

53

53

Ejemplo: Sea la aproximación mediante un modelo polinomial, tal que

$$y(x_i, w) = w_0 + w_1 x_i + w_2 x_i^2 + \cdots + w_n x_i^n$$

en donde  $i = 1, 2, \dots, m$ . En este caso el estimador de mínimos cuadrados se obtiene de la siguiente manera, dado el conjunto de datos  $\{x_i, t_i\}_{i=1}^m$ , entonces

$$y(x_i, w) = \sum_{j=0}^n w_j x_i^j, \quad y e_i(w_j) = t_i - y(x_i, w)$$

por lo que el criterio quedaría de la forma

$$J_{MC}(w) = \sum_{i=1}^m [t_i - y(x_i, w_j)]^2,$$

$$J_{MC}(w) = \sum_{i=1}^m \left[ t_i - \sum_{j=0}^n w_j x_i^j \right]^2,$$

$$J_{MC}(w) = \sum_{i=1}^m t_i^2 - 2 \sum_{j=0}^n w_j \sum_{i=1}^m t_i x_i^j + \sum_j \sum_k w_j w_k \sum_{i=1}^m x_i^{j+k},$$

Luego, como  $\hat{w}_{MC} = \arg \min_{w_j} \{ J_{MC}(w) \}$ ,

entonces

$$\emptyset = \frac{\partial J_{MC}(w)}{\partial w_j} = \frac{\partial}{\partial w_j} \left[ \sum_{i=1}^m t_i^2 - 2 \sum_{j=0}^n w_j \sum_{i=1}^m t_i x_i^j + \sum_j \sum_k w_j w_k \sum_{i=1}^m x_i^{j+k} \right]$$

(54)

(55)

Se obtienen las ecuaciones normales

$$-2 \sum_{i=1}^m t_i x_i^j + 2 \sum_{k=0}^n w_k \sum_{i=1}^m x_i^{j+k} = 0,$$

y reacomodando las ecuaciones, tenemos que

$$\sum_{k=0}^n w_k \sum_{i=1}^m x_i^{j+k} = \sum_{i=1}^m t_i x_i^j,$$

el cual corresponde a un sistema de ecuaciones lineal dado por

$$j=0 : w_0 \sum_{i=1}^m x_i^0 + w_1 \sum_{i=1}^m x_i^1 + \dots + w_n \sum_{i=1}^m x_i^n = \sum_{i=1}^m t_i x_i^0$$

$$j=1 : w_0 \sum_{i=1}^m x_i^1 + w_1 \sum_{i=1}^m x_i^2 + \dots + w_n \sum_{i=1}^m x_i^{n+1} = \sum_{i=1}^m t_i x_i^1$$

$$\vdots \quad \vdots \quad \vdots \quad \vdots$$

$$j=n : w_0 \sum_{i=1}^m x_i^n + w_1 \sum_{i=1}^m x_i^{n+1} + \dots + w_n \sum_{i=1}^m x_i^{2n} = \sum_{i=1}^m t_i x_i^n$$

en donde se conocen los  $\{x_i, t_i\}_{i=1}^m$  que son los datos de entrenamiento, y entonces las incógnitas serán

$$\hat{w}_{\text{MC}} = [\hat{w}_0, \hat{w}_1, \dots, \hat{w}_n].$$

Entonces, tendremos la forma matricial

$$A \hat{w}_{\text{MC}} = b \rightarrow R w = b$$

que podemos resolver mediante métodos como la eliminación Gaussiana con pivoteo parcial

(55)

(56)

o el método de Gauss-Jordan, de tal manera que

$$y(x_i, \hat{W}_{MC}) = \hat{W}_{0MC} + \hat{W}_{1MC}x_i + \hat{W}_{2MC}x_i^2 + \dots + \hat{W}_{nMC}x_i^n$$

La metodología de mínimos cuadrados, puede ser utilizada para otros modelos lineales, como el ARX, AR, ARMA, ARMAX, etc. También puede ser extendida para modelos no lineales en los parámetros, aunque en este caso el sistema de ecuaciones resultante es no lineal y se requiere de métodos iterativos para la obtención de la solución, por ejemplo, el método de Newton-Raphson, métodos del gradiente, del descenso, etc.

También, el método de mínimos cuadrados se puede implementar de manera recursiva para el caso de estimación de parámetros en modelos LPC o estimación de estados en el caso de filtros de Kalman.

