Aproximación probabilística al RP

Alfredo Cuesta Infante

E. T. S. Ingeniería Informática
Universidad Rey Juan Carlos

Master Univ. en Visión Artificial
Reconocimiento de Patrones

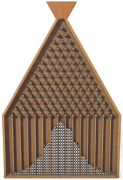
Contenidos

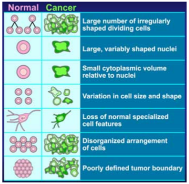
Contenidos

Eleméntos básicos de Probabilidad

Variables aleatorias

- Variables sobre las que tenemos una incertidumbre se denominan *variables aleatorias*, *v.a.*
- Pueden ser continuas o discretas.
- Sólo podemos decir de qué manera se distribuyen sus posibles valores, es decir qué probabilidad tienen.





Distribución de probabilidad

- Función que modela:
  - la masa de probabilidad de una variable discreta *PMF*
  - la densidad de probabilidad de una variable continua *PDF*
  - la masa/densidad acumulada de una variable continua *CDF*

Eleméntos básicos de Probabilidad

Ejemplos de PMFs y PDFs

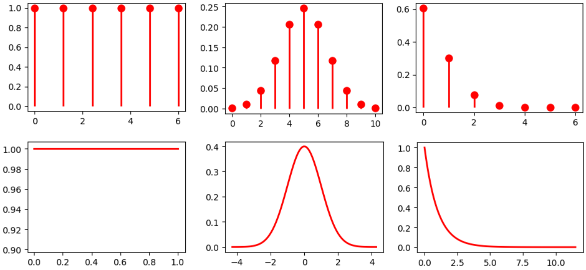


Figura: **Arriba**: PMF Uniforme, Binomial y Poisson. **Abajo**: PDF Uniforme, Normal o Gaussiana, Exponencial. [Fuente: Original de A. Cuesta]

Eleméntos básicos de Probabilidad

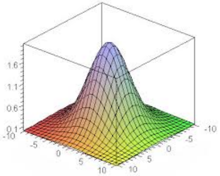
Resumen

Discreta	**Continua**
pmf $p(x = x^*) \geq 0$	**pdf** $p(x = x^*) = 0$
cdf	
$P(a \leq x \leq b) = \sum_{x=a}^{x=b} p(x)$	$P(a \leq x \leq b) = \int_{x=a}^{x=b} p(x)$
$P(-\infty < x < +\infty) = 1$	
Ejemplos	
Uniforme Binomial Poisson Dirichlet $\vdots$	Uniforme Normal o Gaussiana Exponencial Weibull $\vdots$

Eleméntos básicos de Probabilidad

Probabilidad Conjunta

- Dos o más *v.a.* tendrán, en general, una distribución de probabilidad **conjunta**.



Marginales

Sea la distribución conjunta de $(x, y) = P(x, y)$

- Distribución marginal de $x = P(x)$
  - *v.a.* discretas: $P(x) = \sum_{y=-\infty}^{y=+\infty} P(x, y)$.
  - *v.a.* continuas: $P(x) = \int_{y=-\infty}^{y=+\infty} P(x, y)$.
- Distribución marginal de $y = P(y)$
  - *lo mismo que para x*

Eleméntos básicos de Probabilidad

Distribución condicionada

- Supongamos dos *v.a.* distribuidas conjuntamente.
  Si conocemos el valor de una de ellas, entonces la incertidumbre la tenemos sólo sobre la otra.
  Ej. En $p(x|y)$ hay incertidumbre sobre x pero no sobre y.
- La distribución condicionada y la marginal de una *v.a.* son distribuciones univariadas, ¡pero son cosas muy diferentes!

Regla de la cadena

Ej. 2 variables:

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

Ej. 4 variables:

$$p(x, y, z, w) = p(w)p(z|w)p(y|z, w)p(x|y, z, w),$$

o también:

$$p(x, y, z, w) = p(x, w)p(y, z|x, w) = p(y)p(x, z, w|y) = \cdots$$

Eleméntos básicos de Probabilidad

Independencia

- La distribución conjunta de dos *v.a.* independientes es la multiplicación de sus distribuciones individuales. En general

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i)$$

Independencia condicional

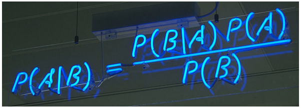
- Cuando el conocimiento de una *v.a.* 'x' convierte a otras dos 'v' y 'w' en independientes, se dice que éstas son **condicionalmente independientes**.

