Aprendizaje automatizado

Clasificador bayesiano ingenuo

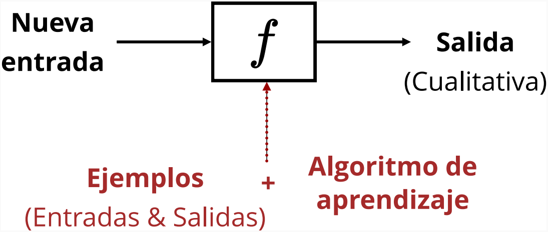
Gibran Fuentes-Pineda

Febrero 2023

Clasificación

- Salida discreta o categórica (cualitativa)
- Conjunto de ejemplos

$$\mathcal{D} = \left\{ \left(\mathbf{x}^{(1)}, y^{(1)} \right), \left(\mathbf{x}^{(2)}, y^{(2)} \right), \ldots, \left(\mathbf{x}^{(n)}, y^{(n)} \right) \right\}$$



Ejemplo de clasificación

- Clasificar sub-especias de la flor Iris basado en el ancho y largo de su pétalo

$$\mathcal{D} = \{([1.4, 0.2], 0), ([1.7, 0.4], 0), \ldots, ([4.7, 1.4], 1), ([4.5, 1.5], 1), \ldots\}$$

Ancho	Largo	Especie
1.4	0.2	Setosa
1.7	0.4	Setosa
1.5	0.1	Setosa
⋮	⋮	⋮
4.7	1.4	Versicolor
4.5	1.5	Versicolor
3.3	1.0	Versicolor
⋮	⋮	⋮

Características o atributo　　Respuesta

Setosa

Versicolor

Imagen tomada de `https://en.wikipedia.org/wiki/Iris_flower_data_set`

Clasificador bayesiano ingenuo: modelo generativo

- Modela distribución conjunta de atributos y clases $P(x_1, \ldots, x_d, y)$, asumiendo independencia condicional de los atributos dada la clase
- **Independencia condicional**: X y Y son condicionalmente independientes dado Z si

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

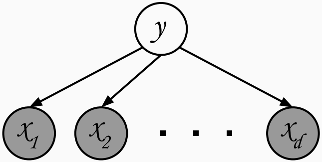
$$P(X|Y, Z) = P(X|Z)$$

- En el clasificador bayesiano ingenuo, la probabilidad conjunta está dada por

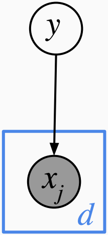
$$P(x_1, \ldots, x_d, y) = \left(\prod_{j=1}^{d} P(x_j|y = c)\right) P(y = c)$$

Clasificador bayesiano ingenuo: representación gráfica

- El clasificador bayesiano ingenuo se puede representar como un modelo gráfico probabilista simple



- De forma más compacta en notación de placas:



Clasificador bayesiano ingenuo: predicción

- Para obtener la probabilidad de cada clase para un nuevo dato $\widetilde{\mathbf{x}} = [\widetilde{x}_1, \ldots, \widetilde{x}_d]$ usamos teorema de bayes

$$P(y = c|\widetilde{x}_1, \ldots, \widetilde{x}_d) = \frac{P(\widetilde{x}_1, \ldots, \widetilde{x}_d|y = c)P(y = c)}{P(\widetilde{x}_1, \ldots, \widetilde{x}_d)}$$

- Debido a que

$$\left(\prod_{j=1}^{d} P(\widetilde{x}_j|y = c)\right) P(y = c) \propto P(y = c|\widetilde{x}_1, \ldots, \widetilde{x}_d)$$

- Podemos obtener la clase más probable como[1]:

$$\hat{y} = \arg\max_{y} \left(\prod_{j=1}^{d} P(\widetilde{x}_j|y = c)\right) P(y = c)$$

[1] En algunas aplicaciones se requiere conocer las probabilidades para la toma de decisiones, por lo que es necesario calcular $P(\widetilde{x}_1, \ldots, \widetilde{x}_d)$

