Aprendizaje Bayesiano

Aprendizaje Automático

Ingeniería Informática

Fernando Fernández Rebollo y Daniel Borrajo Millán

Grupo de Planificación y Aprendizaje (PLG)
Departamento de Informática
Escuela Politécnica Superior
Universidad Carlos III de Madrid

27 de febrero de 2009

En Esta Sección:

Introducción

- La teoría de decisión bayesiana se basa en dos suposiciones:
  - El problema de decisión se puede describir en términos probabilísticos:
    - Dado un conjunto de datos, D, cuál es la mejor hipótesis h del cunjunto de hipótesis H
    - La mejor hipótesis es la hipótesis más probable
  - Todos los valores de las probabilidades del problema son conocidas
- Decisiones tomadas en función de ciertas observaciones

Ejemplo: el Caso de la Moneda Trucada

- Espacio de hipótesis: {cara, cruz}
- Espacio de observaciones: {brillo, mate}
- Lanzo una moneda, recibo la observación, D, y genero una hipótesis, h
- Preguntas:
  - ¿Cuál es la mejor hipótesis?
  - ¿Cuál es la hipótesis más probable?
  - ¿Cuál es la probabilidad de obtener cara?
  - ¿Cuál es la probabilidad de obtener cruz?
  - ¿Cuál es la probabilidad de obtener *cara*, habiéndo recibido como observación *brillo*?
  - ¿Cuál es la probabilidad de obtener *cruz*, habiéndo recibido como observación *mate*?

Probabilidades a Priori

- $P(h)$: Probabilidad de que la hipótesis h sea cierta o *Probabilidad a priori* de la hipótesis h
  - Refleja el conocimiento que tenemos sobre las oportunidades de que la hipótesis h sea cierta antes de recibir ninguna observación
  - Si no tenemos ningún conocimiento a priori, se le podría asignar la misma probabilidad a todas las hipótesis
- $P(D)$: Probabilidad de que recibamos la observación D o *Probabilidad a priori* de la observación D
  - Refleja la probabilidad de recibir la observación D, cuando no tenemos ninguna idea sobre cuál es la hipótesis real

Probabilidades a Posteriori

- $P(D|h)$: Probabilidad de observar el dato D, cuando se cumple la hipótesis h o *Probabilidad a posteriori* de la observación D
- $P(h|D)$: Probabilidad de que se cumpla la hipótesis h, dado que se ha obtenido el dato D, o *Probabilidad a posteriori* de la hipótesis h

Teorema de Bayes

- En aprendizaje inductivo, estamos interesados en calcular las probabilidades de las hipótesis a posteriori, ya que son las que se obtienen tras recibir observaciones o ejemplos de entrenamiento.
- Teorema de Bayes:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

  - Calcula la probabilidad a posteriori de la hipótesis en función de otras probabilidades

Decisor MAP

- Decisor máximo a posteriori:

$$h_{MAP} = \arg\max_{h \in H} P(h|D) =$$

$$\arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} =$$

$$\arg\max_{h \in H} P(h|D) = P(D|h)P(h)$$

- El decisor de máxima verosimilitud, ML (*maximum likelihood*), asume que todas las hipótesis son equiprobables a priori:

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Ejemplo: el Caso de la Moneda Trucada

- Probabilidades a priori:
  - $P(cara) = 0{,}2,\ P(cruz) = 0{,}8$
- Probabilidades a posteriori:
  - $P(brillo|cara) = 0{,}9,\ P(mate|cara) = 0{,}1$
  - $P(brillo|cruz) = 0{,}6,\ P(mate|cruz) = 0{,}4$
- Tiro la moneda y obtengo *brillo*:

$$h_{MAP} = \arg \max_{cara,cruz} P(brillo|h)P(h) =$$

$$\arg \max_{cara,cruz} (P(brillo|cara)P(cara), P(brillo|cruz)P(cruz)) =$$

$$\arg \max_{cara,cruz} (0{,}9 \times 0{,}2, 0{,}6 \times 0,8) =$$

$$\arg \max_{cara,cruz} (0{,}18, 0{,}48) = cruz$$

Ejemplo: el Caso de la Moneda Trucada

- Sin embargo:

$$h_{ML} = \arg \max_{cara,cruz} P(brillo|h) =$$

$$\arg \max_{cara,cruz} (P(brillo|cara), P(brillo|cruz)) =$$

$$\arg \max_{cara,cruz} (0'9, 0'6) = cara$$

Clasificador MAP de Fuerza Bruta

- Caso general:
  1. Para cada hipótesis—clase $h \in H$, calcular la probabilidad a posteriori:
  $$P(h|\vec{D}) = \frac{P(\vec{D}|h)P(h)}{P(\vec{D})}$$
  2. Dar como salida la hipótesis/clase con la mayor probabilidad a posteriori:
  $$h_{MAP} = \arg \max_{h \in H} P(h|\vec{D}) =$$
- Dos hipótesis/clases:

$$g(\vec{D}) = P(\vec{D}|h_1)P(h_1) - P(\vec{D}|h_2)P(h_2)$$

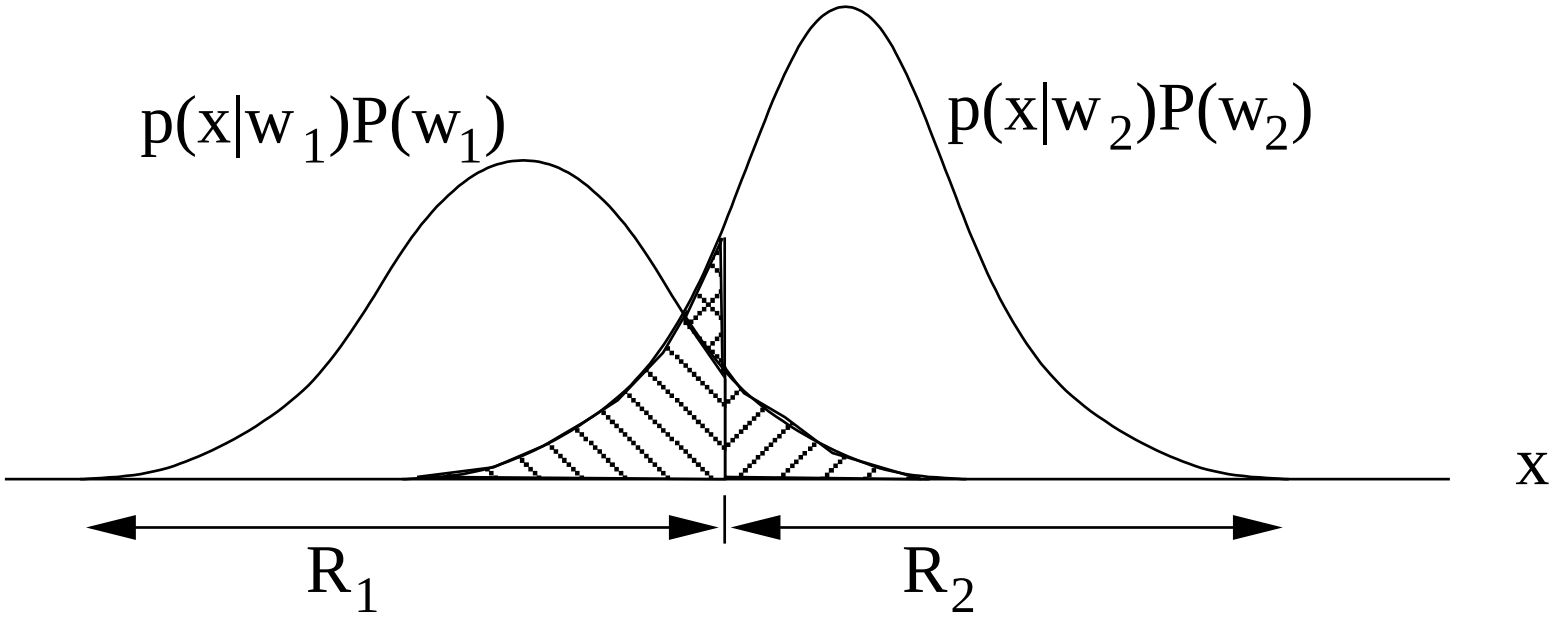
$$h_{MAP} = \begin{cases} h_1 & g(x) \geq 0 \\ h_2 & g(x) < 0 \end{cases} \quad (3)$$

Probabilidades de Error: Ejemplo

- Un clasificador de, por ejemplo, dos categorías, divide el espacio en dos regiones, $\mathcal{R}_1$ para la categoría h_1, y $\mathcal{R}_2$ para la categoría h_2.
- Errores de clasificación de una instancia $\vec{x}$:
  1. $\vec{x}$ pertenece a categoría w_1 pero cae en la región $\mathcal{R}_2$
  2. $\vec{x}$ pertenece a categoría w_2 pero cae en la región $\mathcal{R}_1$
- Probabilidad de Error de un clasificador MAP:

$$\begin{aligned} P(error) &= P(\vec{x} \in \mathcal{R}_2, w_1) + P(\vec{x} \in \mathcal{R}_1, w_2) = \\ & P(\vec{x} \in \mathcal{R}_2 | w_1) P(w_1) + P(\vec{x} \in \mathcal{R}_1 | w_2) P(w_2) = \\ & \int_{\mathcal{R}_2} p(\vec{x}|w_1) P(w_1) dx + \int_{\mathcal{R}_1} p(\vec{x}|w_2) P(w_2) dx \end{aligned} \quad (4)$$

