Lecture 07

Naive Bayes

https://github.com/dalcimar/MA28CP-Intro-to-Machine-Learning

Eventos dependentes

Seja uma caixa contendo **2 bolas brancas** e **3 pretas**. Considere o experimento da retirada aleatória e observância da cor das bolas, com representação de **Bernoulli** (X_1) de ocorrência

- $X_1=0$ caso a bola retirada seja branca
- $X_1=1$ caso a bola retirada seja preta

Pela abordagem clássica, a probabilidade da bola ser branca é:

$$P(X_1 = 0) = \frac{2}{5} \qquad P(X_1 = 1) = \frac{3}{5}$$

Eventos dependentes

Após a primeira retirada, e **sem reposição**, você retira **mais uma bola**. Qual a probabilidade da segunda bola ser branca?

- A resposta mais intuitiva é **DEPENDE**. De fato, a probabilidade da segunda retirada (X_2) está **condicionada** ao resultado da primeira.
  - Caso a primeira seja branca, restam mais 1 branca e as demais 3 pretas, e, nesse caso, a probabilidade de uma segunda retirada branca é ¼
  - Caso a primeira tenha sido preta, seria ¾

De maneira geral, a probabilidade do evento B condicionada ao evento A é representada por P(B/A). Nesse caso, temos que:

$$P(X_2 = 0 | X_1 = 0) = \frac{1}{4} \qquad P(X_2 = 0 | X_1 = 1) = \frac{2}{4}$$

Dessa forma, dizemos que o evento X_2 é dependente do evento X_1.

Probabilidade conjunta

A **probabilidade conjunta** dos eventos A e B é expressa pela regra do produto: $P(B \cap A) = P(B|A)P(A)$

- também pode ser expressa por $P(B \cap A) = P(B,A)$

Eventos independentes

Considere um experimento, o lançamento de uma moeda honesta e a observância da face virada para cima. Considere a representação **Bernoulli** (X_1) de ocorrência:

- $X_1=0$ caso resultado do lançamento seja coroa
- $X_1=1$ caso o resultado do lançamento seja cara

Por ser justa, a probabilidade do resultado cara e o de coroa é

$$P(X_1 = 1) = \frac{1}{2} \qquad P(X_1 = 0) = \frac{1}{2}$$

Eventos independentes

Em seguida, a moeda é lançada novamente e o resultado representado por X2. Perceba que nesse caso, a probabilidade da ocorrência de cara no segundo lançamento também é igual a $P(X_2=1)=½$ por não ser afetado pelo resultado do primeiro lançamento.

Dizemos que X_2 é independente de X_1, e nesse caso, temos que a regra do produto é:

$$P(X_2, X_1) = P(X_2|X_1)P(X_1)$$

$$P(X_2, X_1) = P(X_2)P(X_1)$$

Permutabilidade

Permutabilidade é a propriedade da **alteração no ordenamento de realizações** em uma **sequência de eventos sem que a probabilidade conjunta seja alterada**.

Considere o lançamento de 5 moedas não viciadas, com representação semelhante à apresentada anteriormente, sendo observado o seguinte resultado (1,0,1,1,0), isto é, (cara,coroa,cara,cara,coroa). Sabemos que os lançamentos são independentes e que a probabilidade conjunta desse evento é:

$$P(1,0,1,1,0) = P(X_5 = 1) \times P(X_4 = 0) \times P(X_3 = 1) \times P(X_2 = 1) \times P(X_1 = 0) = \frac{1}{2^5}$$

Nessa situação, caso fosse observado o evento (1,1,1,0,0), a probabilidade não seria alterada, pois $P(1,0,1,1,0)=P(1,1,1,0,0)=1/2^5$

Dessa forma, **a ordem da ocorrências dos resultados cara não altera a probabilidade conjunta**, desde que seja a mesma quantidade de resultados de “sucesso”. Esse é uma versão não rigorosa do Teorema de De Finetti.

De maneira geral, a **independência** de uma sequência de eventos **garante a permutabilidade** da mesma.

Permutabilidade

No entanto, a **independência não é condição necessária** para a permutabilidade, **apenas suficiente**.

- Isto é, **ainda que uma sequência não seja formada por eventos independente, é possível que sejam permutáveis**.

Considere a mesma representação do exemplo da caixa com bolas apresentado anteriormente para o caso da retirada de 5 bolas sem reposição.

- Como visto, **esses eventos não são independentes**, pois a probabilidade de um determinado evento na sequência, depende do resultado observado nos eventos anteriores.

Permutabilidade

Dessa forma, vamos verificar se o evento (1,0,1,1,0) é permutável. Inicialmente, vamos calcular a probabilidade P(1,0,1,1,0):

$$P(1,0,1,1,0) = P(X_1 = 1) \times P(X_2 = 0|X_1 = 1) \times P(X_3 = 1|X_2 = 0, X_1 = 1) \times$$
$$P(X_4 = 1|X_3 = 1, X_2 = 0, X_1 = 1) \times P(X_5 = 0|X_4 = 1, X_3 = 1, X_2 = 0, X_1 = 1)$$

$$P(1,0,1,1,0) = \frac{3}{5} \times \frac{2}{4} \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{10}$$

