Clasificación Bayesiana Estimación paramétrica de densidades desconocidas

FaMAF

2019

Hipótesis sobre datos en un entorno Bayesiano

Podemos diseñar un estimador Bayesiano óptimo si sabemos:

- $p(\omega_i)$ probabilidades a priori
- $p(\mathbf{x}|\omega_i)$ densidades condicionales a la clase.

Desafortunadamente, raramente se tiene toda esta información completa !!
Podemos diseñar un clasificador Bayesiano usando una muestra de entrenamiento D para estimar parámetros no especificados.

- Estimación a priori no es un problema.
- Muestras son a menudo muy chicas para estimación condicional por clase, sobre todo si el espacio de características es de dimensión muy alta.

Diseño

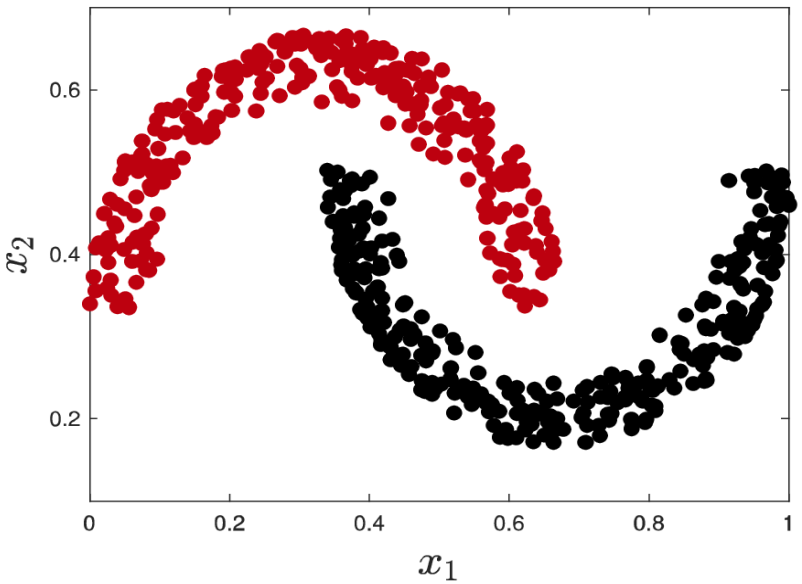
- Para simplificar el problema de diseño del clasificador, se asumen formas paramétricas de las densidades $p(\mathbf{x}|\omega_i, \theta_i)$, donde θ_i son vectores de parámetros desconocidos que cambian por clase.
- Se asume también que se sabe el número de parámetros a estimar, es decir, la dimensión del vector θ por clase.
- Las técnicas que se discutirán son
  - Máxima verosimilitud (MV)
  - Estimación Bayesiana (EB)

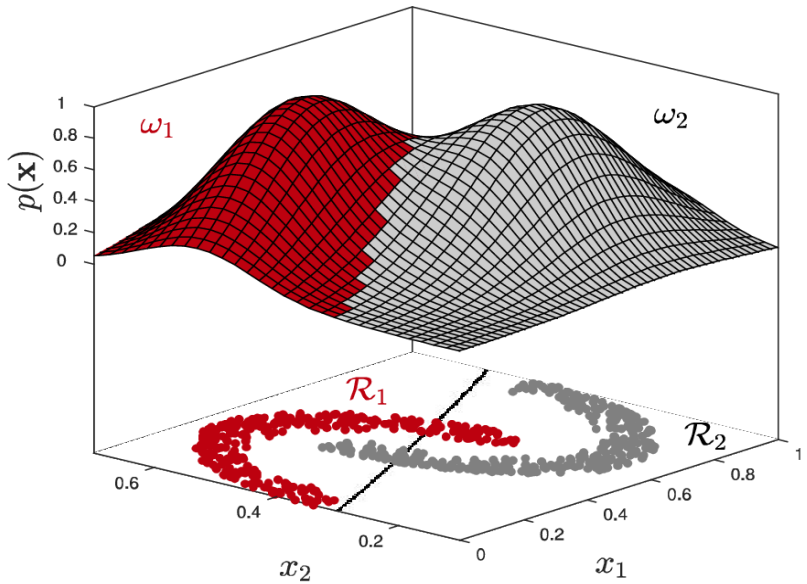
Ejemplo

Supongamos tener k clases diferentes para clasificar.