Probabilidades de Error



¿Cuál es el punto de división entre las regiones que minimiza el error de clasificación?

Fronteras de Decisión

- Dadas dos clases, ω_1 y ω_2, tenemos sus funciones discriminantes:

$$g_i(\vec{x}) = p(\vec{x}|\omega_i)P(\omega_i)$$

$$g_i(\vec{x}) = \log p(\vec{x}|\omega_i) + \log P(\omega_i)$$

- Frontera de decisión:

$$g_1(\vec{x}) = g_2(\vec{x})$$

Función de Densidad Normal

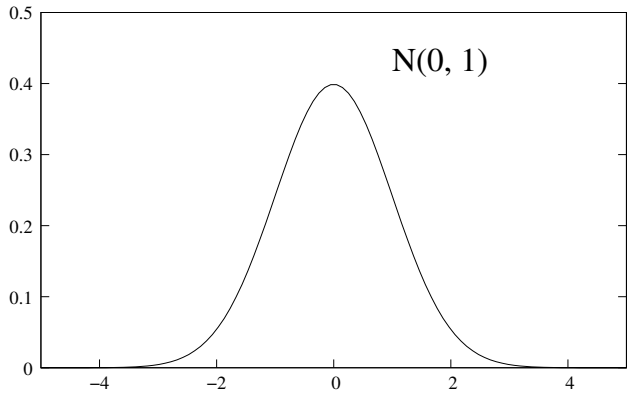
- Función de densidad normal unidimensional:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = N(\mu, \sigma^2) \qquad (5)$$

Donde:

$$E[x] = \int_{-\infty}^{\infty} x p(x) dx = \mu$$

$$E[(x-\mu))^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx = \sigma^2$$



- Con un factor de confianza de aproximadamente el 95 %, $|x - \mu| \leq 2\sigma$

Función de Densidad Normal

- Función de densidad normal multidimensional:

$$p(\vec{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\vec{x}-\vec{\mu})^t \Sigma^{-1}(\vec{x}-\vec{\mu})] = N[\mu, \Sigma] \quad (6)$$

Donde:

- $\vec{\mu}$: vector de medias
- $\vec{\mu} = E[\vec{x}]$; $\mu_i = E[x_i]$
- Σ: matriz de covarianzas $d \times d$
- $\Sigma = E[(\vec{x}-\vec{\mu})(\vec{x}-\vec{\mu})^t]$; $\sigma_{ij} = E[(x_i-\mu_i)(x_j-\mu_j)]$
- σ_{ii}: varianza de x_i
- Si $\sigma_{ij} = 0$, para $i \neq j$, entonces x_i es estadísticamente independiente de x_j
- $|\Sigma|$: determinante de Σ
- $(\vec{x}-\vec{\mu})^t$: transpuesta de $(\vec{x}-\vec{\mu})$

Función Discriminante de una Densidad Normal

- Recordamos la función discriminante:

$$g_i(\vec{x}) = \log p(\vec{x}|\omega_i) + \log P(\omega_i)$$

- Si asumimos que $p(\vec{x}|\omega_i) = N(\vec{\mu}_i, \Sigma_i)$:

$$g_i(\vec{x}) = -\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t\Sigma_i^{-1}(\vec{x}-\vec{\mu}_i)-\frac{d}{2}\log 2\pi-\frac{1}{2}\log|\Sigma_i|+\log P(\omega_i) \quad (7)$$

Caso Particular: $\Sigma_i = \vec{\sigma}^2 I$

- Si $\Sigma_i = \vec{\sigma}^2 I$:
  - Las características son estadísticamente independientes
  - Todas las características tienen la misma varianza, σ^2
  - $|\Sigma_i| = \sigma^{2d}$
  - $\Sigma_i^{-1} = (1/\sigma^2) I$
- Entonces:

$$g_i(\vec{x}) = -\frac{||\vec{x} - \vec{\mu}_i||^2}{2\sigma^2} + \log P(\omega_i) \tag{8}$$

