CLASIFICADOR BAYES INGENUO

Dr. Jorge Hermosillo
Laboratorio de Semántica Computacional

Centro de Investigación en Ciencias

Universidad Autónoma del Estado de Morelos

REGLAS DE DECISIÓN PARA MODELOS PROBABILISTAS

REFERENCIAS BIBLIOGRÁFICAS

📖 Referencias bibliográficas:

- Flach, Peter (2012). *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge University Press.
- Bishop, Christopher M. (2006). *Pattern recognition and machine learning*. New York. Springer.
- Mitchell, Tom (1997). *Machine Learning*. McGraw-Hill, New York.

REGLAS DE DECISIÓN PARA MODELOS PROBABILISTAS

- ¿Cómo decidir si una instancia X pertenece a una clase $Y = 0$ o $Y = 1$?
- Una primera forma es con $P(Y|X) > 0.5$
- La regla de Bayes:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{P(XY)}{P(X)}$$

nos proporciona otra opción:

- La función de ***verosimilitud***

$$P(X|Y)$$

nos dice qué tanto una clase Y explica (hace creíbles) los datos X.

REGLAS DE DECISIÓN PARA MODELOS PROBABILISTAS

- Usando la verosimilitud, podemos decidir si la instancia pertenece más a una clase que a otra calculando su razón:

$$LR = \frac{P(X|Y=0)}{P(X|Y=1)}$$

- La ***razón de verosimilitud*** (LR – $Likelihood\ Ratio$) nos permite establecer otra de decisión:

$$LR > 1 \ \rightarrow Y = 0$$
$$LR < 1 \ \rightarrow Y = 1$$

REGLAS DE DECISIÓN PARA MODELOS PROBABILISTAS

- **MAP** (***Maximum A Posteriori***): Regla Bayesiana

$$y_{MAP} = \operatorname*{argmax}_{Y} P(Y|X) = \operatorname*{argmax}_{Y} \frac{P(X|Y)P(Y)}{P(X)} = \operatorname*{argmax}_{Y} P(X|Y)\, P(Y)$$

si $P(X)$ se ignora o se estima en un paso posterior.

- **ML** (***Maximum Likelihood***): Regla frecuentista

$$y_{ML} = \operatorname*{argmax}_{Y} P(X|Y)$$

- Usar:
  - *ML si no importa el a priori sobre Y, $P(Y)$, o se considera uniforme*
  - *MAP en caso contrario*

CRITERIO DE OPTIMALIDAD DE BAYES

- Hasta ahora nos hemos hecho la pregunta de "¿Cuál es la hipótesis más probable, dados los datos de entrenamiento?"
- Una pregunta que es seguido de mayor importancia es "¿Cuál es la clasificación más probable de una nueva instancia, dados los datos de entrenamiento?"
- A esta pregunta, responde el criterio de clasificación óptima de Bayes.

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Donde v_j representa cualquier valor de clasificación posible para una nueva instancia, y h_i es una hipótesis que explica la verosimilitud de v_j.

CRITERIO DE OPTIMALIDAD DE BAYES

- Cualquier sistema que clasifique nuevas instancias de acuerdo con la ecuación anterior se le llama clasificador "***Bayes-óptimo***" (*Bayes-optimal*).
- Decimos que un modelo de clasificación es Bayes-óptimo, si siempre asigna $\mathrm{argmax}_{\mathrm{Y}} P^{*}(Y = y|X = x)$ a una instancia x, donde P^{*} representa la verdadera distribución a posteriori.
- Calcular la ecuación anterior puede ser muy ineficiente, por lo que este concepto guarda un interés teórico debido a que ningún método de clasificación puede superar a este.

