INCERTIDUMBRE

PROBABILIDADES Y PLAUSIBILIDAD

$\mathbb{P}(A|I) :=$ probabilidad de una aserción A, dada la información

$\mathbb{P}(\text{Verdad}|I) = 1 \qquad \mathbb{P}(\text{Falso}|I) = 0$

$0 \leq \mathbb{P}(A|I) \leq 1$

Regla de la suma: $\mathbb{P}(A|I) + \mathbb{P}(\bar{A}|I) = 1$

Regla del producto: $\mathbb{P}(A,B|I) = \mathbb{P}(B|A,I) \cdot \mathbb{P}(A|I)$

Teorema de Bayes: $\mathbb{P}(B|A,I) = \dfrac{\mathbb{P}(A|B,I) \cdot \mathbb{P}(B|I)}{\mathbb{P}(A|I)}$

PROBABILIDAD CONDICIONAL

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$$

$$\left.\begin{array}{l} \mathbb{P}(A,B) = \mathbb{P}(A|B) \cdot \mathbb{P}(B) \\ \quad \parallel \\ \mathbb{P}(B,A) = \mathbb{P}(B|A) \cdot \mathbb{P}(A) \end{array}\right\} \Longrightarrow \overbrace{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A) \cdot \mathbb{P}(A)}{\mathbb{P}(B)}}^{\text{TEOREMA DE BAYES}}$$

verosimilitud de hipótesis

$$\mathbb{P}(\text{clase}|\text{datos}) = \frac{\mathbb{P}(\text{datos}|\text{clase}) \cdot \mathbb{P}(\text{clase})}{\mathbb{P}(\text{datos})}$$

prior

evidencia

posterior

Terminología

clase = hipótesis = etiqueta = variables → dependientes → a predecir

datos = atributos = características = variables → independientes → predictoras

INFERENCIA A PARTIR DE LOS DATOS

- <u>CLASIFICADOR MÁXIMA VEROSIMILITUD</u> (ML): selecciona la hipótes que maximiza la verosimilitud de la hipótesis dados los datos.

$$H^* = \arg\max_H \mathbb{P}(D|H)$$

ML no usa información a prioris (equivalente a asumir un priori constante)

- <u>CLASIFICADOR MÁXIMA PROBABILIDAD A POSTERIORI</u> (MAP): selecciona la hipótesis que maximiza la probabilidad a posteriori.

$$H^* = \arg\max_H \mathbb{P}(H|D) = \arg\max_H \frac{\mathbb{P}(D|H).\mathbb{P}(H)}{\underbrace{\mathbb{P}(D)}_{\text{igual } \forall H \text{ hipótesis}}} = \arg\max_H \mathbb{P}(D|H).\mathbb{P}($$

Este es el llamado CLASIFICADOR DE BAYES

- <u>INFERENCIA BAYESIANA</u>: promediar sobre todas las hipótesis con probabilidades.

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H).\mathbb{P}(H)}{\mathbb{P}(D)}$$

$$\mathbb{E}[f(H)] = \int dH\, f(H).\mathbb{P}(H|D)$$

- <u>CLASIFICADOR A PRIORI DE LA CLASE</u>: selecciona la hipótesis que maximiza la probabilidad de la clase

$$H^* = \arg\max_H \mathbb{P}(H)$$

⊛ Prioris uniformes ⟹ Clasificador ML = Clasificador MAP (Bayes)

⊛ Bayes es óptimo (minimiza el error)

REDES BAYESIANAS ("Modelos gráficos")

Regla de la cadena:

$$\mathbb{P}(x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}) = \mathbb{P}(x^{(1)} | x^{(2)}, x^{(3)}, x^{(4)}) \cdot \mathbb{P}(x^{(2)}, x^{(3)}, x^{(4)}) =$$

$$= \mathbb{P}(x^{(1)} | x^{(2)}, x^{(3)}, x^{(4)}) \cdot \mathbb{P}(x^{(2)} | x^{(3)}, x^{(4)}) \cdot \mathbb{P}(x^{(3)}, x^{(4)}) =$$

$$= \mathbb{P}(x^{(1)} | x^{(2)}, x^{(3)}, x^{(4)}) \cdot \mathbb{P}(x^{(2)} | x^{(3)}, x^{(4)}) \cdot \mathbb{P}(x^{(3)} | x^{(4)}) \cdot \mathbb{P}(x^{(4)})$$

Representación gráfica de la regla de la cadena:

Nodos: variables

Aristas dirigidas: dependencias

Una flecha hacia un nodo dado corresponde a variables dependientes de la variable que origina la flecha.