Agora, vamos calcular a probabilidade P(1,1,1,0,0), alterando o resultado do segundo e o quarto evento:

$$P(1,1,1,0,0) = P(X_1 = 1) \times P(X_2 = 1|X_1 = 1) \times P(X_3 = 1|X_2 = 1, X_1 = 1) \times$$
$$P(X_4 = 0|X_3 = 1, X_2 = 1, X_1 = 1) \times P(X_5 = 0|X_4 = 0, X_3 = 1, X_2 = 1, X_1 = 1)$$

$$P(1,1,1,0,0) = \frac{3}{5} \times \frac{2}{4} \times \frac{1}{3} \times \frac{2}{2} \times \frac{1}{1} = \frac{1}{10}$$

Teorema de Bayes

Considerando os eventos A e B permutáveis, o termo P(A∩B) é igual a P(B∩A) e, dessa forma, pode ser escrita como:

$$P(A \cap B) = P(B \cap A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

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

Nesse caso, o probabilidade **P(A)** é denominada **probabilidade a priori**, isto é, a informação sobre o evento A antes que se soubesse algo sobre o evento B. Mais adiante, quando se tenha conhecimento sobre B, a probabilidade relacionada ao evento A deve ser atualizada pela probabilidade do evento B. A probabilidade **P(A/B)** é agora denominada **probabilidade a posteriori**.

Teorema de Bayes

Example

- A doctor knows that meningitis causes stiff neck 50% of the times
- Prior probability of any patient having meningitis is 1/50000
- Prior probability of any patient having stiff neck is 1/20

If a patient has stiff neck (evidence), what's the posterior probability he/she has meningitis?

$$P(M \mid S) = \frac{P(S \mid M)P(M)}{P(S)} = \frac{0.5 \times 1/50000}{1/20} = 0.0002$$

Bayes Theorem

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Likelihood
Probability of collecting this data when our hypothesis is true

Prior
The probability of the hypothesis being true before collecting data

$$P(H|D) = \frac{P(D|H)\, P(H)}{P(D)}$$

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?

Naive Bayes classifier

There **is not a single algorithm**, but a **family of algorithms** based on **a common principle**:

- all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable
    - a fruit may be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier considers each of these features to contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.
- In many practical applications, parameter estimation for naive Bayes models uses the method of **maximum likelihood**; in other words, one can work with the naive Bayes model without accepting Bayesian probability or using any Bayesian methods.

Naive Bayes classifier

Abstractly, naïve Bayes is a conditional probability model: given a problem instance to be classified, represented by a vector **x** representing some n features (independent variables), it assigns to this instance probabilities

$$p(C_k \mid x_1, \ldots, x_n)$$

Using Bayes' theorem, the conditional probability can be decomposed as

$$p(C_k \mid \mathbf{x}) = \frac{p(C_k)\, p(\mathbf{x} \mid C_k)}{p(\mathbf{x})}$$

Naive Bayes classifier

The numerator is equivalent to the joint probability model

$$p(C_k, x_1, \dots, x_n)$$

which can be rewritten as follows, using the chain rule for repeated applications of the definition of conditional probability:

$$\begin{aligned}
p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \\
&= \cdots \\
&= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k)
\end{aligned}$$

Naive Bayes classifier

Now the "naïve" conditional independence assumptions come into play: assume that all features in **x** are mutually independent, conditional on the category C_k. Under this assumption,

$$\begin{aligned} p(C_k \mid x_1, \dots, x_n) &\propto p(C_k, x_1, \dots, x_n) \\ &\propto p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &\propto p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

$$\boxed{P(B \mid A_1, \dots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}}$$

$$\hat{y} = \underset{k \in \{1, \dots, K\}}{\operatorname{argmax}}\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k).$$

Naive Bayes classifier

Para várias variáveis aleatórias A_1, A_2, ..., A_n, e B: raciocínio análogo

- $P(B|A_1,A_2,...,A_n) = P(A_1,A_2,...,A_n|B)*P(B) / P(A_1,A_2,...,A_n)$

Se as vars aleatórias $A_1,A_2,...,A_n$ forem independentes entre si tem-se:

- $P(A_1,A_2,...,A_n) = P(A_1)*P(A_2)* \ldots *P(A_n)$ [independência]
- $P(A_1,A_2,...,A_n|B) = P(A_1|B)*P(A_2|B)* \ldots * P(A_n|B)$ [independência condicional]

Naive Bayes classifier

Naive Bayes é um algoritmo que utiliza o Teorema de Bayes com a hipótese de independência entre atributos

- Porque assumir independência entre atributos $A_1,\ldots,A_n$?
  - estimar probabilidades conjuntas $P(A_1,A_2,...,A_n)$ e $P(A_1,A_2,...,A_n|B)$ demandaria uma quantidade mínima de exemplos de cada combinação possível de valores de $A_1,A_2,...,A_n$
  - impraticável, especialmente para quantidades elevadas de atributos !
- Apesar da hipótese ser quase sempre violada, o método (Naive Bayes) se mostra bastante competitivo na prática!

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

$$P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	???

P(Yes|Sunny, Cool, High, True) = (2/9 × 3/9 × 3/9 × 3/9 × 9/14) / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = (3/5 × 1/5 × 4/5 × 3/5 × 5/14) / P(Sunny, Cool, High, True)

P(Yes|Sunny, Cool, High, True) = **0.0053** / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = **0.0206** / P(Sunny, Cool, High, True)

Play = No

$$\hat{y} = \underset{k\in\{1,\ldots,K\}}{\operatorname{argmax}}\ p(C_k)\prod_{i=1}^{n} p(x_i \mid C_k).$$

Problema da Frequência Zero

O que acontece se um determinado valor de atributo não aparece na base de treinamento, mas aparece no exemplo de teste?