CLASIFICADOR BAYES INGENUO

CLASIFICADOR BAYES INGENUO

- Un método muy práctico de aprendizaje Bayesiano es el clasificador Ingenuo de Bayes (*Naïve Bayes*).
- Este clasificador se aplica en tareas de aprendizaje donde cada instancia $\mathbf{x}$ está descrita por una conjunción de valores de atributos y donde la función objetivo $f(\mathbf{x})$ puede tomar cualquier valor y_j en un conjunto finito $\boldsymbol{C}$.
- Desde un enfoque Bayesiano de la clasificación de una nueva instancia $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, el clasificador debe asignar un y_{MAP} tal que:

$$y_{MAP} = \underset{y_j \in \boldsymbol{C}}{\operatorname{argmax}} P(y_j | x_1, x_2, \dots, x_N)$$

CLASIFICADOR BAYES INGENUO

- Se puede usar el teorema de Bayes para reescribir esta expresión como:

$$
\begin{aligned}
y_{MAP} &= \underset{y_j \in \mathcal{C}}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_N | y_j) P(y_j)}{P(x_1, x_2, \ldots, x_N)} \\
&= \underset{y_j \in \mathcal{C}}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_N | y_j) P(y_j)
\end{aligned}
$$

- El clasificador Bayes Ingenuo (BI) se basa en el supuesto de que los valores de atributos son condicionalmente independientes dada la clase (y_j).

$$
y_{BI} = \underset{y_j \in \mathcal{C}}{\operatorname{argmax}} P(y_j) \prod_i P(x_i | y_j) \qquad (1)
$$

- Nota que el número de términos $P(x_i | y_j)$ que deben estimarse de los datos de entrenamiento es sólo el número de valores de atributos por el número de clases.

CLASIFICADOR BAYES INGENUO

- El método de aprendizaje BI involucra un paso de aprendizaje en el cual los distintos términos $P(y_j)$ y $P(x_i \mid y_j)$ se estiman, con base en sus frecuencias en los datos de entrenamiento.
- El conjunto de estas estimaciones constituye la hipótesis aprendida.
- Esta hipótesis es la que se utiliza luego para clasificar cada nueva instancia, aplicando la ecuación (1).
- Siempre que la independencia condicional del supuesto de Bayes Ingenuo se cumpla, la clasificación y_{BI} es idéntica a la clasificación MAP.
- Una diferencia interesante entre el método de aprendizaje BI y otros métodos que hemos visto es que no hay una búsqueda explícita a través del espacio de hipótesis posibles.
- En este caso, el espacio de hipótesis posibles, es el espacio de valores posibles que pueden asignarse a los términos $P(y_j)$ y $P(x_i \mid y_j)$, mediante simple conteo de las combinaciones de datos en los ejemplos de entrenamiento.

APLICACIÓN A DETECCIÓN DE SPAM

Aplicación de un Clasificador Bayes Ingenuo

¿SPAM O HAM?

- En este caso, la VA’s son discretas, por lo que sus distribuciones de probabilidad son Tablas.
- Un ejemplo de cómo se vería la distribución $P(y|\mathbf{x})$ podría ser el siguiente:

$x_1 = Viagra$	$x_2 = Lotería$	$P(y = spam\|x_1, x_2)$	$P(y = ham\|x_1, x_2)$
0	0	0.31	**0.69**
0	1	**0.65**	0.35
1	0	**0.80**	0.20
1	1	0.40	**0.60**

“Viagra” y “Lotería” son atributos booleanos. En cada fila, la clase más probable se resalta en negrillas.

MODELACIÓN DEL PROBLEMA

- Problema:

"Dado un correo electrónico conteniendo ciertas palabras, determinar si el correo es spam o ham"

- Las instancias $\mathbf{x}$ son correos electrónicos, cuyas características ($atributos\ x_1, x_2 \cdots, x_N$) son palabras, que sugieren "spam", contenidas en cada instancia, y las clases $\mathrm{y_j}$ son *spam* y *ham*.
- El problema consiste ahora en modelar $P(y_j)$ y $P(x_i\,|y_j)$ para calcular y_{BI}:

$$y_{BI} = \underset{y_j \in \{\boldsymbol{spam},\boldsymbol{ham}\}}{\operatorname{argmax}} \underbrace{P(y_j)}\prod_i P(x_i\,|y_j)$$

- Alternativamente, podemos modelar solamente la *likelyhood* $\prod_i P(x_i\,|y_j)$ y calcular:

$$LR = \frac{\prod_i P(x_i\,|y_j = spam)}{\prod_i P(x_i\,|y_j = ham)}$$

- Identifiquemos a Spam con $\oplus$ y Ham con $\ominus$

ENFOQUE 1: OCURRENCIA DE PALABRAS (DISTRIBUCIÓN DE BERNOULLI)