## 2.3. Máximo de Verosimilitud (ML)

En este caso diremos que  $\hat{W}_{ML}$  es un estimador en el sentido de máximo de verosimilitud (ML: Maximum of Likelihood), si este maximiza el siguiente criterio

$$J_{ML}(w) = \prod_t (t|w).$$

En donde  $\prod_t$  es una distribución de probabilidad que caracteriza la similitud o verosimilitud entre  $t_i$  y  $y(x_i, w)$ . Si  $w$  estuviese fijo, entonces  $\prod_t (t|w)$  es una densidad de probabilidad del vector aleatorio  $t_i$  asociado al modelo  $y(x_i, w)$ .

Este método consiste en buscar los valores de los parámetros  $w$  que atribuyen a los datos  $t_i$  la más grande verosimilitud. Se tiene siempre cuenta de la información disponible sobre la naturaleza del ruido que perturba al sistema de tal forma que se puede construir un criterio de calidad adaptado.

En la práctica, es más fácil buscar  $\hat{W}_{ML}$  maximizando la log-verosimilitud, de modo que redefinimos el criterio de la sig. manera

(57)

(82)

$$J_{ML}(w) = \ln \left[ \Pi_t (t|w) \right],$$

lo que conduce al mismo resultado que con el criterio inicial, ya que la función logarítmico es monótona creciente.

Ejemplo 1: Observaciones repetidas de una v.a. Gaussiana.

Consideremos un sistema en el cual observamos experimentalmente una salida escalar  $t(x_i)$  para  $i = 1, 2, \dots, n$ . Supongamos que fue posible repetir las observaciones a cada tiempo  $x_i$  para estimar las características de los ruidos de medición.

Bajo la hipótesis de resultados de medidas al instante  $x_i$  independientes, Gaussianas y con distribución idéntica, podemos estimar su media  $\mu_i$  y su varianza  $\sigma_i^2$ , en el sentido de máximo de verosimilitud (ML). El conjunto de resultados de medidas disponibles para el  $i$ -ésimo tiempo  $x_i$  nos da el vector

$$t(x_i) = [t_1(x_i), t_2(x_i), \dots, t_n(x_i)]^T,$$

en donde  $t_j(x_i)$  es el resultado de la medición en un tiempo  $x_i$  para el  $j$ -ésimo experimento.

Dado que se ha supuesto que  $t(x_i)$  sigue una ley normal  $N(\mu_i, \sigma_i^2)$ , la densidad de probabilidad de  $t_j(x_i)$  estará dada por

$$\Pi_t(t_j(x_i) | \mu_i, \sigma_i^2) = \Pi_t(t_j(x_i) | W)$$

entonces

$$\Pi_t(t_j(x_i) | W) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{t_j(x_i) - \mu_i}{\sigma_i}\right)^2\right]$$

en donde  $W = [\mu_i, \sigma_i^2]^T$ . Como las  $n$  observaciones para  $x_i$  se han supuesto independientes, entonces

$$\Pi_t(t(x_i) | W) = \prod_{j=1}^n \Pi_t(t_j(x_i) | W) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{1}{2}\left(\frac{t_j(x_i) - \mu_i}{\sigma_i}\right)^2\right]$$

$$\exp\left[-\frac{1}{2}\left(\frac{t_j(x_i) - \mu_i}{\sigma_i}\right)^2\right]$$

Y aplicando la log-verosimilitud, tenemos que el criterio se escribe como

$$J_{ML}(W) = \ln \left[ \Pi_t(t(x_i) | W) \right] = -\frac{n}{2} \ln(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2} \sum_{j=1}^n (t_j(x_i) - \mu_i)^2$$

(59)

(60)

Para estimar  $\hat{W}_{ML}$ , o sea  $\hat{\mu}_i$  y  $\hat{\sigma}_i^2$  minimizamos el criterio  $J_{ML}(w)$  con respecto de  $w$ . En general, no es posible resolver este problema de optimización de forma explícita, y se hace uso de algoritmos iterativos, aunque si podemos dejar indicadas las condiciones para cada uno de los estimadores

$$\frac{\partial J_{ML}(w)}{\partial \mu_i} = \frac{1}{\hat{\sigma}_{iML}^2} \sum_{j=1}^n (t_j(x_i) - \hat{\mu}_{iML}) = 0,$$

$$\frac{\partial J_{ML}(w)}{\partial \sigma_i^2} = -\frac{n}{2\hat{\sigma}_{iML}^2} + \frac{1}{2\hat{\sigma}_{iML}^4} \sum_{j=1}^n (t_j(x_i) - \hat{\mu}_{iML})^2$$

La primera de las dos ecuaciones implica lo siguiente

$$\hat{\mu}_{iML} = \frac{1}{n} \sum_{j=1}^n t_j(x_i).$$

El estimador de la media en el sentido ML, es la media aritmética de las observaciones. Por otro lado, la segunda ecuación implica que

$$\hat{\sigma}_{iML}^2 = \frac{1}{n} \sum_{j=1}^n (t_j(x_i) - \hat{\mu}_{iML})^2.$$

En general, el método de ML es la base de un gran número de técnicas de estimación y posee propiedades teóricas bastante atractivas.

