ESTIMACIÓN BAYESIANA

RECONOCIMIENTO DE PATRONES

AUTOR: DR. JIMENA OLVERES MONTIEL

LaPI
LABORATORIO AVANZADO DE PROCESAMIENTO DE IMÁGENES

1

Var. Aleatorias Continuas

Fun. de Densidad de Probabilidad

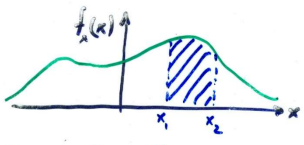
$f_x(x)$

Propiedades

$f_x(x) \geq 0$

$\int_{-\infty}^{\infty} f_x(x)\,dx = 1$

$P(x \leq a) = F_x(a) = \int_{-\infty}^{a} f_x(x)\,dx$



$P(x_1 < x < x_2) = \int_{x_1}^{x_2} f_x(x)\,dx$

$f_x(x) = \frac{dF_x(x)}{dx}$

LaPI

2

Fun. de Den. de Prob. Conjunta

$f_{X,Y}(x,y)$

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx$$

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \qquad f_Y(y) > 0$$

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \qquad f_X(x) > 0$$

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)\, f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|\kappa) f_Y(\kappa)\,d\kappa} \quad \text{(Bayes)} \qquad f_X(x) \neq 0$$

Promedios Estadísticos

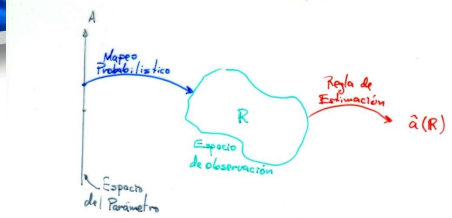
$$E\{g(x)\} = \int_{-\infty}^{\infty} g(x) f(x)\,dx$$

$$E\{g(x,Y)\} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y) f_{XY}(x,y)\,dx\,dy$$

$$\rho_{XY} = \frac{E\{(X-\mu_X)(Y-\mu_Y)\}}{\sigma_X \sigma_Y}$$

3

Estimación de Parámetros



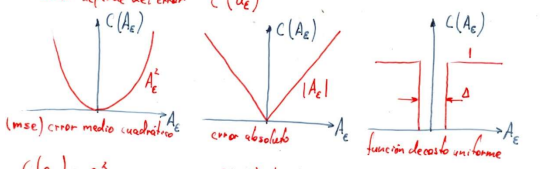
a y $\hat{a}(R)$ son variables continuas

I. Caso: a es una variable aleatoria

A cada par $[a, \hat{a}(R)]$ se asigna un costo $C(a,\hat{a})$

Error de la estimación $a_\varepsilon(R) \triangleq \hat{a}(R) - a$

Costo depende del error $C(a_\varepsilon)$



$C(a_\varepsilon) = a_\varepsilon^2$ $\qquad C(a_\varepsilon) = |a_\varepsilon|$ $\qquad C(a_\varepsilon) = \begin{cases} 0 & |a_\varepsilon| \leq \frac{\Delta}{2} \\ 1 & |a_\varepsilon| > \frac{\Delta}{2} \end{cases}$

4

Función de costo debe

a) Satisfacer los requerimientos del usuario

b) Permitir resolver el problema

$P_a(A)$ Densidad de probabilidad a priori (conocida)

Si $P_a(A)$ es desconocida, entonces se sigue un procedimiento MINIMAX

Riesgo $R \triangleq E\{C[a, \hat{a}(R)]\} = \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} C[A - \hat{a}(R)]\, P_{a,r}(A,R)\, dR$

Estimación de Bayes minimiza el riesgo:

a) Criterio del error medio cuadrático:

$$R_{ms} = \int_{-\infty}^{\infty} dA \int_{-\infty}^{\infty} dR\, [A - \hat{a}(R)]^2\, P_{a,r}(A,R)$$

donde $P_{a,r}(A,R) = P_r(R)\, P_{a|r}(A|R)$

$$\therefore R_{ms} = \int_{-\infty}^{\infty} dR\, P_r(R) \int_{-\infty}^{\infty} dA\, [A - \hat{a}(R)]^2\, P_{a|r}(A|R)$$

Para minimizar R_{ms} hay que minimizar esta integral, o sea, la varianza a posteriori ó varianza condicional

Derivando con respecto a $\hat{a}(R)$:

$$\frac{d}{d\hat{a}} \int_{-\infty}^{\infty} dA\, [A - \hat{a}(R)]^2\, P_{a|r}(A|R) = -2 \int_{-\infty}^{\infty} A\, P_{a|r}(A|R)\, dA + 2\hat{a}(R) \underbrace{\int_{-\infty}^{\infty} P_{a|r}(A|R)\, dA}_{=1}$$

Igualando a cero, obtenemos $\hat{a}_{ms}(R)$:

$$\hat{a}_{ms}(R) = \int_{-\infty}^{\infty} dA\, A\, P_{a|r}(A|R)$$

media de la densidad a posteriori, media condicional.

