

ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL



Clasificador de Bayes Ingenuo

Andrés G. Abad, Ph.D.

Agenda

Introducción al problema de clasificación

Clasificador de bayes ingenuo (naive bayes)

Clasificador de bayes para variables categóricas

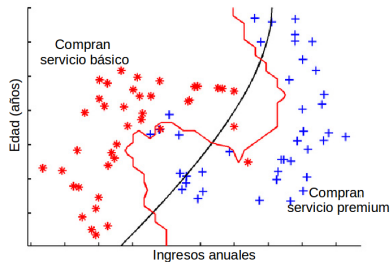
Definición del problema de clasificación I

- ▶ Un objeto $\mathbf{x} = [x_1, \dots, x_p]$, con características x_i , pertenece exactamente a una clase $c \in \{1, 2, \dots, C\}$.
- ▶ Asumimos que tenemos un conjunto de datos

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, c^{(1)}), \dots, (\mathbf{x}^{(n)}, c^{(n)})\}$$

- ▶ Buscamos una función \hat{f} que asigne $\mathbf{x}^{(i)}$ a $c^{(i)}$ lo mejor posible:

$$\hat{f} = \arg \min_f \mathbb{P}_{(\mathbf{x}, c)}[\mathbb{1}(f(\mathbf{x}) \neq c)]$$



- ▶ Objeto \mathbf{x} pertenece a una de dos clases: $\{\text{Basico}, \text{Premium}\}$
- ▶ Objeto \mathbf{x} medidos en dos características: x_1 ingresos anuales, y x_2 edad en años
- ▶ Dos clasificadores \hat{f} 's: convexo-cuadrático (línea negra) y no-convexo (línea roja)

Clasificador de bayes I

Teorema de Bayes

$$p(c|\mathbf{x}) = \frac{p(\mathbf{x}|c)p(c)}{p(\mathbf{x})} \quad \text{para } c \in \{1, \dots, C\}$$

- ▶ $p(c|\mathbf{x})$ es la probabilidad de que objeto \mathbf{x} pertenezca a clase c
- ▶ $p(\mathbf{x}|c)$ es la probabilidad de que la clase c genere al objeto \mathbf{x}
- ▶ $p(c)$ es la probabilidad de observar un objeto cualquiera de la clase c
- ▶ $p(\mathbf{x})$ es la probabilidad de observar un objeto con las características de \mathbf{x}



Thomas Bayes
(1701-1761)

Clasificador de bayes II

Para clasificar un objeto utilizando el teorema de Bayes:

1. Encuentre la probabilidad de que el objeto x pertenezca a cada clase (es decir $p(c|x)$ para $c \in \{1, \dots, C\}$)
2. Asigne el objeto a la clase más probable

$$\begin{aligned}c^* &= \arg \max_c p(c|x) \\ &= \arg \max_c p(x|c)p(c)\end{aligned}$$

Esto se conoce como el criterio de *maximo a posteriori* (MAP)

Clasificador de bayes para variables categóricas I

Declaraciones tributarias deben ser clasificadas como fraudulentas o no.

- ▶ La clase $c \in \{si, no\}$
- ▶ Las características del objeto x son $\{Ind, Dep\}$
- ▶ Los datos son:

Employment	Fraud
Ind	Si
Dep	No
Ind	No
Ind	No
Dep	Si
Ind	Si
Dep	No
Ind	No

- ▶ Queremos clasificar:

Employment	Fraud
Ind	?

- ▶ $p(c = Si) = 3/8$
- ▶ $p(c = No) = 5/8$
- ▶ $p(x = Ind) = 5/8$
- ▶ $p(x = Dep) = 3/8$
- ▶ $p(x = Ind|c = Si) = 2/3$
- ▶ $p(c = Si|x = Ind) = ?$

$$\begin{aligned} p(c = Si|x = Ind) &= \frac{p(x = Ind|c = Si)p(c = Si)}{p(x = Ind)} \\ &= \frac{2/3 \cdot 3/8}{5/8} = 2/5 \end{aligned}$$



Clasificador de bayes para variables categóricas II

Consideremos ahora que $\mathbf{x} = [x_1, \dots, x_p]$, tenemos que

$$c^* = \arg \max_{c_j \in C} P(x_1, \dots, x_p | c_j) P(c_j).$$

Bajo el supuesto de independencia entre características x_i 's

$$c^* = \arg \max_{c_j \in C} P(c_j) \prod_{i=1}^p P(x_i | c_j).$$

- ▶ No considera interacciones entre variables
- ▶ No sufre de la maldición de la dimensionalidad
- ▶ Si la clase correcta tiene probabilidad alta es robusto al supuesto de independencia

Clasificador de bayes para variables categóricas III

Tenemos ahora

$$\mathbf{x} = [\textit{Employment}, \textit{Education}, \textit{Marital}, \textit{Occupation}, \textit{Gender}, \textit{Fraud}].$$

Employment	Education	Marital	Occupation	Gender	Fraud
Dep	Bachelor	Married	Professional	Male	Si
Dep	Bachelor	Married	Sales	Male	Si
Ind	College	Divorced	Sales	Female	No
Dep	College	Married	Sales	Male	No
Dep	HSgrad	Married	Sales	Male	No
Dep	College	Married	Professional	Male	No
Ind	Bachelor	Married	Professional	Female	Si
Dep	HSgrad	Married	Professional	Male	Si
Ind	College	Divorced	Sales	Female	No
Ind	Bachelor	Married	Professional	Male	No

Para calcular $p(\mathbf{x}|c)$ se asumirá independencia de las características, tal que

$$p([\textit{Employment}, \textit{Education}, \textit{Marital}, \textit{Occupation}, \textit{Gender}, \textit{Fraud}]|c)$$

es calculado utilizando

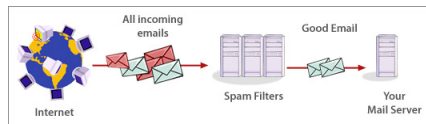
$$p(\textit{Employment}|c)p(\textit{Education}|c)p(\textit{Marital}|c)p(\textit{Occupation}|c)p(\textit{Gender}|c)p(\textit{Fraud}|c)$$

Clasificador de bayes para variables categóricas IV

Dear SIR,

I am Mr. John Coleman and my sister is Miss Rose Coleman, we are the children of late Chief Paul Coleman from Sierra Leone. I am writing you in absolute confidence primarily to seek your assistance to transfer our cash of twenty one Million Dollars (\$21,000.000.00) now in the custody of a private Security trust firm in Europe the money is in trunk boxes deposited and declared as family valuables by my late father as a matter of fact the company does not know the content as money, although my father made them to understand that the boxes belongs to his foreign partner.

...



Clasificador de bayes para variables categóricas V

Para utilizar un clasificador bayesiano debemos extraer características del email.

NIGERIAN_SUBJECT2	Subject is indicative of a Nigerian spam
FROM_ENDS_IN_NUMS	From: ends in numbers
URGENT_BIZ	BODY: Contains urgent matter
US_DOLLARS_3	BODY: Nigerian scam key phrase
DEAR_SOMETHING	BODY: Contains "Dear (something)"
...	...

Clasificador bayesiano dice probabilidad de que sea spam entre 30 y 40 % [score: 0.3728]