Ejemplo 2º Variables aleatorias independientes y aditivas en la salida. Supongamos una señal de salida observada que sigue el siguiente modelo.

Ejemplo 2º Ruido Gaussiano escalar con varianza conocida o estacionaria. Sea el modelo

$$f(x_i) = y(x_i, w) + b_i, \quad b_i \sim N(0, \sigma_i^2).$$

Suponemos que la varianza  $\sigma_i^2$  del ruido  $b_i$  es constante e independiente de  $i$  y además es conocida. La pudimos haber estimado de manera previa utilizando el estimador del ejemplo 1. Si ahora deseamos estimar  $w$  de  $y(x_i, w)$  utilizando también el método ML. De acuerdo a las hipótesis sobre el ruido, tenemos que  $\pi(b_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{b_i}{\sigma}\right)^2\right]$ .

(61)

Entonces, la verosimilitud de las  $n$  observaciones de salida estará dada por

$$\prod_{i=1}^n \pi(t(x_i) | w) = \prod_{i=1}^n \frac{1}{(2\pi\sigma_i^2)^{1/2}} \exp\left[-\frac{1}{2} \left(\frac{t(x_i) - y(x_i, w)}{\sigma_i}\right)^2\right]$$

lo cual quiere decir que los  $t(x_i)$  son v. a. i. i. d. que siguen una ley normal

$$\mathcal{N}(y(x_i, w), \sigma_i^2).$$

La log-verosimilitud puede escribirse entonces como

$$\ln [\pi(t(x_i) | w)] = \text{término indep. } w - \frac{1}{2} \sum_{i=1}^n \left(\frac{t(x_i) - y(x_i, w)}{\sigma_i}\right)^2$$

El estimador de  $w$  en el sentido de máximo de

verosimilitud  $\hat{w}_{ML}(t)$  es entonces el argumento

máximo del siguiente criterio

$$J_{ML}(w) = -\frac{1}{2} \sum_{i=1}^n \left(\frac{t(x_i) - y(x_i, w)}{\sigma_i}\right)^2,$$

es decir, el argmin del criterio cuadrático, para este caso en particular el estimador ML coincide con el estimador MC.

$$J_{MC}(w) = \sum_{i=1}^n w_i [t(x_i) - y(x_i, w)]^2$$

en donde tendremos las ponderaciones  $w_i = \frac{1}{\sigma_i^2}$ .

Lo cual no da Mínimos Cuadrados ponderados.

El error en un tiempo  $x_i$  es ponderado por el inverso de la varianza del ruido asociado: se le da menor peso a una medida con mayor ruido.

Podemos notar que las v.a. dadas por

$$(t_i - y(x_i, w))^2 / \sigma_i^2$$

están identicamente distribuidas de acuerdo a  $N(\phi, 1)$ . En este caso, se puede hablar de igualación de varianzas. Este estimador

puede ser utilizado aunque el ruido no siga una Ley normal a condición de que la varianza sea un parámetro conocido. Podemos hablar

de un estimador Gauss-Markov para este caso anterior. Si tenemos un modelo de

estructura lineal con respecto de  $w$ , el mejor estimador lineal in-sesgado será obtenido (BLUE: Best Linear Unbiased Estimator).

(63)

(43)

Si todos los  $t_i^2$  son iguales, se puede utilizar el estimador de Mínimos Cuadrados no ponderados en donde  $w_i = 1$ .

Propiedades de los estimadores ML:

Propiedad 1: Estos estimadores son convergentes,

$$\forall \delta > 0, \lim_{n \rightarrow \infty} \Pr(\|\hat{w}_{ML} - w\| \leq \delta) \rightarrow 1$$

Se deben cumplir las siguientes hipótesis,

H1: Los datos son generados a partir de un modelo  $M(w^*) \rightarrow$  No existe error de caracterización.

H2: El modelo  $M(\cdot)$  es globalmente identificable de acuerdo a las condiciones experimentales consideradas.

H3: Las perturbaciones que interactúan sobre los datos pueden (ser) modeladas como v.a.i.i.d. eventualmente propagadas a través del modelo  $M(w^*)$ .