5

b) Criterio del error absoluto

$$R_{abs} = \int_{-\infty}^{\infty} dR\, P_r(R) \int_{-\infty}^{\infty} dA\, [|A - \hat{a}(R)|]\, P_{a|r}(A|R)$$

Para Minimizar la integral interior:

$$I(R) = \int_{-\infty}^{\hat{a}(R)} dA\, [\hat{a}(R) - A]\, P_{a|r}(A|R) + \int_{\hat{a}(R)}^{\infty} dA\, [A - \hat{a}(R)]\, P_{a|r}(A|R)$$

Derivando e igualando a cero:

$$\int_{-\infty}^{\hat{a}_{abs}(R)} dA\, P_{a|r}(A|R) = \int_{\hat{a}_{abs}(R)}^{\infty} dA\, P_{a|r}(A|R)$$

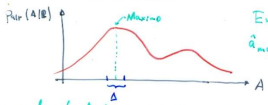
Mediana de la densidad a posteriori

c) Función de costo uniforme

$$R_{uni} = \int_{-\infty}^{\infty} dR\, P_r(R) \left[1 - \int_{\hat{a}_{map}(R) - \Delta/2}^{\hat{a}_{map}(R) + \Delta/2} P_{a|r}(A|R)\, dA\right]$$

Para minimizar R_{uni} hay que maximizar esta integral

Si $\Delta \to 0$, entonces:



En este caso la estimación es $\hat{a}_{map}(R)$ (est. a posteriori máxima)

$\hat{a}_{map}$ es el punto donde $P_{a|r}(A|R)$ es máxima

Si el máximo es interior a A y $\ln P_{a|r}(A|R)$ es continua, entonces una condición necesaria para encontrar $\hat{a}_{map}(R)$ se puede obtener así:

$$\left.\frac{\partial \ln P_{a|r}(A|R)}{\partial A}\right|_{A = \hat{a}(R)} = 0$$ Ecuación MAP

6

En el caso de la ecuación MAP, hay que estar seguros que la solución es un máximo absoluto

Separando las contribuciones del vector $\mathbb{R}$ y de la información a priori de $P_{A|R}(A|R)$

$$P_{A|R}(A|R) = \frac{P_{R|A}(R|A)\,P_A(A)}{P_R(R)}$$

$$\ln P_{A|R}(A|R) = \ln P_{R|A}(R|A) + \ln P_A(A) - \underbrace{\ln P_R(R)}_{\text{independiente de } A}$$

$$l(A) \triangleq \underbrace{\ln P_{R|A}(R|A)}_{\text{dependencia de } R \text{ de } A} + \underbrace{\ln P_A(A)}_{\text{conocimiento a priori}}$$

Ecuación MAP:

$$\left.\frac{\partial l(A)}{\partial A}\right|_{A=\hat{a}(R)} = \left.\frac{\partial \ln P_{R|A}(R|A)}{\partial A}\right|_{A=\hat{a}(R)} + \left.\frac{\partial \ln P_A(A)}{\partial A}\right|_{A=\hat{a}(R)} = 0$$

Ejemplo $r_i = a + n_i \qquad i = 1, 2, \ldots, N$

a es una var. aleat. Gaussiana $N(0, \sigma_a)$ y n_i son variables aleat. Gaussianas independientes $N(0, \sigma_n)$

$$P_{R|A}(R|A) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n} e^{-\frac{(R_i - A)^2}{2\sigma_n^2}}$$

$$P_A(A) = \frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{A^2}{2\sigma_a^2}}$$

7

a) Encontrar $\hat{a}_{mas}(R)$

Necesitamos $P_{A|R}(A|R) = \frac{P_{R|A}(R|A)\,P_A(A)}{P_R(R)} = \frac{P_{R,A}(R,A)}{P_R(R)}$ } cte que hace que $\int_{-\infty}^{\infty} P_{A|R}(A|R)\,dA = 1$

$$\therefore P_{A|R}(A|R) = \frac{\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_n}\right)\frac{1}{\sqrt{2\pi}\sigma_a}}{P_R(R)} e^{-\frac{1}{2}\left[\sum_{i=1}^{N}\frac{(R_i - A)^2}{\sigma_n^2} + \frac{A^2}{\sigma_a^2}\right]}$$