- Si $p(\mathbf{x}|\omega_i)$ es una densidad normal, entonces se puede modelar estimando los parámetros media $\boldsymbol{\mu_i}$ y matriz de covarianza $\boldsymbol{\Sigma}_i$ a partir de datos etiquetados D_i.
- La densidad $p(\mathbf{x}|\omega_i)$ se puede modelar adecuadamente con una única distribución normal.
- El clasificador decide por la clase i-ésima si $p(\omega_i|\mathbf{x}, \hat{\boldsymbol{\mu}}_{\boldsymbol{i}}, \hat{\boldsymbol{\Sigma}}_i)$ es máxima.

Ejemplo





Si la distribución de las clases no siguen una distribución normal (izquierda), entonces la densidad $p(\mathbf{x}|\omega_i)$ puede sub-estimarse cuando se modela con una sola densidad Gaussiana, generando un clasificador con baja capacidad de generalización (derecha).

Ejemplo

Solución:

- Combinar varias densidades normales para mejorar la estimación de la función de verosimilitud $p(\mathbf{x}|\omega_i)$ de acuerdo a la distribución real de los datos:

$$p\left(\mathbf{x}|\omega_i\right) = \alpha_1 \mathcal{N}\left(\boldsymbol{\mu}_{i1}, \boldsymbol{\Sigma}_{i1}\right) + \ldots + \alpha_m \mathcal{N}\left(\boldsymbol{\mu}_{im}, \boldsymbol{\Sigma}_{im}\right) \qquad (1)$$

- Problemas:
  - ¿Cómo determinar los parámetros $\boldsymbol{\mu}_{ik}, \boldsymbol{\Sigma}_{ik}$ y α_k
  - ¿Cuál es el número óptimo m^* de densidades normales?

Estimación de Máxima Verosimilitud

- Estimar el vector de parámetros desconocidos $\boldsymbol{\theta}$ de la distribución $p(\mathbf{x}|\boldsymbol{\theta})$ a partir del conjunto de datos $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
- Función de verosimilitud de con respecto de $\boldsymbol{X}$:

$$p(\boldsymbol{X}|\boldsymbol{\theta}) = p\left(\mathbf{x}_1, \ldots, \mathbf{x}_n|\boldsymbol{\theta}\right) = \prod_{i=1}^{n} p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) \tag{2}$$

- La estimación de máxima verosimilitud (MV) es un método para encontrar los parámetros que mejor ajusten la función $p(\mathbf{x}|\boldsymbol{\theta})$.