Donde $||.||$ es la norma euclídea:

$$||\vec{x} - \vec{\mu}_i||^2 = (\vec{x} - \vec{\mu}_i)(\vec{x} - \vec{\mu}_i)^t = \sum_{j=1}^{d} (x_j - \mu_{ij})^2$$

Caso Particular: $\Sigma_i = \vec{\sigma}^2 I$

- Si desarrollamos la función discriminante:

$$g_i(\vec{x}) = -\frac{1}{2\sigma^2}[\vec{x}^t\vec{x} - 2\vec{\mu}_i^t\vec{x} + \vec{\mu}_i^t\vec{\mu}_i] + \log P(\omega_i) \qquad (9)$$

- De donde se deriva un discriminador lineal (dado que $\vec{x}^t\vec{x}$ es independiente de i):

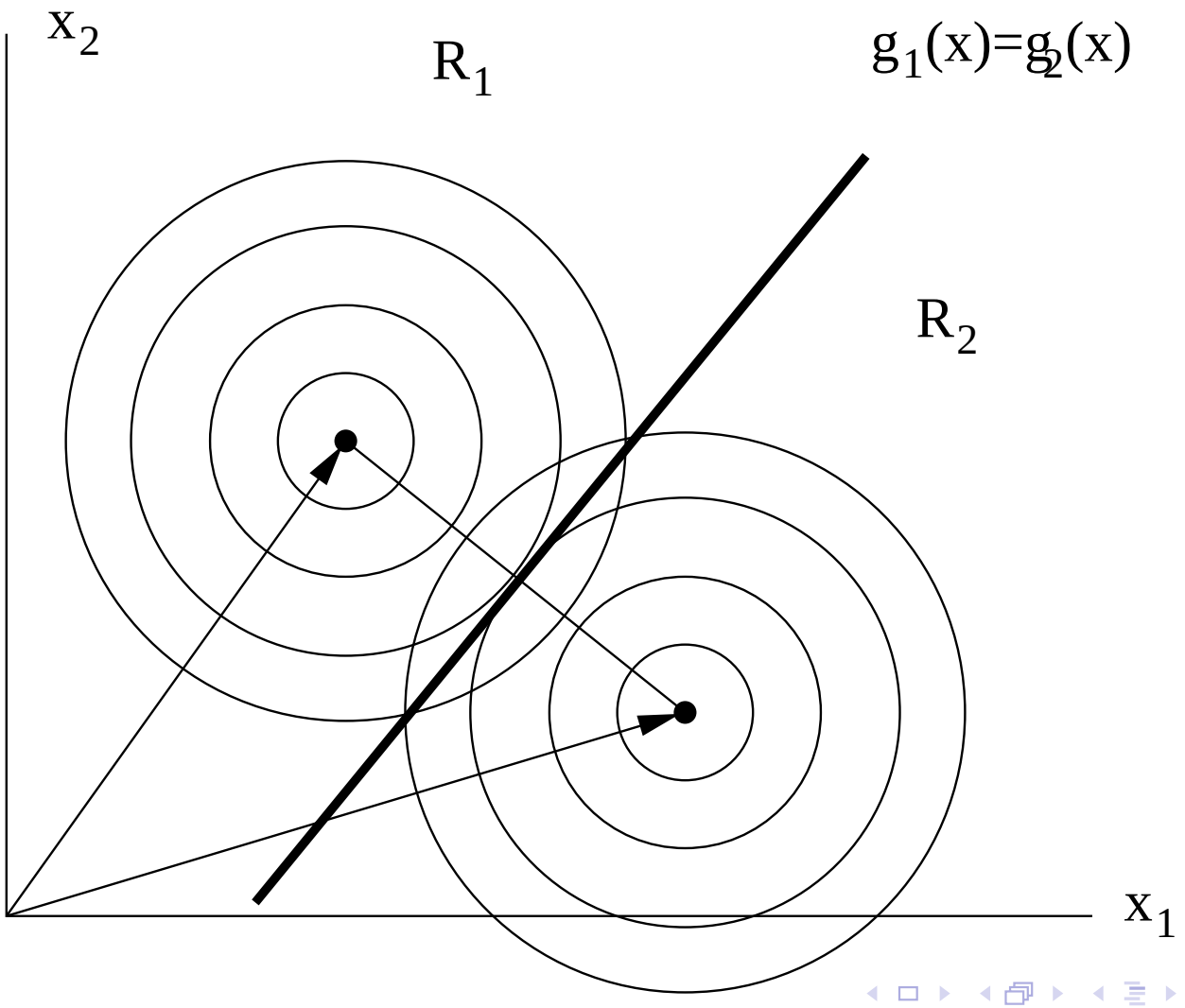
$$g_i(\vec{x}) = \vec{w}_i^t\vec{x} + w_{i0} \qquad (10)$$