Reagrupando:

$$P_{A|R}(A|R) = k(R)\, e^{-\frac{1}{2\sigma_p^2}\left[A - \frac{\sigma_a^2}{\sigma_a^2 + \sigma_n^2/N}\left(\frac{1}{N}\sum_{i=1}^{N} R_i\right)\right]^2}$$ } F. de densidad gaussiana.

donde $\sigma_p^2 \triangleq \left(\frac{1}{\sigma_a^2} + \frac{N}{\sigma_n^2}\right)^{-1} = \frac{\sigma_a^2\sigma_n^2}{N\sigma_a^2 + \sigma_n^2}$

Varianza a posteriori

$\hat{a}_{mas}(R)$ es la media condicionada:

$$\hat{a}_{mas}(R) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_n^2/N}\left(\frac{1}{N}\sum_{i=1}^{N} R_i\right)$$

El riesgo m.s.e es la varianza a posteriori

Observaciones:

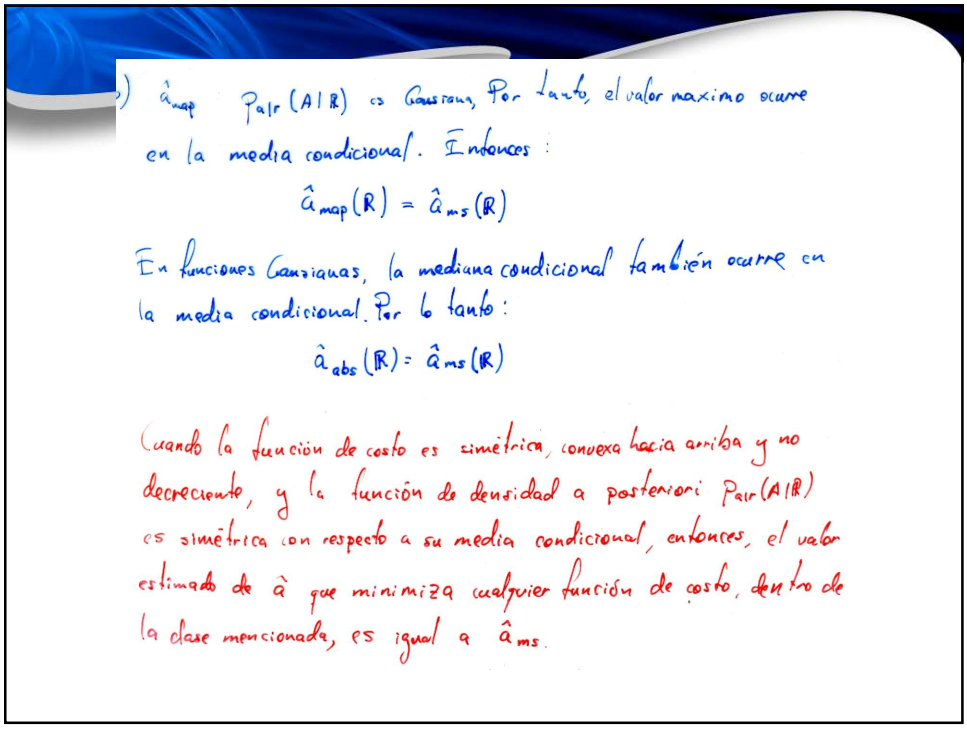
1) $l(R) = \sum_{i=1}^{N} R_i \Rightarrow$ Estadística suficiente

4) a) Si $\sigma_a^2 \ll \frac{\sigma_n^2}{N} \Rightarrow$ Conocimiento a-priori es mejor que los datos observados y el valor estimado es parecido a la media a-priori. (0)

b) Si $\sigma_a^2 \gg \frac{\sigma_n^2}{N} \Rightarrow$ Conocimiento a priori es de poco valor, y el valor esperado son los datos recibidos

c) En el límite: $\lim_{\frac{\sigma_n^2}{N\sigma_a^2} \to 0} \hat{a}_{mas}(R) = \frac{1}{N}\sum_{i=1}^{N} R_i$ } Promedio aritmético de R_i

8



9

Teorema de Bayes para Clasificación

- Sea "x" un vector de características y "Ci" una clase:

$$P(C_i|x) = \frac{P(x|C_i)P(C_i)}{P(x)}$$

10

- Si tenemos dos clases, C1 y C2, y queremos saber a cuál clase pertenece x, simplemente usamos la anterior fórmula y vemos cuál de las dos clases tiene mayor probabilidad.