H4: El conjunto  $\{t(x_i) \mid \Pr(t(x_i) \mid w) > 0\}$  es independiente de  $w$ ; la log-verosimilitud  $\ln [\Pr(t(x_i) \mid w)]$  es dos veces derivable con

(64)

respecto de  $w$  y de derivadas segundas uniformemente continuas en  $t(x_i)$  para llevar al cabo desarrollos al rededor de un vecindario de  $w^*$ .

H5:

$$E_{t|w^*} \left\{ \frac{\partial \ln [\pi(t(x_i)|w)]}{\partial w_i} \right\} < \infty \text{ y}$$

$$E_{t|w^*} \left\{ \frac{\partial^2 \ln [\pi(t(x_i)|w)]}{\partial w_i \partial w_j} \right\} < \infty, \forall w.$$

Propiedad 2:  $\hat{w}_{ML}$  son asintóticamente eficaces, es decir que no existe un estimador convergente que tenga una covarianza cada vez más pequeña cuando  $n \rightarrow \infty$ .

Propiedad 3:  $\hat{w}_{ML}$  son asintóticamente normales e insesgadas, es decir que  $\hat{w}_{ML}$  tenderá a estar distribuido según la ley

$N(w^*, F^{-1}(w^*))$ , para  $n \rightarrow \infty$ , en donde  $F(w^*)$  es la matriz de información de Fisher, la cual está dada por

$$F(w^*) = E_{t|w^*} \left\{ \frac{\partial}{\partial w} \ln [\pi(t(x_i)|w)] \frac{\partial}{\partial w^T} \ln [\pi(t(x_i)|w)] \right\}$$

$w=w^*$

(65)

$$F(w^*) = -E_{t|w^*} \left\{ \frac{\partial^2}{\partial w \partial w^T} \ln [\pi(t(x_i)|w)] \right\}$$

en donde  $F(w)$  es positiva definida para todo  $w$ .

Propiedad 4: Si  $g(\cdot)$  es una función de  $w$

únicamente, por ejemplo corresponde a una reparametrización, entonces el estimador

de  $g$  en el sentido ML estará dado

por  $\hat{g}_{ML} = g(\hat{w}_{ML})$  por principio de invarianza.

Para calcular la estimación ML de toda cantidad que puede ser deducida a partir del conocimiento

de los parámetros  $\hat{w}_{ML}$ , no es necesario desarrollar un estimador en particular. Los  $\hat{w}_{ML}$  pueden

servir para calcular el estimador asociado a una nueva parametrización de la misma estructura de modelos como es el caso de  $g$ .

2.4. Estimadores Bayesianos.

La aproximación de Máximo de Verosimilitud considera a  $w$  como desconocido pero constante.

66

Mientras que las aproximaciones Bayesianas consideran que  $W$  es un vector aleatorio, de densidad de probabilidad "a priori" supuesta conocida y denotada por  $\Pi(W)$ . De acuerdo a la regla de Bayes y de probabilidades condicionales, la densidad de probabilidad conjunta de  $t(x_i)$  y  $W$  se puede escribir de la siguiente manera

$$\Pi(t(x_i), W) = \Pi(t(x_i) | W) \Pi(W) = \Pi(W | t(x_i)) \Pi(t),$$

Luego, la densidad de probabilidad de  $W$  "a posteriori" verifica la siguiente ecuación

$$\Pi(W | t(x_i)) = \frac{\Pi(t(x_i) | W) \Pi(W)}{\Pi(t(x_i))}.$$

Ésta es mejor conocida como la regla de Bayes, que brinda su nombre a esta categoría de estimadores y expresa cuantitativamente un proceso de aprendizaje. Como  $t(x_i)$  es el resultado numérico conocido, las medidas sobre el sistema a modelar,  $\Pi(t(x_i))$  consisten

(67)

(33)

de una constante de normalización destinada a asegurar que  $\Pi(t(x_i)|w)$  sea una densidad de probabilidad, tal que

$$\Pi(t(x_i)) = \int_P \Pi(t(x_i)|w) \Pi(w) dw.$$

Para calcular la densidad de probabilidad de  $w$  dados los datos  $t$ , es suficiente expresar  $\Pi(t(x_i)|w)$  sacando provecho de toda la información que tengamos disponible sobre el ruido  $b$ , también debemos disponer de  $\Pi(w)$  que traducirá nuestro conocimiento a priori sobre los parámetros. Este conocimiento puede provenir de mediciones efectuadas anteriormente sobre el mismo proceso o sobre procesos similares.

Las técnicas de máximo de entropía permiten seleccionar una distribución  $\Pi(w)$  coherente con un conocimiento a priori evitando introducir información parásita.