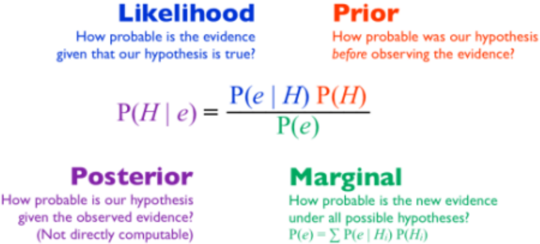
$$p(v, w|x) = p(v|x)p(w|x)$$

Teorema de Bayes

El teorema surge directamente de aplicar la regla de la cadena a la densidad conjunta.

$$\begin{cases} p(A,B) = p(B)p(A|B) & \text{y despejando } p(A|B) = p(A,B)/p(B) \\ p(A,B) = p(A)p(B|A) & \text{que introducimos arriba} \end{cases}$$





Contenidos

Aproximación probabilística al ML

Modelo

- PDF conjunta de los datos y las etiquetas $p(\mathbf{X}, \mathbf{t}; \mathbf{w})$, que dependerá del vector de parámetros $\mathbf{w}$.

Aprendizaje

- Los parámetros serán **óptimos** cuando la probabilidad conjunta de los datos y las etiquetas sea máxima.

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{X}, \mathbf{t}; \mathbf{w})$$

Inferencia

- La etiqueta estimada para un nuevo ejemplo $\mathbf{z}$ es:

$$\hat{t} = \arg\max_{t} p(t|\mathbf{z}).$$

Aplicando el teorema de Bayes:

$$\hat{t} = \arg\max_{t} \left(\frac{p(t)p(\mathbf{z}|t)}{p(\mathbf{z})} \right) = \arg\max_{t} \left(p(t)p(\mathbf{z}|t) \right).$$

Contenidos

Naive Bayes

Modelo

- Asumir 2 condiciones para simplificar la construcción de $p(\mathbf{X}, \mathbf{t}; \mathbf{w})$:
    1. El conjunto de datos es iid
       (independiente y está idénticamente distribuido)
       $$p(\mathbf{X}, \mathbf{t}) = \prod_{i=1}^{m} p(\mathbf{x}^{(i)}, t_i), \text{ y también } p(\mathbf{X}|\mathbf{t}) = \prod_{i=1}^{m} p(\mathbf{x}^{(i)}|t_i)$$
    2. Las características son condicionalmente independientes
       dada la etiqueta.
       $$p(\mathbf{x}|t) = \prod_{j=1}^{n} p(x_j|t).$$

Por tanto:

$$\begin{aligned} p(\mathbf{X}, \mathbf{t}) = p(\mathbf{t})p(\mathbf{X}|\mathbf{t}) &= \prod_{i=1}^{m} p(t_i)p(\mathbf{x}^{(i)}|t_i) \\ &= \prod_{i=1}^{m} p(t_i) \prod_{i=1}^{m}\prod_{j=1}^{n} p(x_j^{(i)}|t_i). \end{aligned}$$

Naive Bayes

Aprendizaje

- Las distribuciones del modelo son familias paramétricas $\Rightarrow p(\mathbf{X}, \mathbf{t}; \mathbf{w})$, donde $\mathbf{w}$ es el conjunto de parámetros.
- En vez de optimizar $p(\mathbf{X}, \mathbf{t}; \mathbf{w})$, vamos a optimizar

$$\ell(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \log p(\mathbf{X}, \mathbf{t}; \mathbf{w})$$

Aplicando las propiedades de los logaritmos tenemos:

$$\ell(\mathbf{w}; \mathbf{X}, \mathbf{t}) = \sum_{i=1}^{m} \log p(t_i) + \sum_{i=1}^{m} \sum_{j=1}^{n} \log p(x_j^{(i)} | t_i). \tag{1}$$

- El óptimo de la función (1) en el vector de parámetros $\mathbf{w}^*$.

$$w_k^* = \underset{w_k}{\arg \mathrm{cero}} \frac{\partial}{\partial w_k} \ell(\mathbf{w}; \mathbf{X}, \mathbf{t})$$

- Por la construcción de $\log p(\mathbf{X}, \mathbf{t}; \mathbf{w})$:

$$\mathbf{w} = (\overbrace{\eta_1, \eta_2, \ldots}^{\text{parm. de } p(t)} , \overbrace{\xi_1, \xi_2, \ldots}^{\text{parm. de } p(x_1|t)} , \overbrace{\zeta_1, \zeta_2 \ldots}^{\text{parm. de } p(x_2|t)} , \ldots)$$

Naive Bayes

Inferencia

- Aplicando la suposición 2:

$$\hat{t} = \arg\max_t \left(p(t) \prod_{i=1}^{n} p(z_i|t) \right)$$

- Añadir el logaritmo a la función que queremos optimizar para convertir productos en sumas:

$$\hat{t} = \arg\max_t \left(\log p(t) + \sum_{i=1}^{n} \log p(z_i|t) \right) \tag{2}$$

- Los parámetros de las distribuciones a priori y de verosimilitud han sido aprendidos previamente.

Vamos a ver dos ejemplos concretos de clasificador Naive Bayes (NBC):

- Para características continuas *NBC Gaussiano*
- Para características discretas *NBC Multinomial*

Resumen visual de NBC