Fundamentos para al Análisis de Datos y la Investigación

Tema 4 - Aplicaciones: clasificación mediante Naïve Bayes

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad

Contents

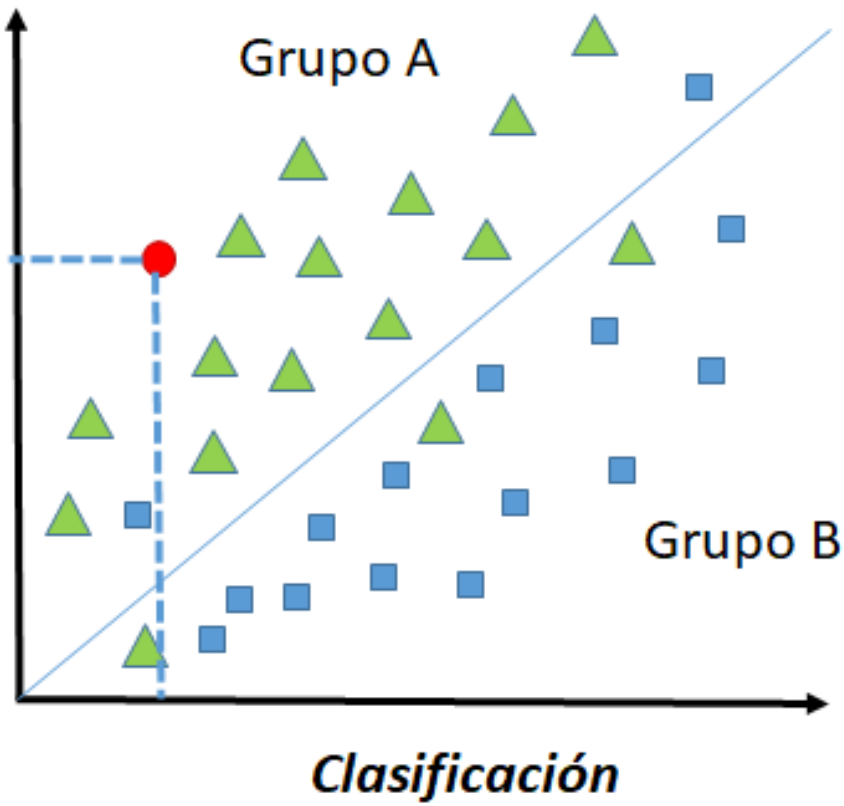
1 Introducción a la Clasificación

- **El problema de la clasificación**. Disponemos a priori de dos grupos diferentes de individuos definidos a partir de una serie de características comunes. Llega un nuevo individuo, y el problema consiste en asignarle o clasificarle en uno de los dos grupos.
- Si disponemos de dos grupos, se trata de un problema de *clasificación binaria*.
- Si en vez de partir de dos grupos disponemos de k grupos, se trata de un problema de *clasificación multiclase o multigrupo*.
- Otra forma de ver el problema desde el punto de vista analítico, es disponer de una variable dependiente categórica con valores $\{0, 1, \ldots, k\}$ y y tenemos que asignar un dato a partir de una serie de variables dependientes $x_1, x_2, \ldots, x_m$

1 Introducción a la Clasificación

Input

$(x_0,y_0)=(1,4)$



Output

(x_0,y_0) Grupo A

1 Introducción a la Clasificación

Ejemplos

- Credit Scoring: se trata de conceder o no conceder un crédito a un cliente. Previamente se clasifican a los clientes en morosos/no morosos y se dispone de información financiera sobre dichos clientes
- Un cliente comprará o no comprará un nuevo producto
- ¿Qué tipos o categorías de productos son de mayor interés para un cliente?
- ¿Cómo clasificar una película (romántica, comedia, thriller, drama,...) con objeto de predecir?

1 Introducción a la Clasificación

Métodos y modelos de clasificación:

- Tenemos los siguientes modelos:
  1. Regresión logística (junto con otros modelos de elección binaria o multiclase)
  2. Análisis discriminante
  3. Arboles de decisión - Random Forest
  4. Método de Bayes ingenuo (Naïve Bayes)
  5. Algoritmo kNN (k vecinos más próximos)
  6. Máquinas de soporte vectorial SVM
  7. Redes Neuronales
  8. Otros métodos
- Los modelos 1 a 4 proceden del ámbito de la Estadística mientras que 5 a 7 del ámbito de la inteligencia artificial.

Contents

2 Clasificador Naive Bayes

- El método de clasificación **Naive Bayes** en uno de los métodos más utilizados por su facilidad de implementación y su rapized
- Se trata de una técnica de clasificación y predicción supervisada que construye modelos de predicción, de modo que se predice la probabilidad de los posible grupos o resultados
- El método hace uso del Teorema de Bayes
- Se trata de predecir una variable categórica (con dos o más categorías) en términos de una serie de predictores. Los predictores pueden ser de naturaleza cuantitativa o cualitativa.