Donde:

$$\vec{w}_i = \frac{1}{\sigma^2}\vec{\mu}_i$$

$$w_{i0} = -\frac{1}{2\sigma^2}\vec{\mu}_i^t\vec{\mu}_i + \log P(\omega_i)$$

Ejemplo



Estimación de Parámetros y Aprendizaje Supervisado

- Hemos visto que se pueden construir clasificadores óptimos si se conocen las probabilidades a priori, $P(\omega_j)$, y las densidades de clases condicionales, $p(\vec{x}|\omega_j)$
- Desafortunadamente, esas probabilidades son raramente conocidas
- En cambio, en la mayoría de las ocasiones se dispone de cierto conocimiento del modelo, así como de un número de ejemplos representativos de dicho modelo
- Por tanto, una buena aproximación es utilizar el conocimiento del dominio y los ejemplos para diseñar el clasificador:
  - Probabilidades a priori parece sencillo
  - Para las probabilidades a posteriori, se necesita demasiada información, sobre todo cuando crece la dimensión de los datos de entrada.

Estimación de Parámetros y Aprendizaje Supervisado

- Utilizar conocimiento del problema para parametrizar las funciones de densidad
  - Asumir que la función de densidad sigue una distribunción dada, por ejemplo $N(\vec{\mu}_j, \Sigma_j)$
  - Traspasar el problema de aproximar la función $p(\vec{x}|\omega_j)$ a estimar los parámetros $\vec{\mu}_j$ y Σ_j.
  - Añadir más simplificaciones, por ejemplo, que Σ_j es conocida.

Estimación de Máxima Verosimilitud (*Maximum Likelihood Estimation*)

- Suponer que podemos separar todas las instancias de acuerdo con su clase, de forma que generamos c conjuntos de ejemplo, $\chi_1, \ldots, \chi_c$
- Los ejemplos en χ han sido generados independientes siguiendo una distribución $p(\vec{x}|\omega_j)$
- Asumimos que $p(\vec{x}, \omega_j)$ se puede parametrizar unívocamente por un vector de parámetros $\vec{\theta_j}$
  - Por ejemplo, podemos asumir que $p(\vec{x}, \omega_j) \sim N(\vec{\mu_j}, \Sigma_j)$, donde $\vec{\theta_j} =< \vec{\mu_j}, \Sigma_j >$
  - Esa dependencia de $p(\vec{x}|\omega_j)$ con $\vec{\theta_j}$ la representamos explícitamente con $p(\vec{x}|\omega_j, \vec{\theta_j})$
- Objetivo: utilizar los conjuntos de ejemplos $\chi_1, \ldots, \chi_c$ para estimar $\vec{\theta_1}, \ldots, \vec{\theta_c}$