- La siguiente Tabla muestra la ocurrencia de tres palabras a, b y c en un conjunto de *emails* que han sido clasificados como *spam* (+) y *ham* (−).

email	$a?$	$b?$	$c?$	***Clase***
e1	0	1	0	+
e2	0	1	1	+
e3	1	0	0	+
e4	1	1	0	+
e5	1	1	0	−
e6	1	0	1	−
e7	1	0	0	−
e8	0	0	0	−

MODELO CON DIST. DE BERNOULLI

- ¿A qué corresponde $\mathbf{x}$ de nuestro modelo en esta Tabla?
  - *La conjunción de ocurrencias de a, b y c* $\rightarrow \mathbf{x} = (a,b,c)$.

↑ x no se observa

- ¿Qué dimensión tiene $\mathrm{P}(\mathbf{x}|y = \oplus)$?
  - $\mathbf{P}(\mathbf{x}|\mathrm{y} = \oplus) = \theta^{\oplus} = (\theta_1,\theta_2,\theta_3)$.
- ¿Cómo se calcula $\mathrm{P}(\mathbf{x}|y = \oplus)$?
  - *Aplicando la corrección de Laplace* $P(x_i|y = \oplus) = \frac{n_i+1}{|S|+k}$ *con* $k = 2$ *(regla de sucesión de Laplace);* $\mathrm{S} \equiv \{\mathbf{x} \in \oplus\} \rightarrow |S| = 4$

$$\theta^{\oplus}_{Blli} = \left(\frac{1}{2},\frac{2}{3},\frac{1}{3}\right) \qquad \frac{2+1}{4+2} = \frac{3}{6} = \frac{1}{2}$$

- Calcula $\mathrm{P}(\mathbf{x}|y = \ominus)$

$$\theta^{\ominus}_{Blli} = \left(\frac{2}{3},\frac{1}{3},\frac{1}{3}\right)$$

e	$a?$	$b?$	$c?$	y
e1	0	1	0	+
e2	0	1	1	+
e3	1	0	0	+
e4	1	1	0	+
e5	1	1	0	–
e6	1	0	1	–
e7	1	0	0	–
e8	0	0	0	–

ENFOQUE 2: CONTEO DE PALABRAS (DISTRIBUCIÓN MULTINOMIAL)

- **Distribución Categórica**: Generaliza la distribución de Bernoulli para $k > 2$. El parámetro de la distribución es un vector de tamaño k: $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k)$ tal que $\sum_{i=1}^{k} \theta_i = 1$
- **Distribución Multinomial**: tabula las probabilidades de n ensayos categóricos independientes e idénticamente distribuidos (i.i.d.). Esto es, si $\mathbf{X} = (X_1, \cdots, X_k)$ es un k-vector de conteos enteros, entonces:

$$P(\mathbf{X} = (x_1, \cdots, x_k)) = n! \frac{\theta_1^{x_1}}{x_1!} \cdots \frac{\theta_k^{x_k}}{x_k!}, \quad \text{con } \sum_{i=1}^{k} x_i = n$$

- Si $n = 1$ tenemos una dist. Categórica con exactamente una $x_i = 1$ y el resto igual a 0. Además, para $k = 2$, la ecuación de arriba nos da la expresión de una dist. de Bernoulli: $P(X = x) = \theta^x (1 - \theta)^{1-x}$ para $x \in \{0,1\}$.

ENFOQUE 2: CONTEO DE PALABRAS (DISTRIBUCIÓN MULTINOMIAL)

- Siguiendo el ejemplo de tres palabras a, b y c en un conjunto de *emails* que han sido clasificados como *spam* $(+)$ y *ham* $(-)$. La idea es ahora contar palabras.