Dichas técnicas conducen a obtener como densidad (de probabilidad a priori) aquella distribución que maximiza la entropía, dada por

$$H(\pi(w)) = - \int_{\mathbb{R}} \pi(w) \ln [\pi(w)] dw,$$

donde la densidad óptima se obtiene mediante una técnica de multiplicadores de Lagrange.

Si por ejemplo, sólo conocemos el valor promedio  $\bar{w}$  y la varianza  $\Omega$  a priori de  $w$ , el principio de máxima de entropía nos permite obtener una densidad a priori normal  $N(\bar{w}, \Omega)$ . Si la

sólo tenemos información a priori sobre los parámetros ese que estos están en el siguiente rango

$$w_{\min} \leq w \leq w_{\max},$$

entonces, la densidad  $\pi(w)$  será uniforme sobre el conjunto definido  $[w_{\min}, w_{\max}]$  (a esta se le denomina distribución a priori no informativa).

La explotación de la densidad de probabilidad a posteriori  $\pi(w | t(x_i))$  no siempre es sencilla,

(69)

55

y seguido es preferible partiendo de una regla para deducir una estimación puntual de los parámetros  $\hat{W}$ .

**Comentario:** Cuando el soporte de la distribución a priori es discreto, la aplicación de la regla de Bayes es simple pues los puntos de soporte de la ley de probabilidad a posteriori coinciden con los puntos de la ley a priori. Sólo cambian los pesos asociados. El cálculo de las esperanzas matemáticas también se simplifica, pues las integrales se convierten en sumatorias finitas.

2.3.1. Criterio de Máximo A Posteriori (MAP).

La estimación en el sentido de máximo a posteriori consiste en la búsqueda de  $\hat{W}_{MAP}$  que maximice el criterio

$$J_{MAP}(W) = \ln [\pi(t|W)] + \ln [\pi(W)],$$

que se obtiene a partir de

$$J_{MAP}(W) = \pi(W|t) = \frac{\pi(t|W)\pi(W)}{\pi(t)},$$

70

como  $\pi(t)$  no depende de  $w$ , y se puede convertir en un valor constante unitario y además, dado que la función logaritmo es monótona, entonces se aplica al criterio Bayesiano. Así pues, se brinda un criterio de dos elementos a la derecha de la ecuación en donde el primero es la log-verosimilitud, y el segundo traduce la información a priori con la que contamos sobre  $w$ . En este caso, es posible integrar en el modelado alguna información acerca de los valores posibles de  $w$ . Prácticamente se mantienen las mismas hipótesis hechas para el estimador ML, y sólo agregamos la siguiente:

H6:  $\pi(w)$  es una función continua diferente de cero y se encuentra dentro del vecindario de  $w^*$ .

Esto significa que el estimador MAP comparte las mismas propiedades de convergencia y eficacia asintóticas del estimador ML. Además, también es invariante a re-parametrizaciones.

71

55

Ejemplo 1: Supongamos que los valores a priori de las componentes de  $\mathbf{w}$  están distribuidos uniformemente e independientemente sobre intervalos conocidos, estos son

$$w_{i\min} \leq w_i \leq w_{i\max}, \text{ para } i=1,2,\dots,n$$

Para todo  $\mathbf{w}$  que no satisface estas desigualdades  $\pi(\mathbf{w}) = 0$  o bien  $\ln[\pi(\mathbf{w})] = -\infty$ .

Con esto, estaremos seguros que  $\hat{\mathbf{w}}_{MAP}$  respetará las condiciones que definen el espacio paramétrico admisible a priori. Por el contrario, para todo  $\mathbf{w}$  que satisface las desigualdades,  $\pi(\mathbf{w})$  y  $\ln[\pi(\mathbf{w})]$  tomarán valores independientes de  $\mathbf{w}$ .

Si  $\hat{\mathbf{w}}_{ML}$  pertenece al espacio paramétrico admisible a priori, entonces

$$[\mathbf{w} - \hat{\mathbf{w}}]^\top \mathbf{S}^{-1} [\mathbf{w} - \hat{\mathbf{w}}] - [\mathbf{w} - \hat{\mathbf{w}}]^\top \hat{\mathbf{w}}_{MAP} = \hat{\mathbf{w}}_{ML}^\top \mathbf{S}^{-1} \mathbf{y}$$

Ejemplo 2: Supongamos ahora que los parámetros están distribuidos a priori de acuerdo a una ley normal  $N(\mathbf{w}_0, \mathbf{\Sigma})$ , donde  $\mathbf{w}_0$  y  $\mathbf{\Sigma}$  son conocidos.

El valor promedio a priori  $W_0$  podría corresponder por ejemplo, a alguna estimación hecha en el sentido de Máximo de Verosimilitud obtenida gracias a algunas mediciones anteriores. Por otro lado,  $\Omega$  podría reflejar alguna estimación de la incertidumbre sobre la variabilidad o dispersión de  $W_0$ , también obtenida a partir de mediciones anteriores.