- Por exemplo: “Outlook = Overcast” para classe “No”
  - Probabilidade correspondente será zero
    - P(Overcast | “No”) = 0
  - Probabilidade a posteriori será também zero!
    - P(“No” | Overcast, ...) = 0
- Não importa as probabilidades referentes aos demais atributos !
- Muito radical, especialmente considerando que a base de treinamento pode não ser totalmente representativa
- Por exemplo, classes minoritárias com instâncias raras

Problema da Frequência Zero

Possível solução (Estimador de Laplace):

- Adicionar 1 unidade fictícia para cada combinação de valor-classe
  - Como resultado, probabilidades nunca serão zero!
  - Exemplo (atributo Outlook – classe No):

$$\frac{3+1}{5+3} \qquad \frac{0+1}{5+3} \qquad \frac{2+1}{5+3}$$

Sunny ***Overcast*** ***Rainy***

- Nota: Deve ser feito para todas as classes, para não inserir viés nas probabilidades de apenas uma classe

Problema da Frequência Zero

Solução mais geral (Estimativa m):

- Adicionar múltiplas unidades fictícias para cada combinação de valor-classe
- Exemplo (atributo Outlook – classe No):

$$\frac{3+\frac{m}{3}}{5+m} \qquad \frac{0+\frac{m}{3}}{5+m} \qquad \frac{2+\frac{m}{3}}{5+m}$$

Sunny ***Overcast*** ***Rainy***

Solução ainda mais geral:

- Substituir o termo 1/n no numerador (onde n é o no. de valores do atributo) por uma probabilidade p qualquer

Valores Ausentes

Treinamento:

- excluir exemplo do conjunto de treinamento

Classificação:

- considerar apenas os demais atributos

Exemplo:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	???

Verossimilhança para "Yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Verossimilhança para "No" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

Probabilidade Estimada ("Yes") = 0.0238 / (0.0238 + 0.0343) = 41%

Probabilidade Estimada ("No") = 0.0343 / (0.0238 + 0.0343) = 59%

Atributos Numéricos

Alternativa 1: **Discretização**

Alternativa 2: **Assumir** ou **estimar** alguma **função de densidade de probabilidade** para estimar as probabilidades

- Usualmente distribuição Gaussiana (Normal)

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Atributos Numéricos

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, ...	85, ...	80, ...	95, ...					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

Valor de densidade:

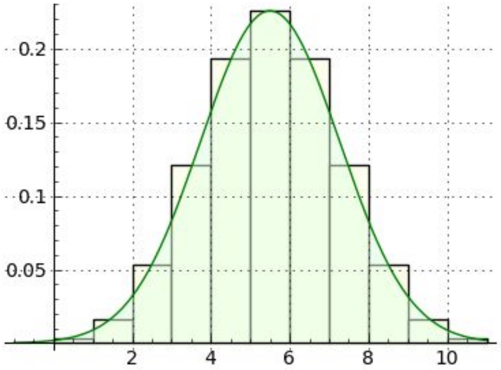
$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}6.2} e^{-\frac{(66-73)^2}{2\times 6.2^2}} = 0.0340$$

Porque o Teorema de Bayes convenientemente permite usar o valor de densidade de probabilidade para estimar a probabilidade de um valor pontual (teoricamente nula)... ?

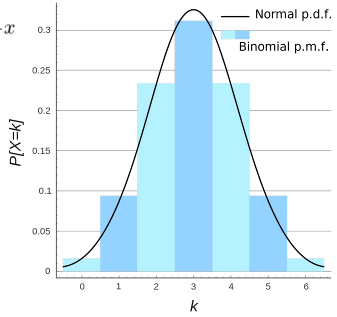
Atributos Numéricos

the binomial distribution with parameters n and p, denoted B(n,p)} is the discrete probability distribution of the number of successes in a sequence of n independent experiments

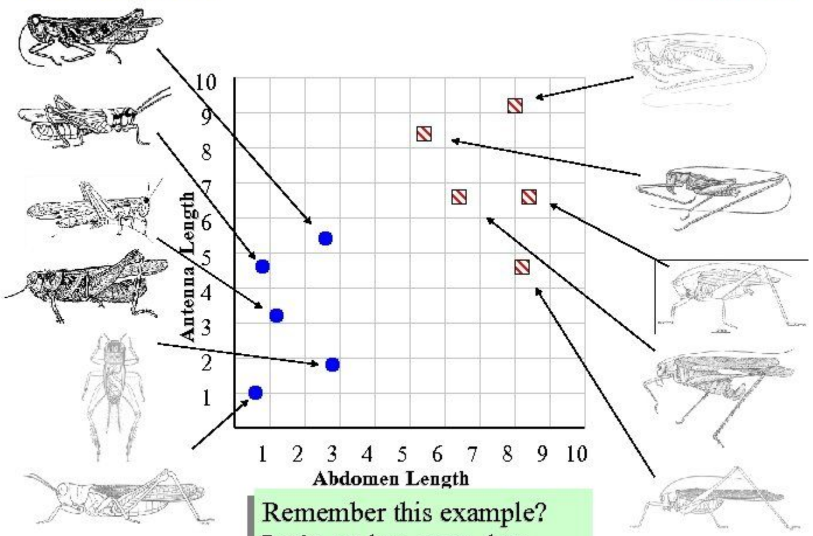
- If n is large enough, then the skew of the distribution is not too great. In this case a reasonable approximation is given by the normal distribution



$$\begin{aligned}\sum_{x=0}^{n} f(x) &= \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \\ &= (p + (1-p))^n \\ &= (1)^n \\ &= 1\end{aligned}$$

