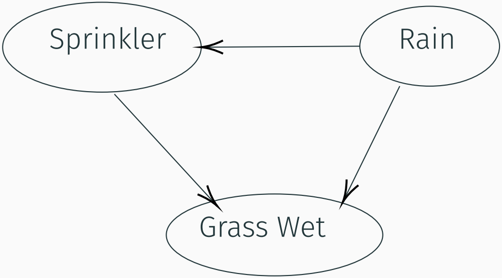
Naive Bayes

Nipun Batra

July 5, 2020

IIT Gandhinagar

Bayesian Networks



- Nodes are random variables.
- Edges denote direct impact

Example

- Grass can be wet due to multiple reasons:
  - Rain
  - Sprinkler
- Also, if it rains, then sprinkler need not be used.

Bayesian Nets

$P(X_1, X_2, X_3, \ldots, X_N)$ denotes the joint probability, where X_i are random variables.

$$P(X_1, X_2, X_3, \ldots, X_N) = \Pi_{k=1}^{N} P(X_k | parents(X_k))$$

$$P(S, G, R) = P(G|S, R)P(S|R)P(R)$$

Spam Email Classification

- $y \in \{0, 1\}$ where 0 means not spam and 1 means spam

Spam Email Classification

- $y \in \{0, 1\}$ where 0 means not spam and 1 means spam
- From the emails construct a vector X.

Spam Email Classification

- $y \in \{0, 1\}$ where 0 means not spam and 1 means spam
- From the emails construct a vector X.
- $$\begin{bmatrix} a \\ an \\ \vdots \\ \text{computer} \\ \vdots \\ \text{lotery} \\ \vdots \\ 200 \end{bmatrix} \} \text{ N words}$$

Spam Email Classification

- $y \in \{0, 1\}$ where 0 means not spam and 1 means spam
- From the emails construct a vector X.
- $$\left.\begin{bmatrix} a \\ an \\ \vdots \\ \text{computer} \\ \vdots \\ \text{lotery} \\ \vdots \\ 200 \end{bmatrix}\right\} \text{N words}$$
- The vector has ones if the word is present, and zeros is the word is absent.

Spam Email Classification

- $y \in \{0, 1\}$ where 0 means not spam and 1 means spam
- From the emails construct a vector X.
- $\left.\begin{bmatrix} a \\ an \\ \vdots \\ \text{computer} \\ \vdots \\ \text{lotery} \\ \vdots \\ 200 \end{bmatrix}\right\}$ N words
- The vector has ones if the word is present, and zeros is the word is absent.
- Each email corresponds to vector/feature of length N containing zeros or ones.

Naive Bayes

- Classification model

Naive Bayes

- Classification model
- Scalable

Naive Bayes

- Classification model
- Scalable
- Generative and Bayesian

Naive Bayes

- Classification model
- Scalable
- Generative and Bayesian
- Usually a simple/good baselines

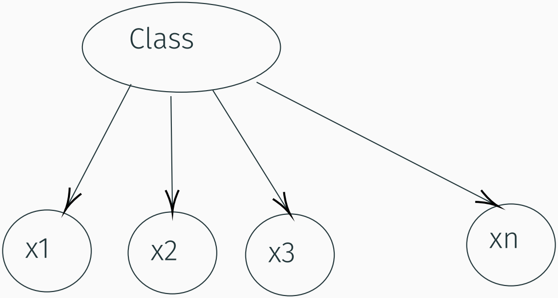
Naive Bayes

- Classification model
- Scalable
- Generative and Bayesian
- Usually a simple/good baselines
- We want to model $P(class(y) \mid \text{features } (\text{x}))$

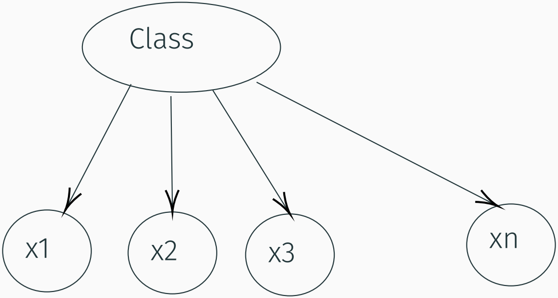
Naive Bayes

- Classification model
- Scalable
- Generative and Bayesian
- Usually a simple/good baselines
- We want to model $P(class(y) \mid \text{features (x)})$
- We can use Bayes rule as follows:
  $P(class(y) \mid \text{ features (x) }) = \frac{P(\text{ features (x) } | class(y))P(class(y))}{P(\text{ features (x) })}$

