tema 10

Aprendizaje ballesiano

la probabilidad de un evento se puede aproximar por la frecuencia de que ocurra (estadistica pero para ballesianos no sirve)

Yo tengo una idea en la cabeza pero los datos luego me darán la razón o no

> ¿cuantos días va a llover este año?
>
> 70% porque lo digo yo, y luego los datos me harán cambiar de opinión

tienen la ventaja de que necesita muy pocos datos

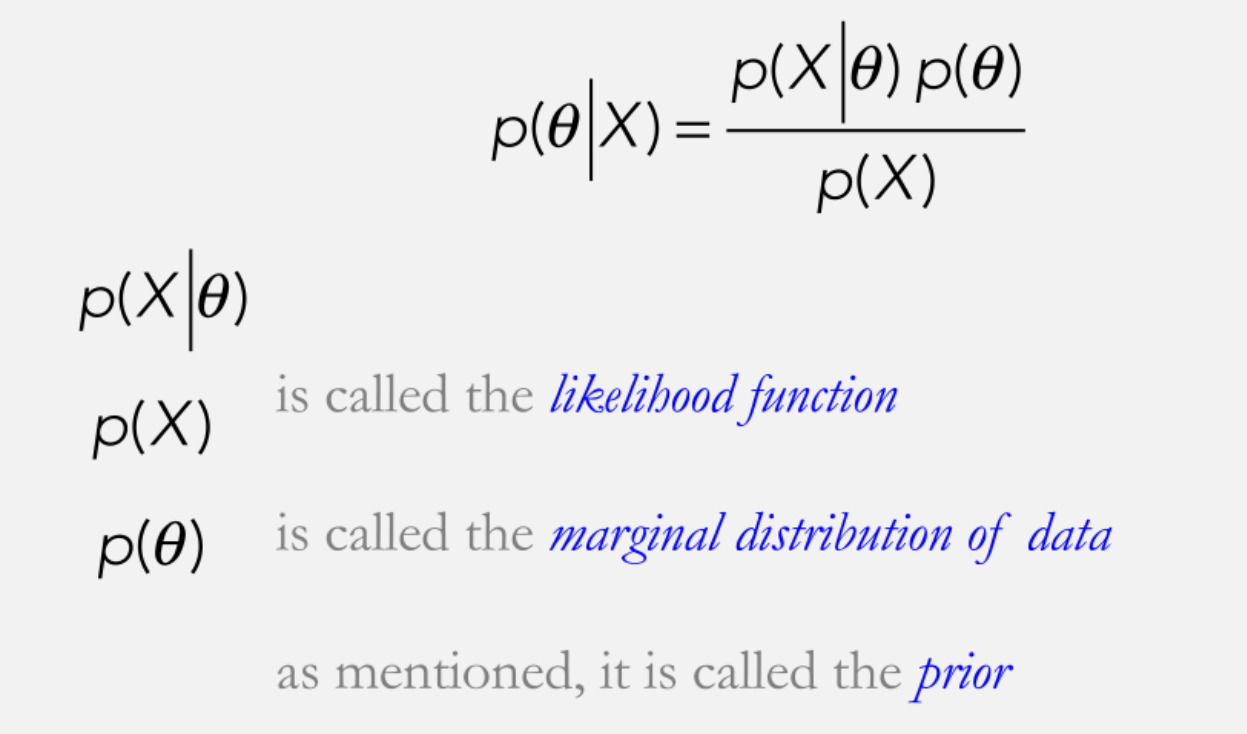
$$P(A|B) = \frac{P(B|A)\mathrm{P(A)}}{P(B)}$$

si es dificil calcularlo de esta forma, puede que al darle la vuelta sea más facil de calcularlo (el 95% de los casos)

Prior no informativa, asigna la misma probabilidad a cada uno de los posibles conjuntos de parámetros

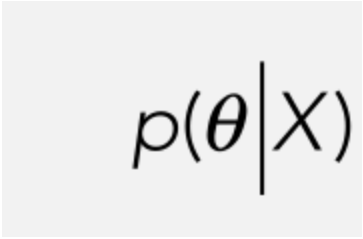
Cual es la probabilidad de unos parametros conocidos unos datos (esto es lo que vamos a calular)

Importante, pregunta segura (el nombre de estas)



- funcion de verosimilitud
- distribución marginal de los datos
- Probabilidad a priori

- Probabilidad a posteriori



Ahora pillo datos de un bd y los pongo para ver como cambia (en el numerador)

p(X) suele ser dificil de calcular entonces nos la vamos a cargar

no me importa el numero en concreto de una probabilidad, me interesa saber si con un conjunto de datos da una mayor probabilidad que dado otro conjunto sobre los mismos datos

$$p(\theta_1|X) = \frac{p(X|\theta_1)\,p(\theta_1)}{p(X)} \quad \text{vs.} \quad p(\theta_2|X) = \frac{p(X|\theta_2)\,p(\theta_2)}{p(X)}$$

miro si lo de la izquierda es mayor que lo da la derecha ,entonces, me da igual el denominador

procedimiento máximo a posteriori (MAP)

$$\theta_{MAP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|X) = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta)p(\theta)$$

(no tiene porque dar la suma de estos 1)