Ojo: la representación gráfica no es única

Interpretación grafo:

Nodo i: variable $x^{(i)}$

$\pi(x^{(i)})$: padres del nodo $x^{(i)}$

$$\mathbb{P}(x^{(1)}, x^{(2)}, \dots, x^{(n)}) = \prod_{i=1}^{n} \mathbb{P}(x^{(i)} | \pi(x^{(i)}))$$

Nota: Si $\pi(x^{(i)}) = \emptyset \implies \mathbb{P}(x^{(i)} | \pi(x^{(i)})) = \mathbb{P}(x^{(i)})$

INDEPENDENCIA CONDICIONAL

A es condicionalmente independiente de B dado C si:

$$\mathbb{P}(A|B,C) = \mathbb{P}(A|C)$$

Que es equivalente a:

$$\mathbb{P}(A,B|C) = \mathbb{P}(A|C) \cdot \mathbb{P}(B|C)$$

CLASIFICADOR DE NAÏVE-BAYES

El Clasificador de Naïve-Bayes asume que todos los atributos son condicionalmente independientes dada la clase.

Atributos: $x = \{x^{(1)}, x^{(2)}, \dots, x^{(D)}\}$

Clase: $c \in \{1, 2, \dots, C\}$

$$\mathbb{P}(c|x) = \frac{\mathbb{P}(x|c) \cdot \mathbb{P}(c)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x^{(1)}, x^{(2)}, \dots, x^{(D)}|c) \cdot \mathbb{P}(c)}{\mathbb{P}(x)} \approx$$

$$\approx \frac{\mathbb{P}(x^{(1)}|c) \cdot \mathbb{P}(x^{(2)}|c) \cdot \dots \cdot \mathbb{P}(x^{(D)}|c) \cdot \mathbb{P}(c)}{\sum_{c'=1}^{C} \mathbb{P}(x^{(1)}|c') \cdot \mathbb{P}(x^{(2)}|c') \cdot \dots \cdot \mathbb{P}(x^{(D)}|c')}$$

ESTIMADOR DE LAPLACE

- Estimación de probabilidades con frecuencias:

$$\mathbb{P}_i = \frac{n_i}{n_{total}} \qquad i = 1, 2, \dots, K$$

- Estimador de Laplace: añadir ejemplos ficticios

$$\mathbb{P}_i = \frac{n_i + \mu/K}{n_{total} + \mu} \qquad i = 1, 2, \dots, K$$

Ventajas:
- evita estimaciones nulas para las probabilidades
- estimaciones más robustas
- asintóticamente pequeño

INCERTIDUMBRE

PROBABILIDAD (ENFOQUE BAYESIANO)

$$\underset{n=1,\dots,N}{X_n} = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \xi_{in}$$

ξ_{in} V.A. independientes e idénticamente distribuida[s] con media cero.

Varianza finita : $\sigma^2 = \mathbb{E}\left[\xi_{in}^2\right] < \infty$

TCL: $X_n \underset{N\to\infty}{\sim} N(0,\sigma^2)$

PROBABILIDAD CONDICIONAL

$$P(A|B) = \frac{\mathbb{P}(A,B)}{\mathbb{P}(B)}$$

$$\left.\begin{array}{l} \mathbb{P}(A,B) = \mathbb{P}(A|B).\,\mathbb{P}(B) \\ \quad \parallel \\ \mathbb{P}(B,A) = \mathbb{P}(B|A).\,\mathbb{P}(A) \end{array}\right\} \Rightarrow$$

TEOREMA DE BAYES

$$\boxed{\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A).\,\mathbb{P}(A)}{\mathbb{P}(B)}}$$

verosimilitud de hipótesis

prior

$$\mathbb{P}(clase|datos) = \frac{\mathbb{P}(datos|clase)\,.\,\mathbb{P}(clase)}{\mathbb{P}(datos)}$$

posterior

evidencia

Terminología:

clase = hipótesis = etiqueta = variables → dependientes / → a predecir

datos = atributos = características = variables → independientes / → predictoras

Ejemplo:

Hipótesis (clase) → llueve / → no llueve

PRIOR $\begin{cases} \mathbb{P}(\text{"llueve"}) = 20\% = 0'2 \\ \mathbb{P}(\text{"no llueve"}) = 80\% = 0'8 \ (= 1 - 0'2 \end{cases}$

Evidencia = $\mathbb{P}(\text{"paraguas"})$
Datos = "alguien llega con paraguas"

VEROSIM: $\mathbb{P}(\text{"paraguas"} \mid \text{"no llueve"}) = 0'1$

$$\mathbb{P}(\text{"paraguas"} \mid \text{"llueve"}) = \frac{\mathbb{P}(\text{"llueve"} \mid \text{"paraguas})\ \mathbb{P}(\text{"paraguas"})}{\mathbb{P}(\text{"llueva"})}$$

sin datos suficientes

↓ nos lo dan

$$\mathbb{P}(\text{"paraguas"} \mid \text{"llueve"}) = 0'7$$

CLASIFICADOR A: basado en PRIORES (máximo entre priores)
CLASIFICADOR B: máxima verosimilitud (ML ≡ max likelihood)
CLASIFICADOR C: MAP (máximo entre posteriores)