2 Descripción del Método

- Para la descripción del método suponemos una variable categórica G con un total de $\{1, 2, \ldots, K\}$ categoría. Suponemos dispoible un total de $X = X_1, X_2, \ldots, X_p$ predictores
- Escribimos de forma abreviada (cuando no exista confusión) G en vez de $G = k$ (grupo k-ésimo) y X en vez de $X_1, \ldots, X_p$
- El Teorema de Bayes establece

$$P(G|X) = \frac{P(X|G)P(G)}{P(X)} = \frac{P(G \cap X)}{P(X)}$$

- Puesto que el denominador es común e independiente de $G = k$ podemos prescindir de él para maximizar las probabilidades
- Tenemos entonces:

$$\begin{aligned} P(G \cap X) &= P(G \cap X_1 \cap \cdots \cap X_p) \\ &= P(G)P(X_1|G)P(X_2|GX_1) \cdots P(X_p|GX_1X_2 \cdots X_{p-1}) \end{aligned}$$

2 Descripción del Método

- Ahora, si suponemos independencia condicional obtenemos:

$$\begin{aligned} P(G \cap X) &= P(G \cap X_1 \cap \cdots \cap X_p) \\ &= P(G)P(X_1|G)P(X_2|G)\cdots P(X_p|G) \end{aligned}$$

- Por tanto, la probabilidad de pertenencia a un grupo dado los valores de X es:

$$P(G = k|X) = P(G = k)\prod_{i=1}^{p} P(X_i|G = k)$$

- Finalmente, el método Naive Bayes clasifica en la clase $G = k$:

$$\text{argmax}_k \left\{ P(G = k)\prod_{i=1}^{p} P(X_i|G = k) \right\}$$

- La predicción de la clase viene dada por:

$$P(G = k|X) = \frac{P(G = k)\prod_{i=1}^{p} P(X_i|G = k)}{P(X)}, \quad k = 1, 2, \ldots, K$$

Contents

3 Estimación de Probabilidades

- Las probabilidades $P(G = k)$, $k = 1, 2, \ldots, K$ se estiman a partir del tamaño de los grupos a priori
- Un aspecto clave en el método es la estimación de las probabilidades $P(X_i|G = k)$, $i = 1, 2, \ldots, p$
- Si la variable X_i es categórica es estima mediante la regla de Laplace corregida, para evitar los casos que en numerador sea 0.
- En el caso que la variable X_i sea continua, una posibilidad es suponer una distribucion normal,

$$X_i|G = k \sim N(\mu_{ik}, \sigma^2_{ik})$$

donde μ_{iG} es la media de la variable X_i en el grupo k y σ^2_{iG} la varianza de la variable X_i en el grupo k

Contents

4 Estimación Naive Bayes

- La estimación Naive Bayes se realiza por medio del comando `naiveBayes` del paquete `e1071`
- Sintaxis:

```
naiveBayes(formula, data, laplace = 0, ..., subset,
                    na.action = na.pass)
```

- Predicción:

```
predict(object, newdata,type = c(``class'', ``raw''))
```

4 Estimación Naive Bayes

```
Naive Bayes Classifier for Discrete Predictors

Call:
naiveBayes.default(x = X, y = Y, laplace = laplace)

A-priori probabilities:
Y
    setosa versicolor  virginica
 0.3333333  0.3333333  0.3333333

Conditional probabilities:
            Sepal.Length
Y             [,1]      [,2]
  setosa     5.006 0.3524897
  versicolor 5.936 0.5161711
  virginica  6.588 0.6358796

            Sepal.Width
Y             [,1]      [,2]
  setosa     3.428 0.3790644
  versicolor 2.770 0.3137983
  virginica  2.974 0.3224966

            Petal.Length
Y             [,1]      [,2]
  setosa     1.462 0.1736640
  versicolor 4.260 0.4699110
  virginica  5.552 0.5518947

            Petal.Width
Y             [,1]      [,2]
  setosa     0.246 0.1053856
  versicolor 1.326 0.1977527
  virginica  2.026 0.2746501
```

Figura: Resultado Naive Bayes

4 Estimación Naive Bayes

```
> #-----------------------------------------------------
> # Matriz de Confusion y Probabilidad de Acierto
> #-----------------------------------------------------
> matrizconfusion <- table(iris$Species, prediccion)
> matrizconfusion
            prediccion
             setosa versicolor virginica
  setosa         50          0         0
  versicolor      0         47         3
  virginica       0          3        47
> # Porcentaje de aciertos
> sum(diag(matrizconfusion))/sum(matrizconfusion)
[1] 0.96
```

Figura: Matriz de Confusión Naïve Bayes

Fundamentos para al Análisis de Datos y la Investigación

Tema 4 - Aplicaciones: clasificación mediante Naïve Bayes

José María Sarabia

Máster Universitario en Ciencia de Datos
CUNEF Universidad