|x) dy dx$$

42

la cual es llamada entropía condicional de  $y$  dado  $x$ . Es fácil ver que utilizando la regla del producto, la entropía condicional satisface la siguiente relación

$$H[x, y] = H[y|x] + H[x],$$

en donde  $H[x, y]$  es la entropía diferencial de  $p(x, y)$  y  $H[x]$  es la entropía diferencial de la distribución marginal  $p(x)$ .

o Entropía relativa e información mutua.

Si ahora relacionamos las ideas de entropía con las de reconocimiento de patrones, tomando en cuenta que la distribución  $p(x)$  es una incógnita, y suponemos que hemos modelado dicha distribución de manera aproximada utilizando la distribución  $q(x)$ . Si utilizamos  $q(x)$  para construir un esquema de codificación con el propósito de transmisión de valores de  $x$  a un receptor, entonces la cantidad promedio adicional de información requerida para especificar el valor de  $x$  como resultado de utilizar  $q(x)$

2/3

HA

en lugar de la verdadera distribución  $p(x)$ ,  
estará dada por la  $q(x)$ .

$$\text{KL}(p||q) = - \int p(x) \ln q(x) dx - \left( - \int p(x) \ln p(x) dx \right)$$

$$\text{KL}(p||q) = - \int p(x) \ln q(x) dx - \left( - \int p(x) \ln(p(x)) dx \right)$$

$$= - \int p(x) \ln \left\{ \frac{q(x)}{p(x)} \right\} dx.$$

esta expresión es conocida como entropía  
relativa o divergencia de Kullback-Leibler,  
o simplemente KL divergence. (1951), entre  
las distribuciones  $p(x)$  y  $q(x)$ .

Nótese que no es una cantidad simétrica:

$$\text{KL}(p||q) \neq \text{KL}(q||p),$$

$$\text{y } \text{KL}(p||q) \geq 0,$$

con igualdad a cero si y sólo si  $p(x) = q(x)$ .

\* Funciones convexas.

44

EP

Si tenemos dos conjuntos de v.a.  $x$  e  $y$ , dados por  $p(x, y)$  y si no son independientes, y considerando la definición de  $KL$ , entonces es posible hablar de una idea de "cercanía a la independencia" a partir de la distribución conjunta y el producto de marginales de cada

$$I[x, y] \equiv KL(p(x, y) \parallel p(x)p(y))$$

$$= - \iint p(x, y) \ln \left\{ \frac{p(x)p(y)}{p(x, y)} \right\} dx dy$$

A esta expresión se le conoce mejor como información mutua entre las variables  $x$  e  $y$ .

$$I(x, y) \geq 0,$$

$$I[x, y] = H[x] - H[x|y],$$

$$= H[y] - H[y|x].$$

que puede verse como la reducción de incertidumbre con respecto de  $x$  una vez que se ha dicho el valor de  $y$ .

45

## 2.2. Clasificación y regresión

Comenzamos con la introducción de un problema de regresión, a partir del cual estaremos desarrollando algunos de los conceptos importantes dentro del capítulo 2. Soponga que estamos observando una variable de entrada valuada en los reales  $x$  y deseamos utilizar esta observación para predecir el valor de una variable objetivo que también toma valores reales,  $t$ . Para los propósitos del curso, es ilustrativo tomar este ejemplo, desde un punto de vista sintético ya que podemos generar datos artificiales a partir de un modelo lo que nos permite conocer el proceso preciso que genera los datos con propósitos de comparar con respecto de algún otro modelo de aprendizaje.

Los datos para nuestro ejemplo, estarán generados a partir de una función senoidal  $\sin(2\pi x)$ , a dichos datos se les agregará ruido aleatorio conformando así los valores para  $t$ , de tal manera que

$$t = \sin(2\pi x) + e,$$

(216)

(217)

en donde  $e$  es ruido aleatorio, puede ser un vector, el cual se obtiene a partir de una distribución Gaussiana, centrada en cero y con varianza  $\sigma^2$  constante (por ejemplo,  $\sigma=0.3$ ).

Ahora, supongamos que contamos con un conjunto de entrenamiento, el cual está compuesto por  $N$  observaciones de  $x$ , o bien el vector  $x = (x_1, x_2, \dots, x_N)$  en conjunto con las observaciones de  $t$ , o el vector  $t = (t_1, t_2, \dots, t_N)^T$ . Dado que  $N$  corresponde a un entero finito, por ejemplo,  $N=10$ , entonces el vector  $x$  se puede formar por los elementos  $x_n$ , para  $n=1, 2, \dots, N$ , que se recomienda estén igualmente espaciados, por ejemplo, en el rango  $[0, 1]$ . Entonces, para cada valor  $x_n$ , se tendrá el siguiente modelo regresivo

$$t_n = \sin(2\pi x_n) + e_n.$$

Generando datos de esta manera, se captura una propiedad de muchos tipos de datos, los cuales poseen una cierta regularidad, a partir de la cual se desea "aprender", aunque las observaciones

417

8A

individuales están corrompidas por ruido abatorio que tiene ciertas características. Dicho ruido puede provenir de procesos intrínsecamente estocásticos (aleatorios) tales como un decaimiento radioactivo o bien que se deban a fuentes de variabilidad las cuales son no observables.

Nuestra meta es explotar este conjunto de entrenamiento con el propósito de hacer predicciones del valor  $\hat{t}$  de la variable objetivo para nuevos valores de  $\hat{x}$  que es la variable de entrada.