Estimación de Máxima Verosimilitud

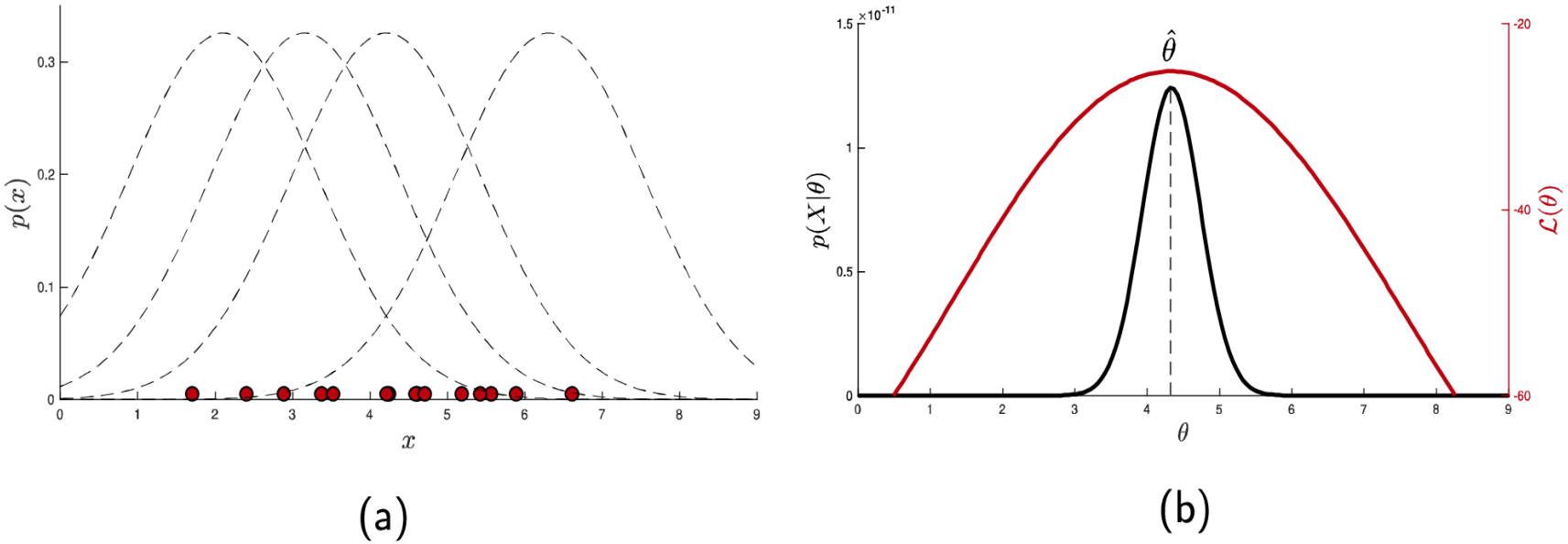
- Suponga que se quiere estimar un sólo parámetro $\boldsymbol{\theta}$, el estimador $\hat{\boldsymbol{\theta}}$ de máxima verosimilitud es el valor de $\boldsymbol{\theta}$ que maximiza $p(\mathbf{x}|\boldsymbol{\theta})$, esto es

$$\hat{\theta} = \arg\max_{\theta} \prod_{i=1}^{n} p\left(\mathbf{x}_i|\theta\right) \tag{3}$$

- Ese estadístico también realiza el máximo de la función de log-verosimilitud $\mathcal{L}(\theta) \equiv \ln p(\mathbf{x}|\theta)$, esto es

$$\hat{\theta} = \arg\max_{\theta} \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i|\theta\right) \tag{4}$$

Estimación de Máxima Verosimilitud



(a) Puntos en una dimensión tomados de una distribución Gaussiana con varianza específica y media desconocida. Se muestran cuatro de un número infinito de distribuciones fuente candidatas. (b) Verosimilitud (curva negra) y log-verosimilitud (curva roja) en función de la media. El valor que maximiza la verosimilitud es señalado con $\hat{\theta}$.

Estimación de Máxima Verosimilitud

Generalizando la verosimilitud como una función de los k parámetros desconocidos $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$; el estimador vectorial de máxima verosimilitud es el vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)$que maximiza la $\mathcal{L}(\boldsymbol{\theta}) \equiv \ln p(\mathbf{x}|\boldsymbol{\theta})$

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n} \ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) \tag{5}$$

En la mayoría de los casos, esto se reduce a resolver la ecuacion

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \ln p\left(\mathbf{x}_i|\boldsymbol{\theta}\right) = \mathbf{0} \tag{6}$$

donde el operador gradiente es

$$\nabla_{\boldsymbol{\theta}} \equiv \left[\frac{\partial}{\partial\theta_1}, \ldots, \frac{\partial}{\partial\theta_k}\right]^T \tag{7}$$