Si parametrizamos la densidad de probabilidad a priori, entonces tendremos que

$$\pi(W) \propto \frac{1}{(2\pi)^n |\Omega|} \exp \left[ -\frac{1}{2} [W - W_0]^T \Omega^{-1} [W - W_0] \right]$$

Sustituyendo esta última expresión en el criterio logarítmico obtenemos el siguiente resultado

$$J_{MAP}(W) = \underbrace{\ln [\pi(W)]}_{\text{log-verosimilitud}} - \underbrace{\frac{1}{2} [W - W_0]^T \Omega^{-1} [W - W_0]}_{\text{término adicional cuadrático}}$$

La introducción de información a priori se traduce en la suma o resta de un término adicional al criterio de Máximo de Verosimilitud, que en este

(73)

(45)

Este caso corresponde a un término de "penalización" cuadrático. Las técnicas de "regularización", utilizadas por ejemplo, en procesamiento de imágenes también conducen a una penalización cuadrática que puede tener una interpretación Bayesiana.

Si nos concentramos en el caso de regresión, para un modelo polinomial (regresión lineal), tenemos el siguiente

$$t = R w^* + b, \quad b \sim N(0, \Sigma)$$

en donde  $\Sigma$  es una matriz conocida, entonces, el estimador MAP,  $\hat{w}_{MAP}$  estará dado por la siguiente fórmula analítica

$$\hat{w}_{MAP} = \arg \max_w \left\{ \ln [\pi(t|w)] + \ln [\pi(w)] \right\},$$

$$\hat{w}_{MAP} = \arg \min_w \left\{ -\ln [\pi(t|w)] + \frac{1}{2} [w - w_0]^T \Sigma^{-1} [w - w_0] \right\}$$

$$\hat{w}_{MAP} = [R^T \Sigma^{-1} R + \Omega^{-1}]^{-1} [R^T \Sigma^{-1} t + \Omega^{-1} w_0].$$

El criterio  $J_{MAP}(w)$  se parece a la ecuación (1.4) propuesta en el libro de C. M. Bishop, al que le denomina suma del cuadrado para todos los coeficientes con un término de penalización, proponiendo una función sobre los errores modificada, de manera que

$$J_{MAP}(w) \approx \tilde{E}(w) = \frac{1}{2} \sum_{n=1}^N [y_n(x_n, w) - t_n]^2 + \frac{\lambda}{2} \|w\|^2$$

en donde  $\|w\|^2 \equiv w^T w$ , y el coeficiente de Lagrange  $\lambda$  gobierna la importancia relativa del término de regularización comparado con la suma de los cuadrados del término del error.

En la literatura de estadística, también se les conoce a estos estimadores como métodos de contracción (shrinkage) dado que reducen los valores de los coeficientes o parámetros. Para el caso particular de un regularizador cuadrático, el método es llamado ridge regresor, y en el contexto de redes neuronales como modelo en donde se estimarán los pesos, se le conoce como: decaimiento de pesos (weight decay).

Una particularidad de los estimadores MAP es que no es necesario disponer de un número de medidas superior o igual al número de parámetros para asegurar la unicidad de la estimación. Por ejemplo, en un caso de medicina clínica podríamos determinar un límite de medidas a realizar por razones de costos y confort del paciente, todo lo anterior con la condición de disponer de información a priori que sea fiable sobre los posibles valores de los parámetros de una cierta población.

#### 2.4.2. Criterio de riesgo mínimo (RM).

Suponemos que es posible evaluar el costo de creer que los parámetros valen  $W$  cuando en realidad valen  $W^*$ , este costo lo notaremos mediante el criterio  $J(W|W^*)$ . Como  $W^*$  es desconocido, podemos buscar  $\hat{W}_{RM}$  que minimiza el riesgo, definido como el costo promedio sobre todos los valores posibles de  $W^*$ , sabiendo que hemos observado  $t$  datos.