Modelo
Bag of Words
BOW

email	$\#a$	$\#b$	$\#c$	*Clase*
e1	0	3	0	$+$
e2	0	3	3	$+$
e3	3	0	0	$+$
e4	2	3	0	$+$
e5	4	3	0	$-$
e6	4	0	3	$-$
e7	3	0	0	$-$
e8	0	0	0	$-$

MODELO CON DISTRIBUCIÓN MULTINOMIAL

- ¿Cuál es, en este caso, la VA $\mathbf{x}$?
  - *La conjunción de las conteos de a, b y c*

- ¿Cómo se obtiene la distribución de probabilidad $\mathrm{P}(\mathbf{x}|y)$?
  - *Aplicando* $P(x_i|y=\oplus)=\frac{n_i+1}{S_\oplus+k}$ *con* $k=3$ *(un pseudo conteo adicional por cada palabra);* $S_\oplus=\sum_\oplus x_i=17$

$$\theta_{Mi}^{\oplus}=\left(\frac{6}{20},\frac{10}{20},\frac{4}{20}\right)$$

- Calcula $\mathrm{P}(\mathbf{x}|y=\ominus)$

e	$a?$	$b?$	$c?$	Cl
e1	0	3	0	+
e2	0	3	3	+
e3	3	0	0	+
e4	2	3	0	+
e5	4	3	0	−
e6	4	0	3	−
e7	3	0	0	−
e8	0	0	0	−

EJERCICIO: COMPARA LOS CLASIFICADORES

- Clasifica un nuevo correo que contiene a, b pero no c usando un modelo multivariable de Bernoulli:

$$\hat{\theta}^{\oplus}_{Blli} = (0.5, 0.\underline{67}, 0.33) \qquad \hat{\theta}^{\ominus}_{Blli} = (0.67, 0.33, 0.33)$$

y? $e = (1, 1, 0) = \mathbf{x}$

$$LR = \frac{P(\mathbf{x} \mid \oplus)}{P(\mathbf{x} \mid \ominus)}$$

$$P(\mathbf{x} \mid \oplus) = P(x_1, x_2, x_3 \mid \oplus)$$

$$= P(x_1 \mid \oplus) P(x_2 \mid \oplus) P(x_3 \mid \oplus) = \prod_{i=1}^{3} P(x_i \mid \oplus)$$

$$= P(x_1 = 1 \mid \oplus) P(x_2 = 1 \mid \oplus) P(x_3 = 0 \mid \oplus)$$

$$= 0.5 \times 0.67 \times (1 - 0.33)$$

$$= 0.222$$

$$P(x \mid \ominus) = 0.67 \times 0.33 \times (1 - 0.33)$$

$$= 0.148$$

$$LR = \frac{0.222}{0.148} > 1$$

$\Rightarrow y = \text{Spam}$

EJERCICIO: COMPARA LOS CLASIFICADORES

- Clasifica un nuevo correo que contiene $3a$, $1b$ y $0c$ usando un modelo multinomial:

$$\hat{\theta}_{Mi}^{\oplus} = (0.3, 0.5, 0.2) \qquad \hat{\theta}_{Mi}^{\oplus} = (0.6, 0.2, 0.2)$$

$y = ?$ $\quad x = (3, 1, 0) \quad n = \sum x_i = 4$

$$P(x \mid \oplus) = n! \cdot \frac{\theta_1^{x_1}}{x_1!} \cdot \frac{\theta_2^{x_2}}{x_2!} \cdot \frac{\theta_3^{x_3}}{x_3!} = 4! \, \frac{0.3^3}{3!} \cdot \frac{0.5^1}{1!} \cdot \frac{0.2^0}{0!}$$

$$P(x \mid -) = \quad = 4! \, \frac{0.6^3}{3!} \cdot \frac{0.2^1}{1!} \cdot \frac{0.2^0}{0!}$$

$$LR = \left(\frac{0.3}{0.6}\right)^3 \left(\frac{0.5}{0.2}\right)^1 \left(\frac{0.2}{0.2}\right)^0 = \frac{5}{16} < 1$$

$y =$ Ham!

CONCLUSIONES

- En un modelo de bernoulli multivariado todas las palabras contribuyen (presentes y ausentes), multiplicando por $\theta_i^{\oplus}/\theta_i^{\ominus}$ si $x_i = 1$ y por $(1-\theta_i^{\oplus})/(1-\theta_i^{\ominus})$ si $x_i = 0$, mientras que en el modelo multinomial sólo las palabras presentes contribuyen con factores $\left(\frac{\theta_i^{\oplus}}{\theta_i^{\ominus}}\right)^{x_i}$.
- El clasificador bayes ingenuo descompone un problema de aprendizaje multivariado en sub-tareas univariadas.
- Este modelo es usualmente utilizado en clasificación de texto, categórica, y mezcla de categórica/datos-numéricos reales.