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x)}$$

$$P(C_2|x) = \frac{P(x|C_2)P(C_2)}{P(x)}$$

Se observa que como el denominador es el mismo, entonces el denominador no importa y típicamente solo se calcula el numerador y se ve cuál es más grande.

11

- Entonces el Clasificador solo necesitaría calcular esto para clasificar a un vector "x": $P(C_i|x) \propto P(x|C_i)P(C_i)$

- Y, por tanto, en el entrenamiento, solamente se necesita obtener:

$$P(x|C_i) \qquad P(C_i)$$

12

Bayes Ingenuo
(Naive Bayes)

Imaginando que "x" es un vector de 3 componentes, tendríamos la siguiente probabilidad conjunta:

$$P(x|C_i) = P(x_1, x_2, x_3|C_i)$$

13

- Y recordando la expansión de la probabilidad conjunta en probabilidades condicionales, tendríamos:

$$P(x_1, x_2, x_3|C_i)$$

$$= P(x_1|C_i)\,P(x_2|x_1, C_i)\,P(x_3|x_2, x_1, C_i)$$

14

- Intentar calcular las probabilidades para todo x1, x2 y x3, suele resultar muy complicado, por lo que **se asume que x1, x2 y x3 son independientes**.
- Asumiendo que son independientes la anterior fórmula se simplifica:

$$P(x_1, x_2, x_3|C_i) = P(x_1|C_i)\,P(x_2|x_1, C_i)\,P(x_3|x_2, x_1, C_i)$$
$$= P(x_1|C_i)\,P(x_2|C_i)\,P(x_3|C_i)$$

15

Bayes Ingenuo

- A este clasificador se le llama **Bayes Ingenuo**, ya que resulta ingenuo pensar que los elementos del vector de características son independientes.
- Sin embargo se simplifica mucho la fórmula, haciendo que sea uno de los clasificadores más rápidos para entrenar.

16

- Generalizando a un vector x de tamaño n, simplemente tendríamos:

$$P(x|C_i)=\prod_{j=1}^{n} P(x_j|C_i)$$

- Y entonces en el entrenamiento tendríamos que calcular

$$P(x_j|C_i)$$

para toda j para todo i.

17

Naive Bayes

Resumiendo:

$$p(C_j \mid x_1, x_2, \ldots, x_d) \propto p(x_1, x_2, \ldots, x_d \mid C_j) p(C_j)$$

$$p(X \mid C_j) \propto \prod_{k=1}^{d} p(x_k \mid C_j)$$

$$p(C_j \mid X) \propto p(C_j) \prod_{k=1}^{d} p(x_k \mid C_j)$$

18

Ejemplo: Clasificador de SPAM

Training Set: Correos preprocesados

Tenemos 8 correos, 3 son spam y 5 no son spam. Tras quitar palabras duplicadas y palabras que no interesan, quedan así:

Spam	No Spam
oferta es secreto	practica deportes hoy
click link secreto	fue practica deportes
link deportes secreto	evento deportes secreto
	deportes es hoy
	deportes cuesta dinero

19

Entrenamiento (Contar)

Palabra	Spam	No Spam
Oferta	1	0
Es	1	1
Secreto	3	1
Click	1	0
Link	2	0
Deportes	1	5
Practica	0	2
Hoy	0	2
Fue	0	1
Evento	0	1
Cuesta	0	1
Dinero	0	1
TOTAL:	**9**	**15**

20

Entrenamiento (Calcular Probabilidades)

Palabra	P(Palabra | Spam)	P(Palabra|No Spam)
Oferta	1 /9	0
Es	1 /9	1 / 15
Secreto	3 /9	1 / 15
Click	1 /9	0
Link	2 /9	0
Deportes	1 /9	5 / 15
Practica	0	2 / 15
Hoy	0	2 / 15
Fue	0	1 / 15
Evento	0	1 / 15
Cuesta	0	1 / 15
Dinero	0	1 / 15
TOTAL:	**1**	**1**

21

Entrenamiento (Calcular Probabilidades)

- Así ya obtuvimos todas las P(xj | Ci)
- Luego para obtener P(Ci) recordamos que teníamos solo 8 correos, 3 eran de spam y 5 no eran de spam, entonces:

P(Spam) = 3/8

P(No Spam) = 5/8

22

Ejemplo de Predicción

- Tenemos un mensaje M, lo preprocesamos y terminamos con "secreto es deportes"
- Queremos clasificarlo, así que buscaremos calcular:

$$P(Spam|\text{secreto es deportes}) \propto P(\text{secreto es deportes}|Spam)\,P(Spam)$$

$$P(No\,Spam|\text{secreto es deportes}) \propto P(\text{secreto es deportes}|No\,Spam)\,P(No\,Spam)$$

23

- Como estamos usando Bayes Ingenuo, calculamos las probabilidades de cada palabra de manera independiente.

$$P(\text{secreto es deportes}\,|Spam)\,P(Spam)$$

$$= P(secreto|Spam)\,P(es|Spam)\,P(deportes|spam)\,P(Spam)$$

$$= (3/9)(1/9)(1/9)(3/8)$$

24

- Haciendo los cálculos para ambas clases, se obtiene:

$$P(\text{secreto es deportes} \mid Spam)P(Spam)=(3/9)(1/9)(1/9)(3/8)=0.00154$$

$$P(\text{secreto es deportes} \mid No\ Spam)P(No\ Spam)=(1/15)(1/15)(5/15)(5/8)=0.00092$$

Por lo tanto se clasifica como Spam

25

Problema: Datos no vistos

- Si tenemos "Oferta practica secreto" sucedería lo siguiente:

$$P(\text{oferta practica secreto} \mid Spam)P(Spam)=(1/9)(0)(3/9)(3/8)=0$$

$$P(\text{oferta practica secreto} \mid No\ Spam)P(No\ Spam)=(0)(2/15)(1/15)(5/8)=0$$

¡No podría clasificarlo!

26

Solución: "Smoothing"

Palabra	P(Palabra | Spam)	P(Palabra|No Spam)
Oferta	(1 + 1) / (9 + 12)	(0 + 1) / (15 + 12)
Es	(1 + 1) /(9 + 12)	(1 + 1) / (15 + 12)
Secreto	(3 + 1) /(9 + 12)	(1 + 1) / (15 + 12)
Click	(1 + 1) /(9 + 12)	(0 + 1) / (15 + 12)
Link	(2 + 1) /(9 + 12)	(0 + 1) / (15 + 12)
Deportes	(1 + 1) /(9 + 12)	(5 + 1) / (15 + 12)
Practica	(0 + 1) / (9 + 12)	(2 + 1) / (15 + 12)
Hoy	(0 + 1) / (9 + 12)	(2 + 1) / (15 + 12)
Fue	(0 + 1) / (9 + 12)	(1 + 1) / (15 + 12)
Evento	(0 + 1) / (9 + 12)	(1 + 1) / (15 + 12)
Cuesta	(0 + 1) / (9 + 12)	(1 + 1) / (15 + 12)
Dinero	(0 + 1) / (9 + 12)	(1 + 1) / (15 + 12)
TOTAL:	**1**	**1**

27

En SCIKIT LEARN

- En scikit learn hay 3 tipos de clasificadores de Bayes Ingenuo:

Clasificador	Tipo de Features en X
Gaussiano	Continuo
Multinomial	Discreto (entero no negativo)
Bernoulli	Binario

Así que para poder usarlo es necesario que todas las características del vector x sean del mismo tipo.

- Scikit hace el smoothing adecuado de manera automática para evitar probabilidades iguales a 0 que impidan clasificar a un vector

28

- Para usarlo, simplemente se utiliza este código:

```
from sklearn.naive_bayes import GaussianNB
gnb = GaussianNB()
gnb.fit(X, y)
y_pred = gnb.predict(X)
```

29

Ventajas y desventajas

Ventajas	Desventajas
Es muy rápido de entrenar	Es necesario aplicar Smoothing para evitar probabilidades iguales a 0
Es fácil de implementar	Aunque sirve bien para clasificar,no es un buen estimador de probabilidades
Funciona bien con muchas dimensiones del vector de característica	Puede no generalizar bien a partir del training set
Como se asume independencia, el cálculo de la probabilidad de cada dimensión se puede hacer de manera independiente	Si las características están correlacionadas, puede tener resultados sesgados

30

Referencias

- https://blog.sicara.com/naive-bayes-classifier-sklearn-python-example-tips-42d100429e44
- https://www.kdnuggets.com/2017/03/email-spam-filtering-an-implementation-with-python-and-scikit-learn.html
- http://scikit-learn.org/stable/modules/naive_bayes.html

31