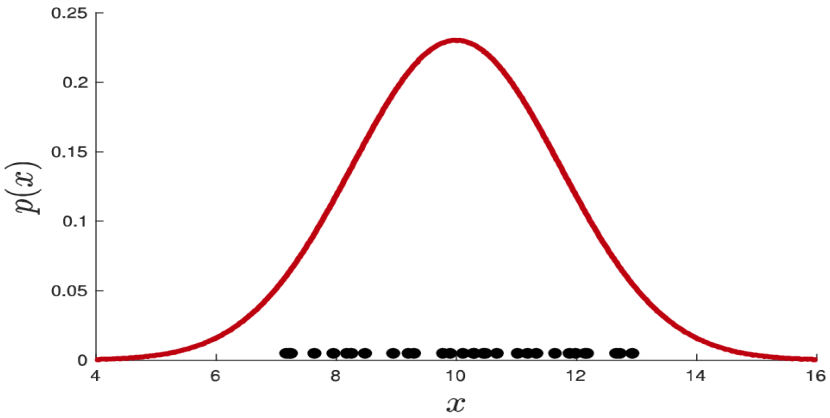
Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$

- Suponga que n muestras $x_1, \ldots, x_n$ fueron generadas a partir de una distribución normal univariada con media μ y varianza σ^2.
- La función de log-verosimilitud de la normal unidimensional es

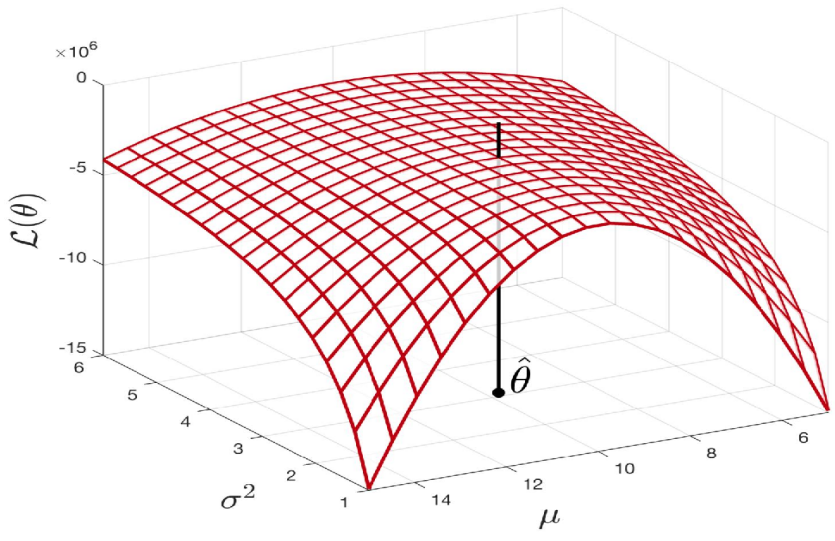
$$\mathcal{L}\left(\mu, \sigma^2\right) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{k=1}^{n} \left(x_k - \mu\right)^2 \tag{8}$$

- Para estimar los parámetros μ y σ^2 por máxima verosimilitud se buscan los valores críticos de la $\mathcal{L}\left(\mu, \sigma^2\right)$

Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$



(a)



(b)

(a) Conjuntos de datos unidimensionales tomados aleatoriamente de una distribución Gaussiana $\mathcal{N}\left(10, 3\right))$. (b) Función de log-verosimilitud en cuyo máximo se encuentran los parámetros estimados $\hat{\theta}$: $\hat{\mu} = 9.7$, $\hat{\sigma}^2 = 3.1$

Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$

Aplicando el método de Máxima verosimilitud dado en (4) a la función de log verosimilitud en (8)

$$\begin{aligned} \frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} &= \sum_{k=1}^{n} \left(x_k - \mu\right) = 0 \\ \frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{k=1}^{n} \left(x_k - \mu\right)^2 = 0 \end{aligned} \tag{9}$$

Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$

Reacomodando los términos en (9) se obtienen los estimadores de máxima verosimilitud para μ y σ^2

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{k=1}^{n} x_k \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{k=1}^{n} (x_k - \hat{\mu})^2 \end{aligned} \tag{10}$$

Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$

Sesgo de un estimador: diferencia entre el valor esperado de $\hat{\theta}$ y el valor real θ.