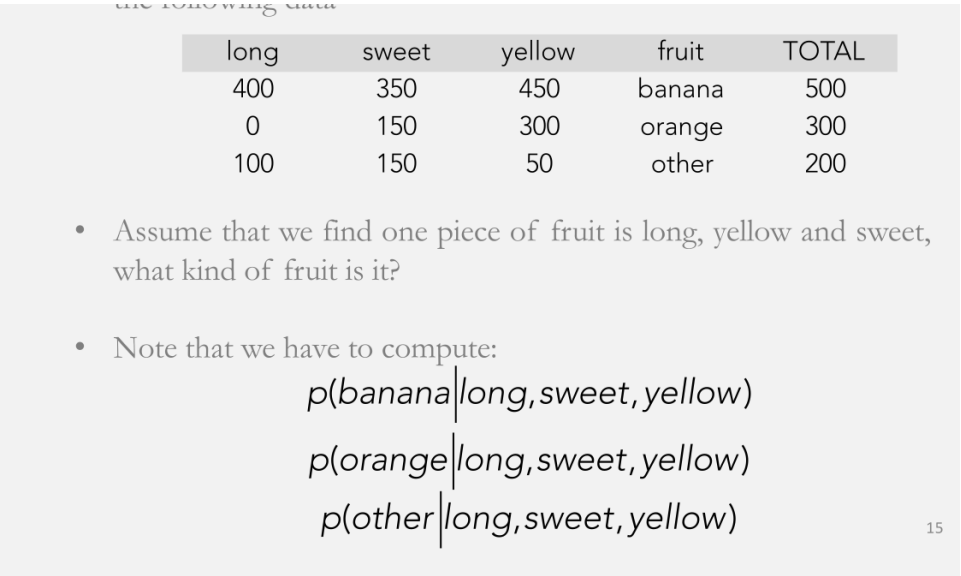
naive bayes

independencia condicional dada la clase

$$p(X_1, X_2|Y) = p(X_1|Y)p(X_2|Y)$$

esto es lo que vamos a utilizar (que no es lo mismo que seria en estadística)

ejemplo



long	sweet	yellow	fruit	TOTAL
400	350	450	banana	500
0	150	300	orange	300
100	150	50	other	200

- Assume that we find one piece of fruit is long, yellow and sweet, what kind of fruit is it?
- Note that we have to compute:

$$p(banana \mid long, sweet, yellow)$$
$$p(orange \mid long, sweet, yellow)$$
$$p(other \mid long, sweet, yellow)$$

15

- We have $p(banana|long,sweet,yellow) = \frac{p(long|banana)p(sweet|banana)p(yellow|banana)p(banana)}{p(long)p(sweet)p(yellow)} = \frac{0.8\times0.7\times0.9x0.5}{0.5\times0.65\times0.8} = 0.96$

$$p(orange|long,sweet,yellow) = \frac{p(long|orange)p(sweet|orange)p(yellow|orange)p(orange)}{p(long)p(sweet)p(yellow)} = \frac{0\times0.7\times0.9x0.3}{0.5\times0.65\times0.8} = 0$$

$$p(other|long,sweet,yellow) = \frac{p(long|other)p(sweet|other)p(yellow|other)p(other)}{p(long)p(sweet)p(yellow)} = \frac{0.5\times0.75\times0.25x0.2}{0.5\times0.65\times0.8} = 0.07$$

16

da igual los denominadores y no deberian estár puestos

- We have $p(banana|long,sweet,yellow) = \frac{p(long|banana)p(sweet|banana)p(yellow|banana)p(banana)}{} =$

$$0.8\times0.7\times0.9x0.5$$

calculo esto para todos y me quedo con el que tenga mayor probabilidad

ventajas de este método, reduce muchisimo la carga computacional

Tenemos el problema de que si una probabilidad es 0, todo va a ser 0

no tengo naranjas largas ya que tengo una muestra muy pequeña que no tiene ese caso

Esto es el problema de 0 probabilidad

solucion:

- For example, assume that $X=\{x_1,x_2\}$, $\theta=\{\theta_1, \theta_2\}$ and we have the sample

X	θ	p(x,θ)
1	1	p(1,1)=3/6
1	2	p(1,2)=1/6
2	1	p(2,1)=2/6
2	1	p(2,2)=0
1	1	p(θ=1)=5/6
1	1	p(θ=2)=1/6

23

me invento datos pero de forma justa

- Alternatively, using *Laplace smoothing* ($k=1$):

X	θ
1	1
1	2
2	1
2	1
1	1
1	1
1	1
1	2
2	1
2	2

p(1,1)=4/10
p(1,2)=2/10
p(2,1)=3/10
p(2,2)=1/10
p(X=1)=6/10
p(X=2)=4/10

ventajas

- Advantages:

1. Easy to understand
2. Can be applied even with small datasets
3. If conditional independence is true, Naive Bayes is the BEST classifier one may build
4. Excellent performance in most of the cases
5. Computationally very efficient (*eager*)
6. Insensitive to irrelevant features
7. Robust to outliers (particularly in the discrete case)
8. Robust to missing data
9. Compared to other algorithms it does not need to be finely "tuned", there are no hiper parameters (except the "prior")
10. New data can easily be incorporated without changing the structure of the model

26

desventajas

- Disadvantages:

1. Zero probability problem requires artificially modifying the distribution
2. The performance degrades when there is a high interacion among the variables
3. It does not provide “true” estimated probabilities
4. For continuous variables one needs to either use some model or discretize the variable
5. Can not be used in regression problems
6. It can not learn intractions between features
7. Different priors will produce different models