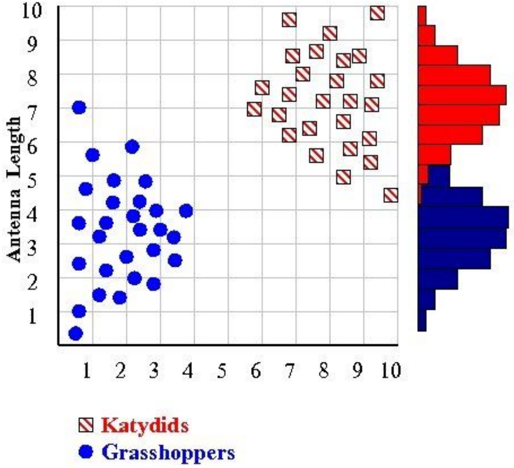


Grasshoppers **Katydids**

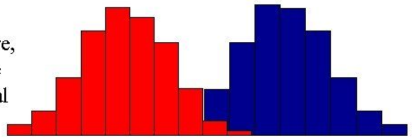


Remember this example?
Let's get lots more data…

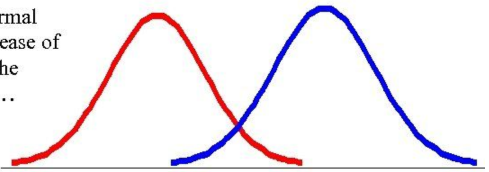
With a lot of data, we can build a histogram. Let us just build one for "Antenna Length" for now…



We can leave the histograms as they are, or we can summarize them with two normal distributions.



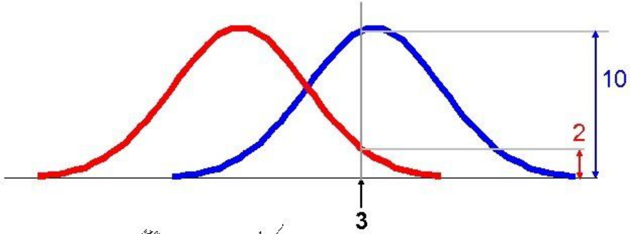
Let us us two normal distributions for ease of visualization in the following slides…

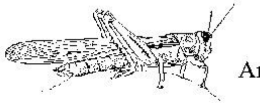


$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **3**) = 10 / (10 + 2) = 0.833

P(**Katydid** | **3**) = 2 / (10 + 2) = 0.166



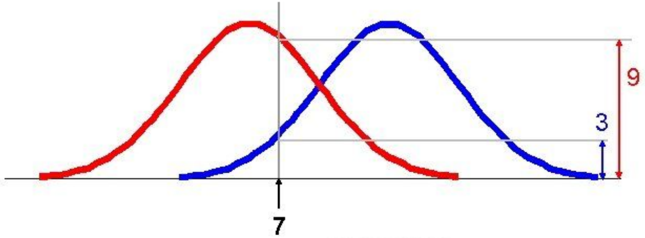


Antennae length is **3**

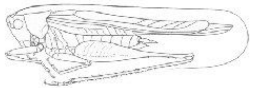
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **7**) = 3 / (3 + 9) = 0.250

P(**Katydid** | **7**) = 9 / (3 + 9) = 0.750



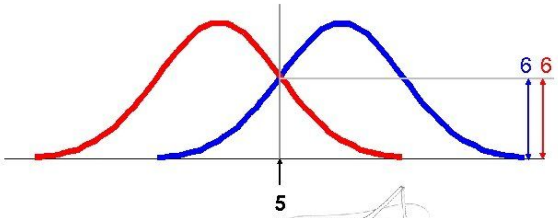
Antennae length is **7**



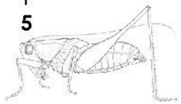
$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **5**) = 6 / (6 + 6) = 0.500