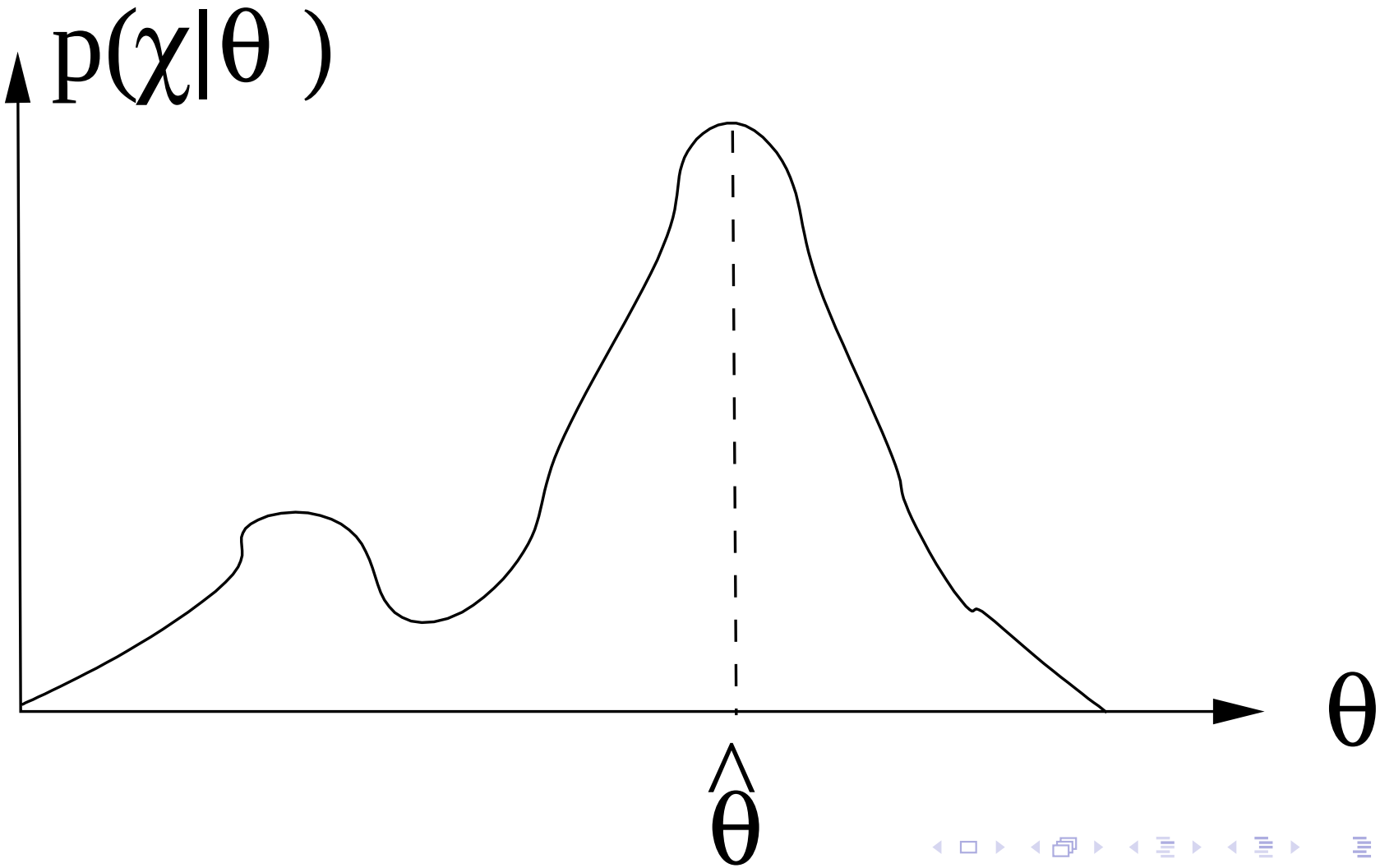
Estimador de Máxima Verosimilitud

- Idea: utilizar el conjunto de ejemplos χ, generados independientemente siguiendo la densidad de probabilidad $p(\chi|\vec{\theta})$, para estimar el vector de parámetros desconocido $\vec{\theta}$.
- Si que χ contiene n ejemplos, $\chi = \{\vec{x}_1, \ldots, \vec{x}_n\}$, dado que fueron generados independientemente:

$$p(\chi|\vec{\theta}) = \prod_{k=1}^{n} p(\vec{x}_k|\vec{\theta}) \tag{11}$$

- Vista como una función, $p(\chi|\vec{\theta})$ puede ser denominada la probabilidad de $\vec{\theta}$ dado el conjunto χ
- El estimador de verosimilitud probabilidad de $\vec{\theta}$ es, por definición, el valor $\hat{\theta}$ que maximiza $p(\chi|\vec{\theta})$

Estimador de Máxima Verosimilitud: Ejemplo



Estimador de Máxima Verosimilitud para una Densidad de Probabilidad Normal, dada Σ

- Como Σ viene dada, intentamos estimar $\hat{\mu}$
- Dado que la función logaritmo es monótona creciente, el estimador de máxima verosimilitud coincide con el estimador de máxima verosimilitud de su logaritmo

$$p(\vec{x_k}|\vec{\mu}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp[-\frac{1}{2}(\vec{x}_k - \vec{\mu})^t \Sigma^{-1}(\vec{x}_k - \vec{\mu})]$$
$$\log p(\vec{x_k}|\vec{\mu}) = -\frac{1}{2}(\vec{x}_k - \vec{\mu})^t \Sigma^{-1}(\vec{x}_k - \vec{\mu}) - \frac{d}{2}\log 2\pi - \frac{1}{2}\log|\Sigma|$$
$$\nabla_{\vec{\mu}} \log p(\vec{x_k}|\vec{\mu}) = \Sigma^{-1}(\vec{x}_k - \vec{\mu}) \quad (12)$$

$$\nabla_{\vec{\mu}} p(\chi|\vec{\theta}) = \nabla_{\vec{\mu}} \prod_{k=1}^{n} p(\vec{x}_k|\vec{\theta}) =$$
$$\sum_{k=1}^{n} \nabla_{\vec{\mu}} \log p(\vec{x_k}|\vec{\mu}) = \quad (13)$$
$$\sum_{k=1}^{n} \Sigma^{-1}(\vec{x}_k - \vec{\mu})$$