Observamos "Una persona entra con paraguas" → DECISIÓN

- $\xrightarrow{\text{clasificador A}}$ "no llueve"
- $\xrightarrow{\text{MaxVero}}$ "llueve" $\left(\mathbb{P}(\text{"paraguas"} \mid \text{"llueve"}) > \mathbb{P}(\text{"paraguas"} \mid \text{"no llueve"})\right)$
- $\xrightarrow{\text{MAP}}$ "llueve" $\left(\mathbb{P}(\text{"llueve"} \mid \text{"paraguas"}) > \mathbb{P}(\text{"no llueve"} \mid \text{"paraguas"})\right)$

Calculamos POSTERIOR:

$$\mathbb{P}(\text{"llueve"} \mid \text{"paraguas"}) = \frac{\mathbb{P}(\text{"paraguas"} \mid \text{"llueve"}) . \mathbb{P}(\text{"llueve"})}{\mathbb{P}(\text{"paraguas"})} = \frac{0'7 . 0'2}{\mathbb{P}(\text{"paraguas"})} = 0'64$$

$$\mathbb{P}(\text{"no llueve"} \mid \text{"paraguas"}) = \frac{\mathbb{P}(\text{"paraguas"} \mid \text{"no llueve"}) . \mathbb{P}(\text{"no llueve"})}{\mathbb{P}(\text{"paraguas"})} = \frac{0'1 . 0'8}{\mathbb{P}(\text{"paraguas"})} = 0'3$$

Observación: $\mathbb{P}(\text{"paraguas"}) = \mathbb{P}(\text{"paraguas"} \mid \text{"llueve"}) . \mathbb{P}(\text{"llueve"}) + \mathbb{P}(\text{"paraguas"} \mid \text{"no llueve"}) . \mathbb{P}(\text{"n}$

$= 0'22$

OJO! → los costes asimétricos varía[n] el umbral

ENFOQUE BAYESIANO

$H \in \{H_1, H_2, \dots, H_c\}$ con una probabilidad $\{\mathbb{P}(H_i|D),\ i=1,\dots,C\}$

$$\mathbb{E}_H[f(H)] = \sum_{i=1}^{C} f(H_i) \cdot \mathbb{P}(H_i|D)$$

(*) Estamos suponiendo costes simétricos en todos estos clasificadores.

NAÏVE - BAYES

$\to$ probabilidad en espacio D-dimensional $\to$ D prob. en 1 dimensió

$$\mathbb{P}(C|\tilde{x}) = \underbrace{\frac{\mathbb{P}(\tilde{x}|C) \cdot \mathbb{P}(C)}{\mathbb{P}(\tilde{x})}}_{\text{exacto si las estimaciones de prob. son correctas}} \approx \frac{\mathbb{P}(x_1|C) \cdot \mathbb{P}(x_2|C) \cdot \dots \cdot \mathbb{P}(x_D|C) \cdot \mathbb{P}(C)}{\text{Normalización}}$$

Normalización $\uparrow$ en la misma aprox.

$$\tilde{x} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_D \end{pmatrix}$$

$\uparrow$ ECTOR EATORIO

$$\mathbb{P}(\tilde{x}|C) \approx \mathbb{P}(x_1|C) \cdot \mathbb{P}(x_2|C) \cdot \dots \cdot \mathbb{P}(x_D|C)$$

$\uparrow$ "los atributos son INDEPENDIENTEMENTE CONDICIONADOS a C"

OJO:

$$\mathbb{P}(A) = \mathbb{P}(A, B) + \mathbb{P}(A, \neg B)$$

$\mathbb{P}(\text{Gripe}) =$

$\cancel{\mathbb{P}(\neg G \mid \neg F, \neg T)}$

$\mathbb{P}(A|B) + \mathbb{P}(A|\neg B) \neq 1$

$\mathbb{P}(\neg G, \neg T \mid \neg F)$ $\qquad$ $\mathbb{P}(\neg G \mid \neg T, \neg F)$

$\mathbb{P}(\neg G, \neg T|\neg F) = \mathbb{P}(\neg G| \neg T, \neg F) \cdot \mathbb{P}(\neg T)$

$$\mathbb{P}(\bar{G}, \bar{F}, \bar{T}) + \underbrace{\mathbb{P}(\bar{G}, F, T)}_{0'1} + \underbrace{\mathbb{P}(\bar{G}, \bar{F}, T)}_{0'05} + \underbrace{\mathbb{P}(\bar{G}, F, \bar{T})}_{0'0} = \mathbb{P}(\bar{G})$$

$\mathbb{P}(\bar{G}) = 0'6$

$\mathbb{P}(\bar{G}, E) + \mathbb{P}(G, E)$

$$\Rightarrow \mathbb{P}(\bar{G}, \bar{F}, \bar{T}) = 0'6 - 0'1 - 0'05 = 0'45$$

Lo mismo para el otro