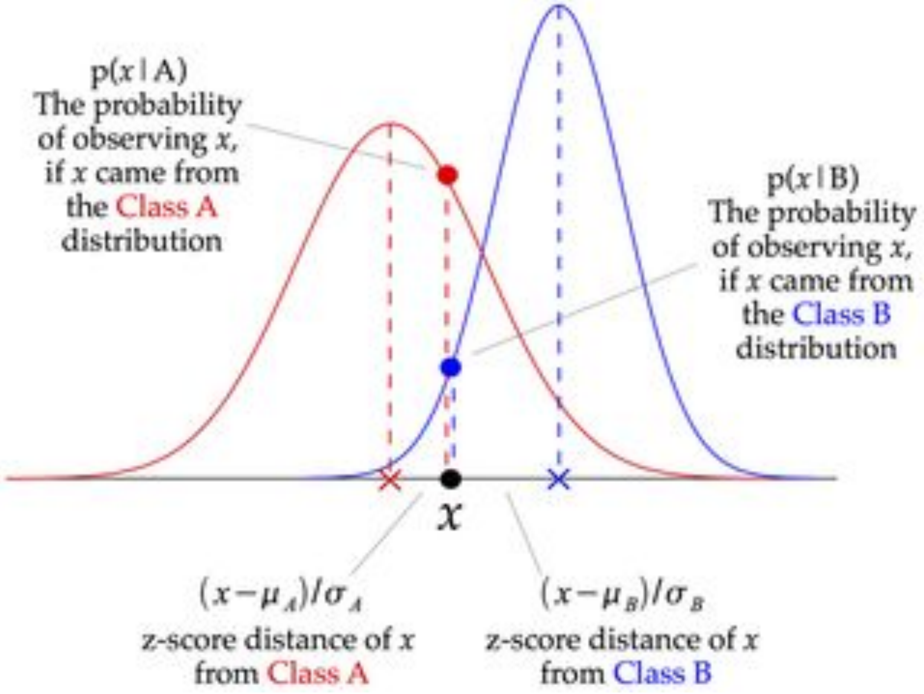
P(**Katydid** | **5**) = 6 / (6 + 6) = 0.500

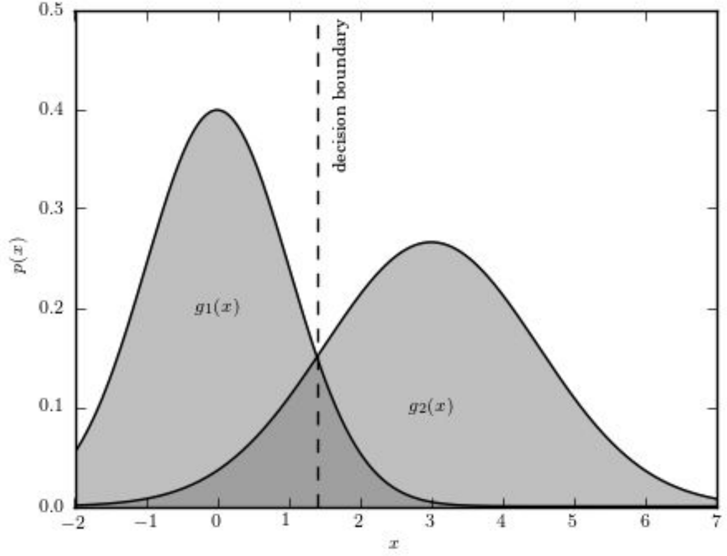


Antennae length is **5**



Atributos Numéricos





Atributos Numéricos

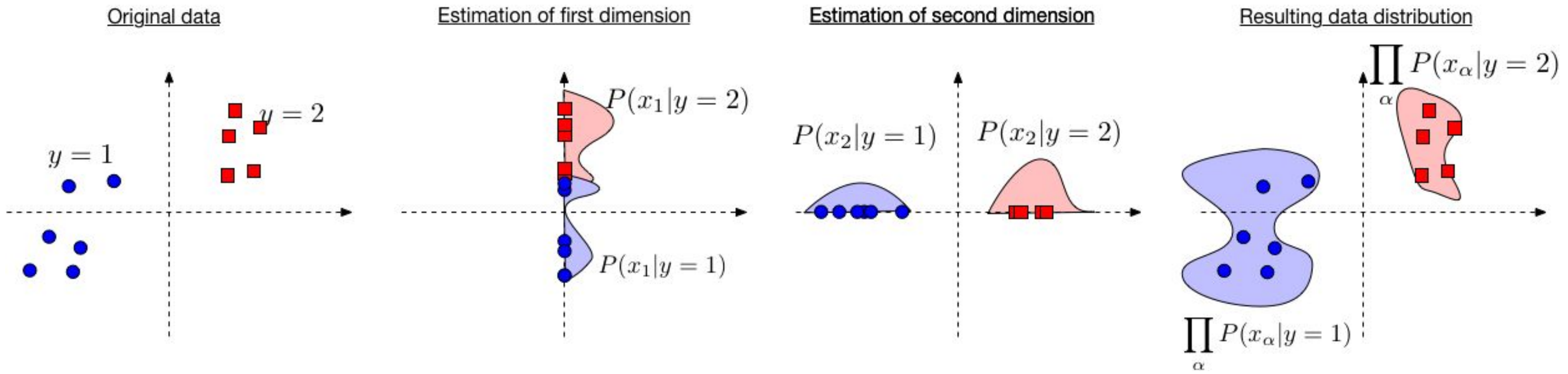
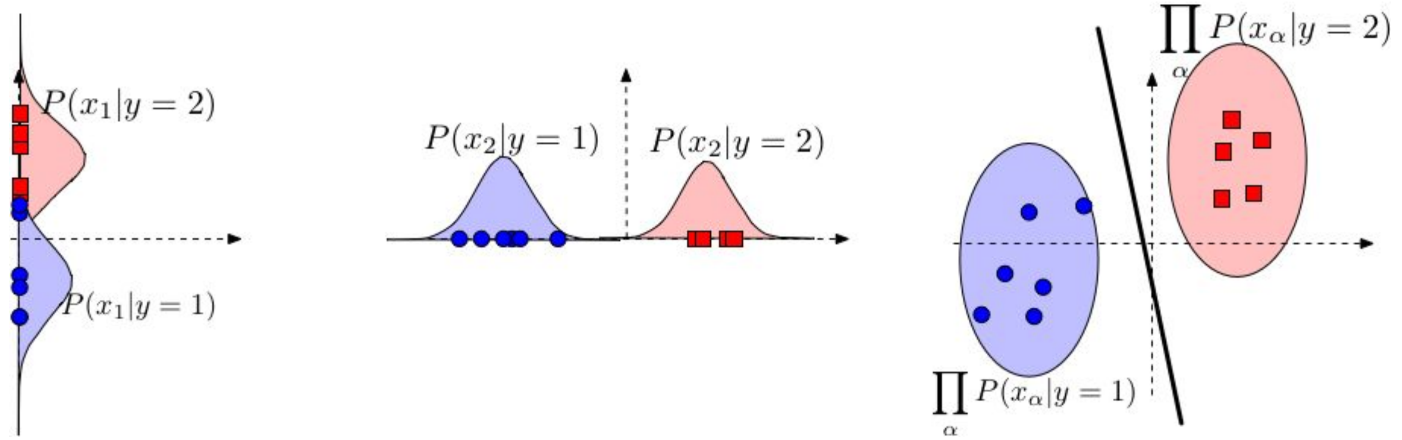