Estimador de Máxima Verosimilitud para una Densidad de Probabilidad Normal, dada Σ

- Ahora igualamos a 0 para obtener el máximo:

$$\sum_{k=1}^{n} \Sigma^{-1}(\vec{x}_k - \hat{\mu}) = 0 \tag{14}$$

$$\hat{\mu} = \frac{1}{n}\sum_{k=1}^{n} \vec{x}_k \tag{15}$$

- Igualmente, se puede calcular el estimador de la matriz de covarianzas cuando dicha matriz es desconocida:

$$\hat{\Sigma} = \frac{1}{n}\sum_{k=1}(\vec{x_k} - \hat{\mu})(\vec{x_k} - \hat{\mu})^t \tag{16}$$

Clasificador Bayesiano

- Necesitamos aprender las probabilidades a posteri, $P(\omega_i|\vec{x})$
- Por teorema de Bayes:

$$P(\omega_i|\vec{x}) = \frac{p(\vec{x}|\omega_i)P(\omega_i)}{P(\vec{x})} = \frac{P(\vec{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\vec{x}|\omega_i)P(\omega_j)} \quad (17)$$

- Las probabilidades anteriores son desconocidas, pero:
  - Diponemos de conocimiento del dominio que nos permite parametrizar esas densidades de probabilidad (por ejemplo, que siguen una distribución normal)
  - Diponemos de un conjunto de entrenamiento, χ, del que podemos aprender los parámetros de las funciones de densidad

Clasificador Bayesiano (caso continuo)

- La regla de Bayes para clasificación desde ejemplos queda como:

$$P(\omega_i|\vec{x},\chi) = \frac{p(\vec{x}|\omega_i,\chi)P(\omega_i|\chi)}{P(\vec{x}|\chi)} = \frac{p(\vec{x}|\omega_i,\chi)P(\omega_i|\chi)}{\sum_{j=1}^{c} p(\vec{x}|\omega_j,\chi)P(\omega_j|\chi)} \quad (18)$$

- Separando las instancias de entrenamiento de χ en c conjuntos, $\chi_1,\ldots,\chi_c$, y asumiendo que las probabilidades a priori son conocidas:

$$P(\omega_i|\vec{x},\chi) = \frac{p(\vec{x}|\omega_i,\chi_i)P(\omega_i)}{\sum_{j=1}^{c} p(\vec{x}|\omega_j,\chi_j)P(\omega_j)} \quad (19)$$

- Por tanto, el clasificador Bayesiano se define como:

$$Bayes(\vec{x}) = \omega = \arg_{\omega_i} \max p(\vec{x}|\omega_i,\chi_i)P(\omega_i) \quad (20)$$

El Caso Discreto

- Para toda clase ω_i, $P(\omega_i|\chi) = \frac{|\chi_i|}{|\chi|}$
- Para toda posible instancia $\vec{x}$
  - Sea $\mathcal{M}_i$ el conjunto de todas las ocurrencias de $\vec{x}$ en χ_i
  - $p(\vec{x}|\omega_i, \chi_i) = \frac{|\mathcal{M}_i|}{|\chi_i|}$
  - $Bayes(\vec{x}) = \arg_{\omega_i} \max |\mathcal{M}_i|$
- El problema de la dimensionalidad:
  - Cada ejemplo $\vec{x}$ debe aparencer en χ_i un número suficientemente grande de veces como para obtener estadísticas significativas.
  - Si la dimensión de $\vec{x}$ crece, el número de posibles valores de $\vec{x}$ crece exponencialmente, haciendo el problema intratable
  - ¿Qué ocurre si el nuevo ejemplo a clasificar, $\vec{x}$, no se había dado en χ?