(76)

(35)

Entonces, el criterio a minimizar está dado por

$$J_{RM}(w) = \int J(w|w^*) \pi(w^*|t) dw^*$$

a la condición de que la integral exista, y donde

$\pi(w^*|t) dw^*$  es la probabilidad de que  $w^*$  pertenece al elemento de hipervolumen  $dw^*$ , sabiendo que contamos con  $t$  datos. Ahora,

es necesario expresar este criterio en función

de los datos que se suponen conocidos. De acuerdo a la regla de Bayes, tenemos

entonces que

$$J_{RM}(w) = \frac{1}{\pi(t)} \int J(w|w^*) \pi(t|w^*) \pi(w^*) dw^*$$

pero como  $\pi(t)$  no depende de  $w$ , y es una constante cercana a 1, entonces

$$J_{RM}(w) = \int J(w|w^*) \pi(t|w^*) \pi(w^*) dw^*$$

77

87

En este criterio,  $J(w|w^*)$  traduce todo lo que conocemos a cerca de la modelización o modelado,  $\Pi(t|w^*)$  traduce la información a cerca de la naturaleza del ruido que perturba al sistema o fenómeno físico, y  $\Pi(w^*)$  traduce la información a priori sobre los parámetros del modelo.

Comentarios: La evaluación de  $J_{RM}(w)$  requiere del cálculo de una integral múltiple (haciendo uso de algoritmos de integración sobre el conjunto  $\mathbb{R}^n$ ) lo cual es un problema desde el punto de vista del cálculo numérico, una alternativa es el uso de técnicas de muestreo estocásticas que permiten calcular  $\hat{W}_{RM}$  sin evaluar  $J_{RM}(w)$

El estimador de riesgo mínimo  $\hat{W}_{RM}$  podría ser muy sensible a alguna modificación en las colas de la distribución  $\Pi(w)$ . Además,  $\hat{W}_{RM}$  si varía con respecto a reparametrizaciones del modelo ya que  $J(w|w^*)$  depende del modelado.

78

55

Ejemplo: Supongamos que existe un costo de equivocarse y este es cuadrático, entonces

$$J(w|w^*) = [w - w^*]^T Q [w - w^*],$$

en donde  $Q$  es una matriz de ponderación simétrica y definida positiva. El riesgo mínimo existente sabiendo que se cuenta con  $t$ , se escribe de la siguiente forma

$$J_{RM}(w) = \int [w - w^*]^T Q [w - w^*] \pi(w|t) dw$$

El estimador de riesgo mínimo debe entonces cumplir con la siguiente ecuación

$$\textcircled{1} = \frac{\partial J_{RM}(w)}{\partial w} = 2Q \int \hat{w}_{RM} - w \pi(w|t) dw$$

de donde, puesto que  $Q$  es inversible, entonces

$$\begin{aligned} \hat{w}_{RM} \int \pi(w|t) dw &= \hat{w}_{RM} = \int w \pi(w|t) dw \\ &= E\{w|t\}. \end{aligned}$$

(79)

El estimador de riesgo mínimo es entonces el valor promedio de  $W$  a posteriori para cualquier valor de  $Q$ . El estimador obtenido, se dice que es el "mejor estimador en promedio cuadrático". Luego, la regla de Bayes permite calcularlo de acuerdo a la siguiente ecuación

$$\hat{W}_{RM} = \int_{\Theta} W \left[ \frac{\pi(t|w)\pi(w)}{\pi(t)} \right] dw.$$

De forma general, no se puede obtener de manera analítica salvo en algunas excepciones. Supongamos nuevamente que tenemos nuestro regresor lineal

$$t = R w^* + b, \quad b \sim N(0, \Sigma)$$

y  $w$  también está distribuido normalmente, es decir:  $w \sim N(w_0, \Sigma)$ .

En este caso, notamos que el estimador  $\hat{W}_{MAP}$  y  $\hat{W}_{RM}$  coinciden, y entonces

$$\hat{W}_{RM} = \hat{W}_{MAP} = [R^T \Sigma^{-1} R + \Sigma^{-1}]^{-1} [R^T \Sigma^{-1} t + \Sigma^{-1} w_0]$$

Para ~~los~~ casos donde se necesita robustez dado que es posible que existan también datos aberrantes, podemos optar por estimadores robustos como los propuestos por Huber y Tukey que se clasifican dentro de lo que se denomina M-estimadores,  $\hat{W}_M$ .

## 2.5. Teoría de decisión Bayesiana

Ya se ha visto como la teoría de probabilidad nos provee de un marco matemático consistente para cuantificar y manipular la incertidumbre. Ahora toca discutir sobre la teoría de la decisión que también se basa en probabilidades y nos permite tomar decisiones de manera óptima en situaciones que involucran incertidumbre como las que se presentan en reconocimiento de patrones.