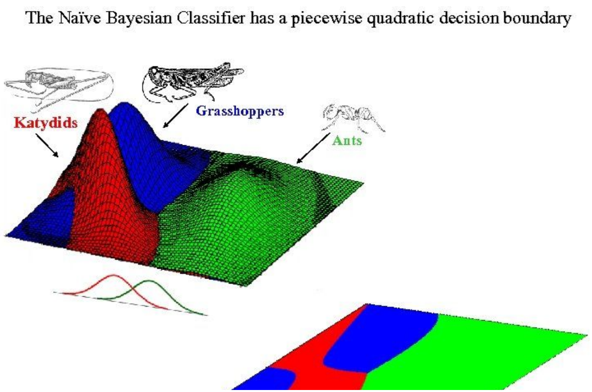
Illustration behind the Naive Bayes algorithm. We estimate P(**x**|y) independently in each dimension (middle two images) and then obtain an estimate of the full data distribution by assuming conditional independence (very right image).

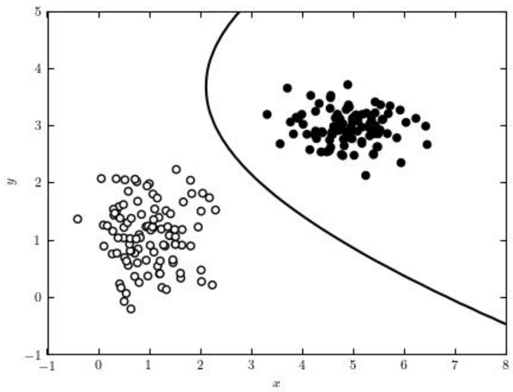
Atributos Numéricos

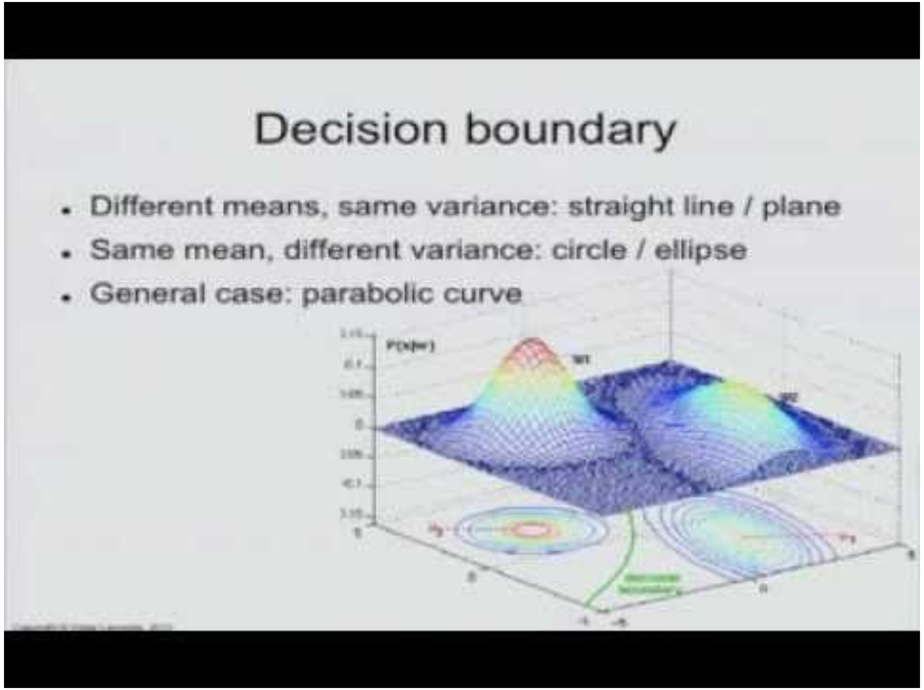


Naive Bayes leads to a linear decision boundary in many common cases. Illustrated here is the case where P(**x**|y) is Gaussian and where **σ** is identical for all classes (but can differ across dimensions). The boundary of the ellipsoids indicate regions of equal probabilities P(x|y). The red decision line indicates the decision boundary where P(y=1|x)=P(y=2|x).