- El estimador MV para la media μ no está sesgado: $E(\hat{\mu}) = \mu$.
- El estimador MV para la varianza σ^2 es sesgado:

$$\mathbb{E}\left[\frac{1}{n}\sum_{k=1}^{n}\left(x_k - \mu\right)^2\right] = \underbrace{\frac{n-1}{n}}_{\text{sesgo}}\sigma^2 \neq \sigma^2 \tag{11}$$

- Varianza de la muestra sin sesgo:

$$S = \frac{1}{n-1}\sum_{k=1}^{n}\left(x_k - \hat{\mu}\right)^2 \tag{12}$$

Estimación de Máxima Verosimilitud para la $\mathcal{N}\left(\mu, \sigma^2\right)$

En el caso más general, la función de log-verosimilitud de la normal multivariada es:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{X} - \boldsymbol{\mu}) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| \tag{13}$$

donde los estimadores MV para $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$ son

$$\begin{aligned} \hat{\boldsymbol{\mu}} &= \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k \\ \hat{\boldsymbol{\Sigma}} &= \frac{1}{n-1}\sum_{k=1}^{n} (\mathbf{x}_k - \hat{\boldsymbol{\mu}})(\mathbf{x}_k - \hat{\boldsymbol{\mu}})^T \end{aligned} \tag{14}$$

Estimacion Bayesiana

- Estimación de máxima verosimilitud supone θ un parámetro fijo.
- Estimación Bayesiana considera θ una variable aleatoria con densidad $p(\theta|\theta_0)$, con θ_0 parámetros conocidos.
- La muestra de entrenamiento permite cambiar el problema de etiquetar clases por el problema de estimar densidades.

Estimacion Bayesiana

- Si tenemos k clases, queremos clasificar el dato x en la clase que produza el máximo valor de $p(\omega_i|x)$, $1 \leq i \leq k$.

$$p(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{\sum_{i=1}^{k} p(x|\omega_i)P(\omega_i)}$$

- Sabemos que $p(x|\omega_i)$ tiene un parámetro θ desconocido.
- Se tiene una muestra aleatoria D_i de la distribución $p(x|\omega_i)$.
- Lo mas cercano que podemos conocer es $p(x|\omega_i, D_i)$. Por lo cual la regla va a clasificar x en la clase i si ésta hace máxima a

$$p(\omega_i|x, D) = \frac{p(x|\omega_i, D_i)P(\omega_i)}{\sum_{i=1}^{k} p(x|\omega_i, D_i)P(\omega_i)}$$

Estimacion Bayesiana

- Sea θ una variable aleatoria con densidad $p(\theta)$ conocida. Para cada θ fijo, sea $p(x|\theta)$ una densidad conocida.
- Podemos encontrar $p(x|\omega_i, D_i)$ integrando la densidad conjunta $P(x, \theta|D)$ sobre θ.

$$p(x|\omega_i, D_i) = \int p(x, \theta|D)d\theta = \int p(x|\theta)p(\theta|D)d\theta)$$

- Si $p(\theta|D)$ tiene un pico agudo en algun valor θ_e resulta $p(x|\omega_i, D) \sim p(x|\theta_e)$
- Si no es asi se promedia $p(x|\theta)$ sobre todos los valores de θ.

Estimacion Bayesiana

- La densidad a posteriori $p(\theta|\mathrm{D})$

$$\begin{aligned} p(\theta|\mathrm{D}) &= \frac{p(\mathrm{D}|\theta)\cdot p(\theta)}{\int p(\mathrm{D}|\theta)\cdot p(\theta)d\theta} \\ &= \alpha \prod_{k=1}^{n} p\left(x_k|\theta\right)\cdot p(\theta) \end{aligned} \tag{15}$$

- $p(\mathrm{D}|\theta) = \prod_{k=1}^{n} p\left(x_k|\theta\right)$ puede calcularse en forma recursiva.