Como se verá más adelante, esto involucra implícitamente que tratamos de descubrir la función verdadera generadora sea  $(2\pi x)$ . Este es un problema intrínsecamente difícil dado que tenemos que generalizar a partir de una cantidad de datos finita. Más aún, los datos observados están contaminados con ruido, y por lo tanto para un  $\hat{x}$  dado existe incertidumbre al igual que para el valor de  $\hat{t}$ . La teoría de probabilidad nos provee de un marco de trabajo para expresar

la incertidumbre en forma precisa y cuantitativa, y por su parte la teoría de la decisión nos permite explotar la representación probabilista con el

(48)

FH

propósito de realizar predicciones que sean óptimas de acuerdo con los criterios o funciones de costo apropiados.

Por el momento, consideremos un problema sencillo de aproximación por curvas. En particular, podríamos estar interesados en ajustar los datos utilizando una función polinomial de la forma

$$y(x, w) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M$$

$$= \sum_{j=0}^M w_j x^j$$

en donde  $M$  es el orden del polinomio,  $x^j$  denota que la variable  $x$  está elevada a la potencia  $j$ -ésima, y  $w$  es el vector de coeficientes  $w_j$  del polinomio. Nótese

que aunque el polinomio  $y(x, w)$  es una función no lineal en  $x$ , ésta es una función lineal en los coeficientes  $w$ . Las funciones tales como los polinomios, que son lineales con respecto de los parámetros desconocidos tienen propiedades importantes y se denominan "modelos lineales" y podrá existir una forma cerrada para su solución.

49

65

Los valores de los coeficientes serán determinados adecuando el polinomio a los datos de entrenamiento. Esto se puede hacer "minimizando" una función de error que mide qué tan adecuados están los datos a la función polinomial  $y(x, w)$ , para algún valor dado de  $w$ . Una elección simple de función de error, que es ampliamente utilizada, está dada por la suma de los cuadrados entre las "predicciones"  $y(x, w)$  para cada dato  $x_n$  y su correspondiente valor  $t_n$ , por lo que se propone minimizar

$$e^2(w) = \frac{1}{2} \sum_{n=1}^N [y(x_n, w) - t_n]^2,$$

en donde el factor de  $1/2$  se incluye por conveniencia. De manera general, se establece una "función criterio" o "función de costo" sobre la función de error, que podemos escribir

$$J(e(w)) = e^2(w),$$

para el caso de aproximar o estimar  $w$ .

Para la estimación de  $w$  existen diferentes métodos propuestos en la literatura, uno de los más clásicos es el de mínimos cuadrados.

Siguiéndole, Máximo de Verosimilitud, y la Estimación Bayesiana.

### MÍNIMOS CUADRADOS (caso lineal).

El estimador de mínimos cuadrados (MC) cuando el modelo a estimar es de estructura lineal, como el polinomio  $y(x_n, w)$ , está dado por la optimización del criterio (forma matricial/vectorial)

$$J_{MC}(w) = e^T(w) Q e(w),$$

de manera que

$$\hat{w}_{MC} = \underset{w}{\operatorname{arg\min}} \{ J_{MC}(w) \},$$

dado que  $e(w) = t_n - y(x_n, w) = y(x_n, w) - t_n$ ,

si además, en forma vectorial  $e(w) = t - R w$ ,

entonces, minimizando el criterio, su derivada primera con respecto de  $w$  es nula. Si hacemos

la hipótesis de que  $Q$  es simétrica, entonces

51

52

el estimador se obtiene de la siguiente manera

$$\frac{\partial J_{MC}(W)}{\partial W} \Big|_{W = \hat{W}_{MC}} = -2R^T Q [t - R\hat{W}_{MC}] = 0$$

en donde, en la medida que  $[R^T Q R]^{-1}$  exista, entonces la solución convergerá a la siguiente forma

$$\hat{W}_{MC} = [R^T Q R]^{-1} R^T Q t.$$

Este estimador lineal, tiene las siguientes propiedades:

Propiedad 1: El vector de parámetros o coeficientes estimados  $\hat{W}_{MC}$  se obtiene de forma cerrada (o analítica) partiendo directamente de los datos  $\{t, x\}$  (No existe problema de inicialización).

Propiedad 2: Se puede verificar ampliamente que  $e^T(\hat{W}_{MC}) Q R \hat{W}_{MC} = 0$ . Si  $Q = \mathbb{I}$ ,

esto implica que los errores son perpendiculares a la salida del modelo,  $y(x_n, W)$  es entonces la proyección ortogonal de  $t$  sobre el espacio de respuestas de los posibles modelos (superficie

esperada) denotado por  $\mathcal{S}_r = \{y(x_n, w), w \in \mathbb{R}^{n_w}\}$ .

Como en este caso el modelo es lineal, el espacio

$\mathcal{S}_r$  es un hiperplano.

$(w)$   $\mathcal{S}$

Propiedad 3: La matriz a invertir  $[R^T Q R]$

es simétrica, de dimensiones  $n_w \times n_w$ . Las dimensiones no dependen del número de datos, sino del orden del modelo o del # de parámetros.

Propiedad 4: Aún si  $\hat{W}_{MC}$  no es la mejor estimación de  $W$ , de acuerdo a la meta del modelado, esta estimación podría servir como una inicialización en el caso de utilizar una metodología iterativa, aún y cuando se cambie de criterio.

Propiedad 5: La matriz inversa  $[R^T Q R]^{-1}$  contiene información importante a cerca del funcionamiento del estimador. Es de buena utilidad saber como fluctúan sus valores numéricos y sobre todo los de la diagonal principal. La precisión sobre

los parámetros estimados será cada vez mayor cuando los elementos de  $[R^T Q R]^{-1}$  son pequeños.