Naive Bayes decision boundary



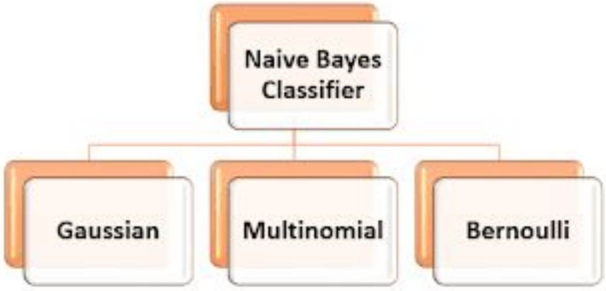




https://www.youtube.com/watch?v=0oca6pC3f0M
IAML5.10: Naive Bayes decision boundary

Parameter estimation

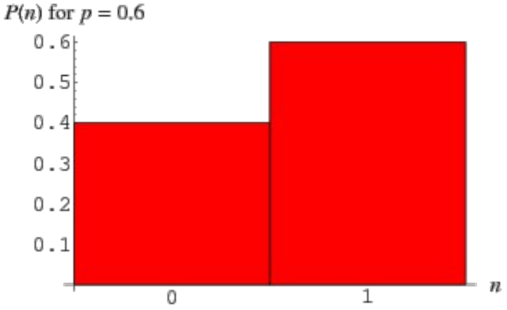
Now that we know how we can use our assumption to make the estimation of P(y|x) tractable. There are 3 (or 4) notable cases in which we can use our naive Bayes classifier.



Case #1: Bernoulli features

Algorithm for data that is distributed according to multivariate Bernoulli distributions; i.e., there may be multiple features but each one is assumed to be a binary-valued (Bernoulli, boolean) variable.

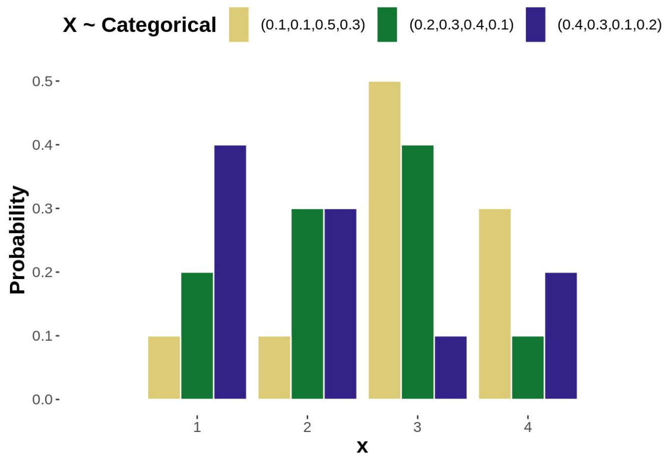
$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i)$$



Case #2: Categorical features

Each feature α falls into one of K categories. (Note that the case with binary features is just a specific case of this, where K=2.) An example of such a setting may be medical data where one feature could be gender (male / female) or marital status (single / married / widowed).

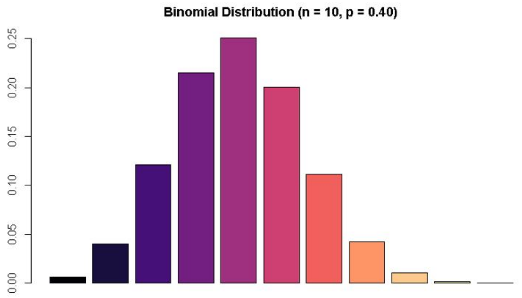
$$P(x_i = t \mid y = c \,;\, \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i},$$

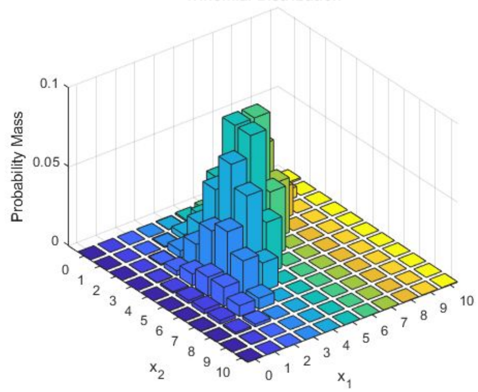


Case #3: Multinomial features

If feature **values don't represent categories** (e.g. male/female) **but counts** we need to use a different model. E.g. in the text document categorization, feature value x_i=j means that in this particular document x the ith word in my dictionary **appears j times**.

- Let us consider the example of spam filtering. Imagine the ith word is indicative towards "spam". Then if x_i=10 means that this email is likely spam (as word α appears 10 times in it). And another email with x'_i=20 should be even more likely to be spam (as the spammy word appears twice as often).
- With categorical features this is not guaranteed. It could be that the training set does not contain any email that contain word i exactly 20 times. In this case you would simply get the hallucinated smoothing values for both spam and not-spam - and the signal is lost. We need a model that incorporates our knowledge that features are counts - this will help us during estimation (you don't have to see a training email with exactly the same number of word occurances)