Quick Question

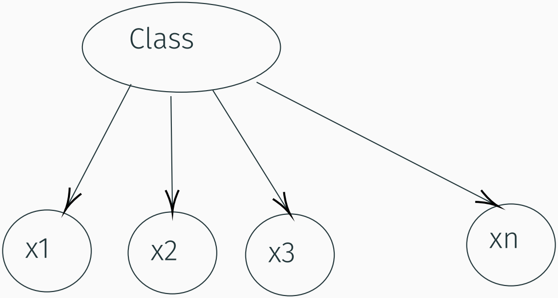


Quick Question



$$P(x_1, x_2, x_3, \ldots, x_N | y) = P(x_1|y)P(x_2|y)\ldots P(x_N|y)$$

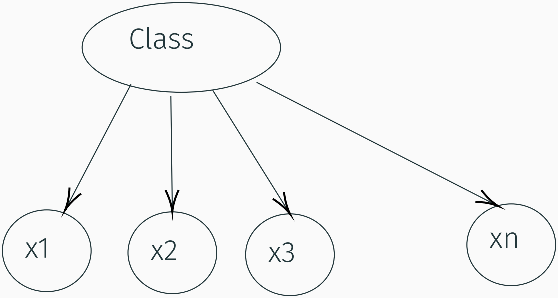
Quick Question



$$P(x_1, x_2, x_3, \ldots, x_N|y) = P(x_1|y)P(x_2|y)\ldots P(x_N|y)$$

Why is Naive Bayes model called Naive?

Quick Question



$$P(x_1, x_2, x_3, \ldots, x_N|y) = P(x_1|y)P(x_2|y)\ldots P(x_N|y)$$

Why is Naive Bayes model called Naive?

Naive assumption x_i and x_{i+1} are independent given y

$$\text{i.e. } p\left(x_2 \mid x_1, y\right) = p\left(x_2 \mid y\right)$$

Frame Title

It assumes that the features are independent during modelling, which is generally not the case.

What do we need to predict?

$$P(y|x_1, x_2, \ldots, x_N) = \frac{P(x_1, x_2, \ldots, x_N|y)P(y)}{P(x_1, x_2, \ldots, x_N)}$$

Spam Mail Classification

Probability of x_i being a spam email

$$P(x_i = 1|y = 1) = \frac{\text{Count}(x_i = 1 \text{ and } y = 1)}{\text{Count } (y = 1)}$$

Similarly,

$$P(x_i = 0|y = 1) = \frac{\text{Count}(x_i = 0 \text{ and } y = 1)}{\text{Count } (y = 1)}$$

Spam Mail classification

$$P(y = 1) = \frac{\text{Count } (y = 1)}{\text{Count } (y = 1) + \text{Count } (y = 0)}$$

Similarly,

$$P(y = 0) = \frac{\text{Count } (y = 0)}{\text{Count } (y = 1) + \text{Count } (y = 0)}$$

Example

lets assume that dictionary is $[w_1, w_2, w_3]$

Index	w_1	w_2	w_3	y
1	0	0	0	1
2	0	0	0	0
3	0	0	0	1
4	1	0	0	0
5	1	0	1	1
6	1	1	1	0
7	1	1	1	1
8	1	1	0	0
9	0	1	1	0
10	0	1	1	1

Spam Classification

if y=0

- $P(w_1 = 0|y = 0)$ = $\frac{3}{5}$ = 0.6
- $P(w_2 = 0|y = 0)$ = $\frac{2}{5}$ = 0.4
- $P(w_3 = 0|y = 0)$ = $\frac{3}{5}$ = 0.6