Estimacion Bayesiana: Caso Normal univariado

- Sea $p(x|\theta_i)$ la densidad $\mathcal{N}\left(\mu_i, \sigma_i^2\right)\big)$, donde μ_i es desconocido y σ_i^2 conocido, $1 \leq i \leq k$.
- Supongamos que $p(\mu_i) \sim \mathcal{N}\left(\mu_{0,i}, \sigma_{0,i}^2\right)\big)$ con $\mu_{0,i}$ y $\sigma_{0,i}^2$ conocidos.
- Entonces

$$\begin{aligned} p(\mu_i|\mathrm{D_i}) &= \frac{p(\mathrm{D_i}|\mu_i) \cdot p(\mu_i)}{\int p(\mathrm{D_i}|\mu_i) \cdot p(\mu_i) d\mu_i} \\ &= \alpha \prod_{k=1}^{n} p\left(x_k|\mu_i\right) \cdot p(\mu_i) \end{aligned} \tag{16}$$

- Reemplazando por las densidades supuestas queda

$$\begin{aligned} p(\mu_{0,i}|\mathrm{D_i}) &\sim N\left(\mu_{i,n}, \sigma_{i,n}^2\right) \\ \mu_{i,n} &= \left(\frac{n\sigma_0^2}{n_0\sigma_{i,0}^2+\sigma_i^2}\right)\overline{\mathrm{x}}_\mathrm{n} + \frac{\sigma_i^2}{n\sigma_{i,0}^2+\sigma_i^2} \cdot \mu_{0,i} \\ \sigma_n^2 &= \frac{\sigma_{i,0}^2\sigma_i^2}{n\sigma_{i,0}^2+\sigma_i^2} \end{aligned} \tag{17}$$

Estimacion Bayesiana: Caso Normal univariado

- Usando la densidad $p(\mu_i|\mathrm{D}_i)$ calculada, y la fórmula

$$p(x|\omega_i, \mathrm{D}_i) = \int p(x|\mu) \cdot p(\mu|\mathrm{D}) d\mu \tag{18}$$

se ve que $p(x|\omega_i, \mathrm{D}_i)$ tiene distribución Normal

$$p(x|\omega_i, \mathrm{D}_i) \sim N\left(\mu_{i,n}, \sigma_i^2 + \sigma_{i,n}^2\right) \tag{19}$$

- Junto a las probabilidades a priori $p(\omega_i)$ se tiene toda la información para diseñar el clasificador.

Estimacion Bayesiana Recursiva

- $D^n = \{x_1, \ldots, x_n\}$ muestra de entrenamiento.
- La ecuación

$$p(\mathrm{D^n}|\theta) = \prod_{k=1}^{n} p\left(x_k|\theta\right) \tag{20}$$

puede calcularse en forma recursiva

$$p\left(\mathrm{D^n}|\theta\right) = p\left(x_n|\theta\right) p\left(\mathrm{D^{n-1}}|\theta\right)$$

- Por lo tanto $p\left(\theta|\mathrm{D}\right)$ tambien puede calcularse en forma recursiva, observando que antes de que se tenga ningun dato,

$$p\left(\theta|\mathrm{D^0}\right) = p(\theta) \tag{21}$$

Estimacion Bayesiana Recursiva

- Cuando se observa el primer dato

$$p\left(\theta|\mathrm{D}^1\right) \propto p(x_1|\theta)p(\theta) = p(x_1|\theta)p\left(\theta|\mathrm{D}^0\right) \tag{22}$$

- Cuando se observa el segundo dato

$$p\left(\theta|\mathrm{D}^2\right) \propto p(x_2|\theta)p(x_1|\theta)p(\theta) = p(x_2|\theta)p\left(\mathrm{D}^1|\theta\right) \tag{23}$$