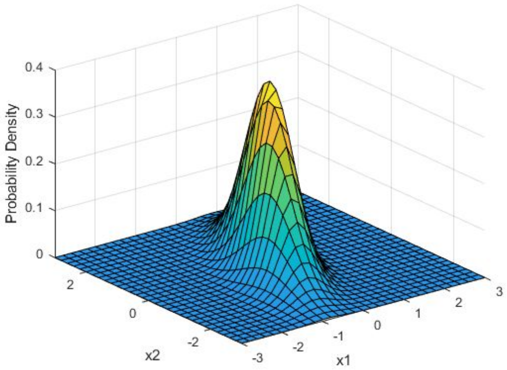


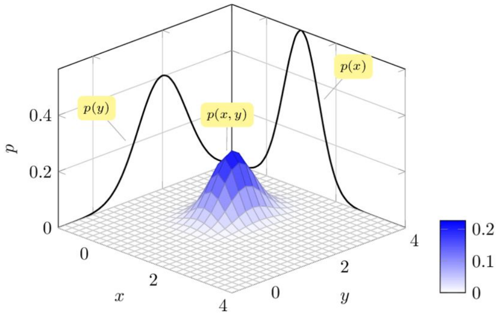
$$\hat{\theta}_{yi} = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

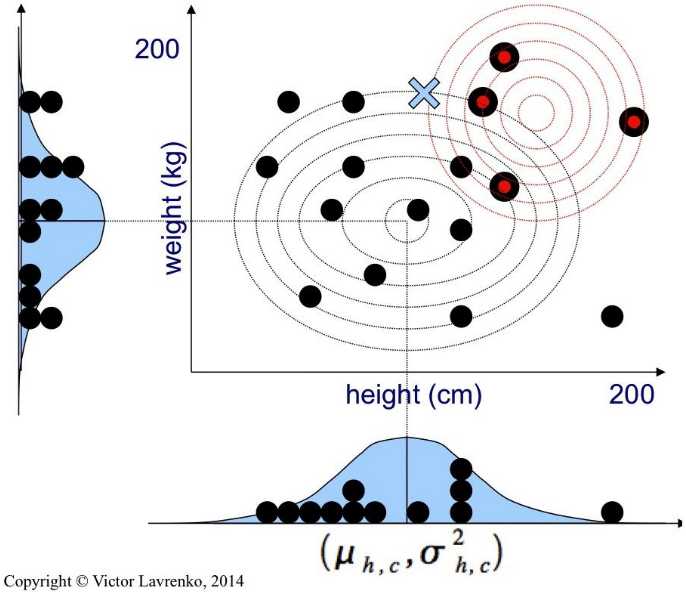
Case #4: Continuous features (Gaussian Naive Bayes)

Note that the model is based on our assumption about the data - that each feature i comes from a class-conditional Gaussian distribution.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$



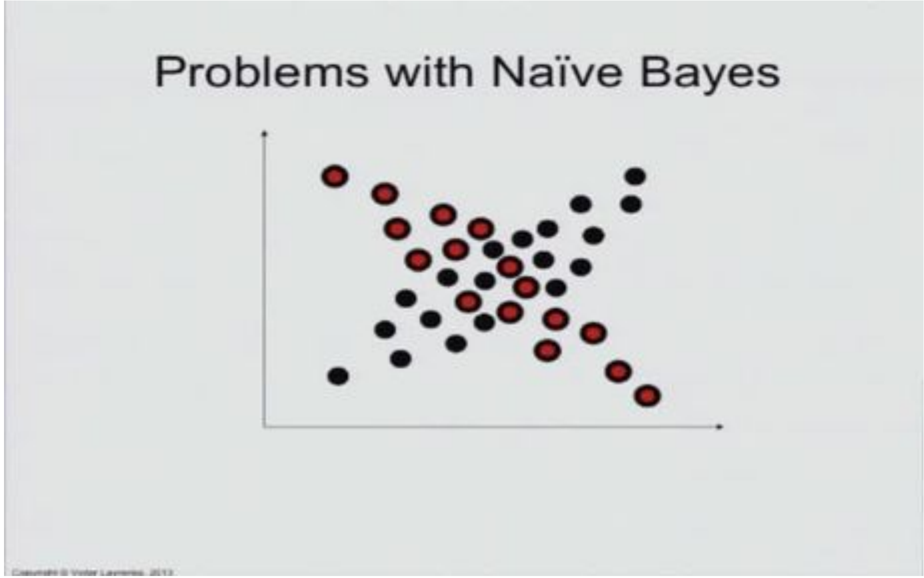




Naive Bayes: Características

- **Robusto a ruídos** isolados
    - Por exemplo, outliers ilegítimos
    - Afetam pouco o cálculo das probabilidades
- Robusto a atributos irrelevantes
    - Afetam pouco as probabilidades relativas entre classes
- Capaz de classificar instâncias com valores ausentes
- Assume que atributos são igualmente importantes
- Desempenho pode ser (mas muitas vezes não é) afetado pela presença de atributos correlacionados

Where naive Bayes fail?



https://www.youtube.com/watch?v=feBKiAdhYkc
IAML5.11: Example where Naive Bayes fails

Seleção de Atributos: Wrapper Naive Bayes

Algoritmo Guloso:

- Selecione o melhor classificador Naive Bayes com um único atributo (avaliando todos em um conjunto de dados de teste)
- Enquanto houver melhora no desempenho do classificador faça
  - Selecione o melhor classificador Naive Bayes com os atributos já selecionados anteriormente adicionados a um dentre os atributos ainda não selecionados

Nota: Apesar de ser um wrapper, o algoritmo acima é relativamente rápido devido à sua simplicidade e à eficiência computacional do Naive Bayes !