P(y=0) = 0.5
Similarly, if y=1

- $P(w_1 = 1|y = 1)$ = $\frac{2}{5}$ = 0.4
- $P(w_2 = 1|y = 1)$ = $\frac{1}{5}$ = 0.2
- $P(w_3 = 1|y = 1)$ = $\frac{3}{5}$ = 0.6

P(y=1) = 0.5

Spam Classification

Given, test email 0,0,1, classify using naive bayes

Spam Classification

Given, test email 0,0,1, classify using naive bayes

$$P(y = 1|w_1 = 0, w_2 = 0, w_3 = 1)$$

$$= \frac{P(w_1 = 0|y = 1)P(w_2 = 0|y = 1)P(w_3 = 1|y = 1)P(y = 1)}{P(w_1 = 0, w_2 = 0, w_3 = 1)}$$

$$= \frac{0.6 \times 0.8 \times 0.6 \times 0.5}{Z}$$

Spam Classification

Given, test email 0,0,1, classify using naive bayes

$$P(y = 1|w_1 = 0, w_2 = 0, w_3 = 1)$$

$$= \frac{P(w_1 = 0|y = 1)P(w_2 = 0|y = 1)P(w_3 = 1|y = 1)P(y = 1)}{P(w_1 = 0, w_2 = 0, w_3 = 1)}$$

$$= \frac{0.6 \times 0.8 \times 0.6 \times 0.5}{Z}$$

Similarly, we can calculate
$P(y = 0|w_1 = 0, w_2 = 0, w_3 = 1) = \frac{0.6 * 0.4 * 0.6 * 0.5}{Z}$

Spam Classification

Given, test email 0,0,1, classify using naive bayes

$$P(y = 1|w_1 = 0, w_2 = 0, w_3 = 1)$$

$$= \frac{P(w_1 = 0|y = 1)P(w_2 = 0|y = 1)P(w_3 = 1|y = 1)P(y = 1)}{P(w_1 = 0, w_2 = 0, w_3 = 1)}$$

$$= \frac{0.6 \times 0.8 \times 0.6 \times 0.5}{Z}$$

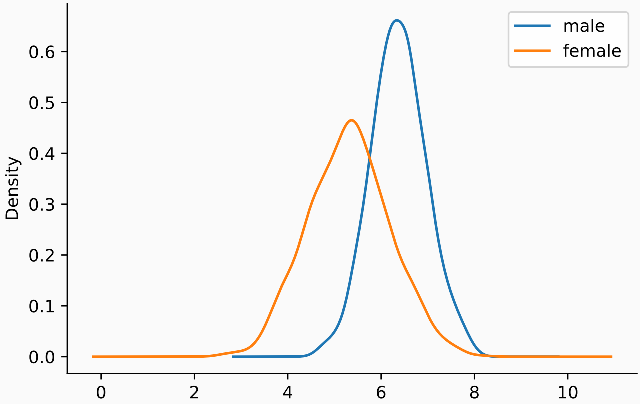
Similarly, we can calculate
$P(y = 0|w_1 = 0, w_2 = 0, w_3 = 1) = \frac{0.6*0.4*0.6*0.5}{Z}$
$\frac{P(y=1|w_1=0,w_2=0,w_3=1)}{P(y=0|w_1=0,w_2=0,w_3=1)} = 2 > 1$. Thus, classified as a spam example.

Naive Bayes for email/sentiment analysis

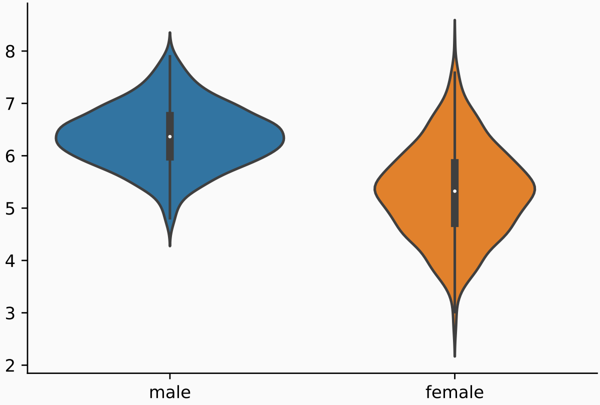
- “This product is pathetic”. We would assume the sentiment of such a sentence to be negative. Why? Presenece of “pathetic”
- Naive bayes would store the probabilities of words belonging to positive or negative sentiment.
- Good is positive, Bad is negative
- What about: This product is not bad. Naive Bayes is very naive and does not account for sequential aspect of data.