Detección de correo basura:

- Considera n correos electrónicos representados como bolsas de palabras y sus correspondientes etiquetas $\mathcal{X} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(n)}, y^{(n)})\}$
    - Vectores $\mathbf{x}^{(i)} = [x_1^{(i)}, \ldots x_d^{(i)}]$
    - $x_j^{(i)}$ es el número de veces que la palabra j ocurre en el correo i ($x_j^{(i)} \in [0, 1]$ si el esquema es binario)

Bolsas de palabras

- Conjunto de documentos

  d_1 *Ella toma café y él toma mate*

  d_2 *Ella toma notas mientras toma café*

- Matriz documento-término

	café	él	ella	mate	mientras	notas	toma	y
d_1	1	1	1	1	0	0	1	1
d_2	1	0	1	0	1	1	1	0

- Bolsas de palabras
  - $d_1 = \{$café, él, ella, mate, toma, y$\}$
  - $d_2 = \{$café, ella, notas, mientras, toma$\}$

Matriz documento-término

- Conjunto de documentos

d_1 *Ella* $\underbrace{\textit{toma}}_{w_4}$ $\underbrace{\text{café}}_{w_1}$ *y* él $\underbrace{\textit{toma}}_{w_4}$ $\underbrace{\textit{mate}}_{w_2}$

d_2 *Ella* $\underbrace{\textit{toma}}_{w_4}$ $\underbrace{\textit{notas}}_{w_3}$ *mientras* $\underbrace{\textit{toma}}_{w_4}$ $\underbrace{\text{café}}_{w_1}$

- Matriz documento-término **sin palabras vacías**

	café	mate	notas	toma
d_1	1	1	0	1
d_2	1	0	1	1

- Bolsas de palabras binaria (conjunto)
  - $\mathsf{d}_1 = \{1, 2, 4\}$
  - $\mathsf{d}_2 = \{1, 3, 4\}$

Matriz documento-término

- Conjunto de documentos

d_1 $\textit{Ella } \underbrace{\textit{toma}}_{\mathsf{w}_4} \ \underbrace{\text{café}}_{\mathsf{w}_1} \ \textit{y}\ \text{él} \ \underbrace{\textit{toma}}_{\mathsf{w}_4} \ \underbrace{\textit{mate}}_{\mathsf{w}_2}$

d_2 $\textit{Ella } \underbrace{\textit{toma}}_{\mathsf{w}_4} \ \underbrace{\textit{notas}}_{\mathsf{w}_3} \ \textit{mientras} \ \underbrace{\textit{toma}}_{\mathsf{w}_4} \ \underbrace{\text{café}}_{\mathsf{w}_1}$

- Matriz documento-término **sin palabras vacías con frecuencias**

	w_1	w_2	w_3	w_4
d_1	1	1	0	2
d_2	1	0	1	2

- Bolsas de palabras con frecuencia
  - $\mathsf{d}_1 = \{1, 2, 4, 4\}$
  - $\mathsf{d}_2 = \{1, 3, 4, 4\}$

Clasificador bayesiano ingenuo

- Presuponiendo esquema binario y clasificación binaria:

$$x_j \sim Ber(q_j), j = 1, \ldots, d$$
$$y \sim Ber(q_y)$$

- Entrenamiento: se estiman los parámetros $q_1, \ldots, q_d$ condicionadas a las clases ($y = 0$ y $y = 1$) y el parámetro de la distribución a priori de la clase q_y

- Predicción: dado un nuevo documento $\widetilde{\mathbf{x}} = [\widetilde{x}_1, \ldots \widetilde{x}_d]$, podemos obtener su clase más probable usando los parámetros estimados

$$\hat{y} = \arg\max_y \left(\prod_{j=1}^{d} Ber(x_j; \hat{q}_j) \right) Ber(y; \hat{q}_y)$$