- En general

$$p\left(\theta|\mathrm{D}^{\mathrm{n}}\right) = \frac{p\left(x_n|\theta\right)p\left(\theta|\mathrm{D}^{\mathrm{n}-1}\right)}{\int p\left(x_n|\theta\right)p\left(\theta|\mathrm{D}^{\mathrm{n}-1}\right)d\theta} \tag{24}$$

Ejemplo Estimacion Bayesiana Recursiva

- Supongamos que las muestras provienen de una distribución uniforme

$$p(x|\theta) \sim U(0,\theta) = \begin{cases} 1/\theta & 0 \leq x \leq \theta \\ 0 & \text{afuera} \end{cases} \tag{25}$$

- Se supone un prior no informativo o flat

$$p(\theta) \sim U(0,10) \tag{26}$$

- Datos seleccionados aleatoriamente de la distribución subyacente D={ 4,7,2,8}

Ejemplo Estimacion Bayesiana Recursiva

- Antes que se tenga ningún dato

$$p\left(\theta|D^0\right) = p(\theta) = U(0, 10) \tag{27}$$

- Cuando el primer dato $x_1 = 4$ se observa, olvidando la renormalización

$$p\left(\theta|D^1\right) \propto p(x|\theta)p\left(\theta|D^0\right) = \begin{cases} 1/\theta & 4 \leq \theta \leq 10 \\ 0 & \text{afuera} \end{cases} \tag{28}$$

- Cuando el siguiente dato $x_2 = 7$ se observa

$$p\left(\theta|D^2\right) \propto p(x|\theta)p\left(\theta|D^1\right) = \begin{cases} 1/\theta^2 & 7 \leq \theta \leq 10 \\ 0 & \text{afuera} \end{cases} \tag{29}$$

Ejemplo Estimacion Bayesiana Recursiva

- La forma general de la solución es

$$p\left(\theta|D^n\right) \propto 1/\theta^n \quad \max_x \left[D^n\right] \leq \theta \leq 10 \tag{30}$$

- Con la muestra conjunta, el estimador MV es $\theta = 8$, lo cual implica una distribución

$$p(\theta|D) \sim U(0, 8) \tag{31}$$

- En cambio, la metodología Bayesiana requiere una integración

$$\mathrm{p(x}|D) = \int p(x|\theta)p(\theta|D)d(\theta) \tag{32}$$

Ejemplo Estimacion Bayesiana Recursiva

- La densidad solución es uniforme hasta $\theta = 8$, y tiene una cola para valores mayores indicando que la influencia del prior $p(\theta)$ no ha sido descartado por la información de la muestra de entrenamiento.
- MV estima un punto en el espacio de θ, el método bayesiano estima una densidad.
- Técnicamente, no se pueden comparar los estimadores pero si se pueden comparar las distribuciones $p(x|D)$

Identificabilidad en el enfoque Bayesiano

- Para la mayoría de las densidades $p(x|\theta)$, la sucesión de densidades $p(\theta|D)$ converge a una delta de Dirac.
- Esto implica, que con un gran número de muestras, habría un solo θ que ajuste a esos datos, por lo cual θ puede ser identificado unívocamente de $p(x|D)$.
- Cuando mas de un valor de θ ajusta los datos, esto es $p(D|\theta)$ tiene el mismo valor para mas de un θ, este resulta no identificable.
- Sin embargo, $p(x|D^n)$ va a converger a $p(x|w_i)$, a pesar de la no identificabilidad de θ.
- El problema de clasificación es mas simple que el de estimación .

Diferencias entre Clasificacion con plug in MV y Aprendizaje Bayesiano

- Asintóticamente, en la mayoría de los casos coinciden.
- En muestra finita, sin embargo, hay diferencias a considerar
- Complejidad computacional
  - MV requiere calculo diferencia o búsqueda por gradiente
  - AB requiere integración multidimensional
- Interpretabilidad
  - MV da un solo valor
  - AB un promedio pesado de modelos, que a menudo son difíciles de interpretar