Suponga que contamos con un vector  $x$  de entrada en conjunto con otro vector  $t$  de variables objetivo, y nuestra meta es predecir  $t$  dado un nuevo valor para  $x$ . Por ejemplo, para el caso de regresión,  $t$  corresponde a variables continuas, mientras que en el caso de clasificación  $t$  podría ser etiquetas de las clases. Entonces, la probabilidad conjunta  $p(x, t)$  provee un resumen completo de la incertidumbre asociada a estas dos variables. La determinación de  $p(x, t)$  a partir de los datos de entrenamiento corresponde a un ejemplo de inferencia y es un problema muy difícil de

de resolver. En aplicaciones prácticas, es posible realizar una predicción específica para el valor de  $t$ , de manera general se toma una acción en específico basada sobre nuestro entendimiento de los valores  $t$ .

Considerese por el momento, un ejemplo sobre diagnóstico médico en el cual se obtienen imágenes de rayos X para un paciente. Se desea determinar si el paciente tiene o no cáncer. Para este caso, el vector de entrada  $x$  es el conjunto de intensidades de los pixeles de la imagen y la variable  $t$  de salida representará la presencia de cáncer, que denotaremos con la clase  $C_1$ , o bien la ausencia de cáncer denotada por la clase  $C_2$ .

Por ejemplo, podríamos pensar en que  $t$  sea una variable binaria de manera que  $t=0$  corresponda a la clase  $C_1$  y entonces para  $t=1$  corresponda a la clase  $C_2$ . Esta elección de clases es particularmente conveniente para modelos probabilísticos.

(B)

(A)

En fincas, el problema de inferencia involucra la determinación de la distribución conjunta  $P(x, C_k) \triangleq p(x, t)$ . Se tiene que "decidir" si se da tratamiento o no al paciente y por lo tanto corresponde a un problema de teoría de la decisión.

Como ya se vió en la sección anterior, la regla de Bayes nos permite establecer una relación basada en probabilidades conjuntas y marginales, de manera que podemos calcular la probabilidad a posteriori

$$P(C_k|x) = \frac{P(x|C_k)p(C_k)}{P(x)}$$

No se sabe que en esta propuesta todas las cantidades de la ecuación de Bayes provienen de la distribución conjunta  $p(x, C_k)$ .

Ahora, supongamos que nuestra meta es realizar la menor cantidad posible de malas clasificaciones (misclassifications).

Entonces, necesitamos una regla que asigne cada valor de  $x$  a una de las clases disponibles. Tal regla dividirá el espacio de entrada en regiones  $R_k$  denominadas regiones de ~~clasificación~~ decisión, una para cada clase, de tal manera que todos los puntos en  $R_k$  serán asignados a la clase  $C_k$ . Los límites entre las regiones de decisión son llamados límites de decisión o superficies de decisión. Nótese que cada región de decisión no debe de ser contigua aunque podría contener algún número de regiones disjuntas.

Para encontrar una regla de decisión óptima, consideremos primero las dos clases del problema de cancer. Un error ocurre cuando un vector de entrada que pertenece a la clase  $C_1$ , es asignado a la clase  $C_2$  o viceversa.

La probabilidad de que ocurra dicho error está dada por

D)  $P(\text{mistake}) = P(x \in R_1, C_2) + P(x \in R_2, C_1)$

$$p(\text{mistake}) = \int_{R_1} p(x, C_2) dx +$$

$$\int_{R_2} p(x, C_1) dx.$$

Para minimizar  $p(\text{mistake})$  se debe reacomodar cada punto en  $x$  de manera que sea asignado a la clase que tiene el valor mas pequeño del integrando de la ecuación anterior. Entonces, si  $p(x, C_1) > p(x, C_2)$  para cualquier  $x$ , entonces  $x$  se asigna a la clase  $C_1$ .  $\Rightarrow p(C_k | x)$  es el mayor.

Para el caso de clasificación correcta,

$$p(\text{correct}) = \sum_{k=1}^K p(x \in R_k, C_k),$$

$$= \sum_{k=1}^K \int_{R_k} p(x, C_k) dx,$$

(E)

que se maximiza cuando las regiones  $R_k$  son elegidas de manera que  $x$  es asignada

a la clase para la cual  $p(x, c_k)$  toma el valor más grande.

Para minimizar  $q$  necesitamos elegir  $x$  de modo que sea considerado el mejor candidato de las  $n$  alternativas. Esto implica que  $b(x)$  sea menor que  $b(c_k)$  para cualquier  $x$  distinto de  $c_k$ .

Por el caso de igualdad,  $b(x) = b(c_k)$ .

Por el caso de igualdad,  $b(x) < b(c_k)$ .

$$\text{Entonces } b(x) < b(c_k) \quad \forall x \in R \setminus \{c_k\}$$

$$b(x) < b(c_k) \quad \forall x \in R \setminus \{c_k\}$$

(F)

Por lo tanto,  $b(x) < b(c_k)$  para todo  $x \in R \setminus \{c_k\}$ .

(T)