Ejemplo: Jugar al tenis

Day	outlook	temperature	humidity	windy	play
D1	sunny	hot	high	weak	no
D2	sunny	hot	high	strong	no
D3	overcast	hot	high	weak	yes
D4	rainy	mild	high	weak	yes
D5	rainy	cool	normal	weak	yes
D6	rainy	cool	normal	strong	no
D7	overcast	cool	normal	strong	yes
D8	sunny	mild	high	weak	no
D9	sunny	cool	normal	weak	yes
D10	rainy	mild	normal	weak	yes
D11	sunny	mild	normal	strong	yes
D12	overcast	mild	high	strong	yes
D13	overcast	hot	normal	weak	yes
D14	rainy	mild	high	strong	no

Solución de la Clasificación Bayesiana

- Consulta: (outlook=sunny, Temperature=cool, Humidity=high, Wind=strong)
- Clasificador Bayesiano:

$$Bayes(\vec{x}) = \omega = \arg_i \max p(\vec{x}|\omega_i, \chi_i)P(\omega_i|\chi) = \omega = \arg_{yes,no} \max(p(\vec{x}|\omega_{yes}, \chi_{yes})P(\omega_{yes}|\chi), p(\vec{x}|\omega_{no}, \chi_{no})P(\omega_{no}|\chi)) \quad (21)$$

- Probabilidades a priori:

$$P(\omega_{yes}) = \frac{9}{14} \quad P(\omega_{no}) = \frac{5}{14} \quad (22)$$

- probabilidades a posteriori:

$$p(< sunny, cool, high, strong > |\omega_{yes}, \chi_{yes}) = ?? \quad p(< sunny, cool, high, strong > |\omega_{no}, \chi_{no}) = ?? \quad (23)$$

El Clasificador *Naive Bayes*

- Naive Bayes asume independencia lineal entre los distintos atributos
- Eso implica que:

$$p(\vec{x}|\omega_i, \chi_i) = p(< x_1, x_2, \ldots, x_k > |\omega_i, \chi_i) = \prod_{k=1}^{K} p(x_k|\omega_i, \chi_i) \quad (24)$$

- Por tanto:

$$NaiveBayes(\vec{x}) = \omega = \arg_i \max \prod_{k=1}^{K} p(x_k|\omega_i, \chi_i) P(\omega_i) \quad (25)$$

Solución al Ejemplo con Naive Bayes

- Consulta: (outlook=sunny, Temperature=cool, Humidity=high, Wind=strong)
- Probabilidades a priori:

$$\begin{aligned} P(\omega_{yes}|\chi) &= \frac{9}{14} = 0{,}64 \\ P(\omega_{no}|\chi) &= \frac{5}{14} = 0{,}35 \end{aligned} \tag{26}$$

Solución al Ejemplo con Naive Bayes

- Probabilidades a posteriori:

$$\begin{aligned} p(< outlook = sunny > |\omega_{yes}, \chi_{yes}) &= \frac{2}{9} = 0{,}22 \\ p(< outlook = sunny > |\omega_{no}, \chi_{no}) &= \frac{3}{5} = 0{,}6 \\ p(< Temperature = cool > |\omega_{yes}, \chi_{yes}) &= \frac{3}{9} = 0{,}33 \\ p(< Temperature = cool > |\omega_{no}, \chi_{no}) &= \frac{1}{5} = 0{,}2 \\ p(< Humidity = high > |\omega_{yes}, \chi_{yes}) &= \frac{3}{9} = 0{,}33 \\ p(< Humidity = high > |\omega_{no}, \chi_{no}) &= \frac{4}{9} = 0{,}44 \\ p(< Wind = strong > |\omega_{yes}, \chi_{yes}) &= \frac{3}{9} = 0{,}33 \\ p(< Wind = strong > |\omega_{no}, \chi_{no}) &= \frac{3}{5} = 0{,}6 \end{aligned} \tag{27}$$

- Entonces:

$$\begin{aligned} &P(yes)p(sunny|yes)p(cool|yes)p(high|yes)p(strong|yes) = \\ &0{,}64 \times 0{,}22 \times 0{,}33 \times 0{,}33 \times 0{,}33 = 0{,}005 \\ &P(no)p(sunny|no)p(cool|no)p(high|no)p(strong|no) = \\ &0{,}35 \times 0{,}6 \times 0{,}2 \times 0{,}44 \times 0{,}6 = 0{,}01 \\ &NaiveBayes(< sunny, cool, high, strong >) = no \end{aligned} \tag{28}$$

Resumen

- Teoría Bayesina nos da mecanismos para generar clasificadores basándose en las probabilidades a priori y las distribuciones de probabilidad a posteriori
- Las probabilidades pueden ser desconocidas: aprendizaje paramétrico
- Estimación de parámetros en densidades de probabilidad conocidas
- Clasificador Bayesiano
- Naive Bayes

Bibliografía

- Pattern Classification and Scene Analysis, Duda and Hart. Capítulo 2
- Tom Mitchell. Machine Learning, capítulo 6, McGraw-Hill 1997