Gaussian Naive Bayes

Let us generate some normally distributed height data assuming Height (male) $\sim \mathcal{N}(\mu_1 = 6.1, \sigma_1^2 = 0.6)$ and Height (female) $\sim \mathcal{N}(\mu_2 = 5.3, \sigma_2^2 = 0.9)$

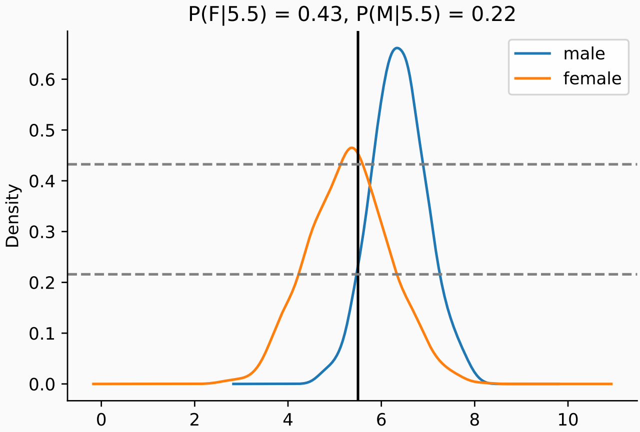


Let us generate some normally distributed height data assuming Height (male) $\sim \mathcal{N}(\mu_1 = 6.1, \sigma_1^2 = 0.6)$ and Height (female) $\sim \mathcal{N}(\mu_2 = 5.3, \sigma_2^2 = 0.9)$



Gaussian Naive Bayes

Would you expect a person to height 5.5 as a female or male? And why?



Gaussian Naive Bayes

We have classes $C_1, C_2, C_3, \ldots, C_k$
There is a continuous attribute x
For Class k

- $\mu_k = Mean(x|y(x) = C_k)$
- $\sigma_k^2 = Variance(x|y(x) = C_k)$

Guassian Naive Bayes

Now for x = some observation 'v'

$$P(x = v|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp^{\frac{-(v-\mu_k)^2}{2\sigma_k^2}}$$

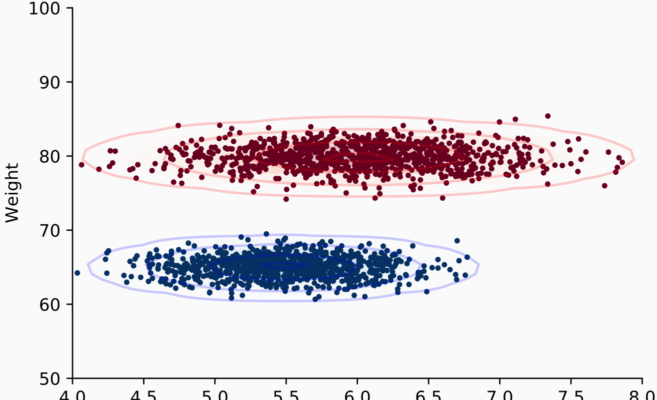
Gaussian Naive Bayes (2d example)

Would you expect a person to height 5.5 and weight 80 as a female or male? And why?

Gaussian Naive Bayes (2d example)

Would you expect a person to height 5.5 and weight 80 as a female or male? And why?

Note: no cross covariance! Remember all features are independent.



Wikipedia Example

Height	Weight	Footsize	Gender
6	180	12	M
5.92	190	11	M
5.58	170	12	M
5.92	165	10	M
5	100	6	F
5.5	100	6	F
5.42	130	7	F
5.75	150	7	F

Example

	Male	Female
Mean (height)	5.855	5.41
Variance (height)	3.5×10^{-2}	9.7×10^{-2}
Mean (weight)	176.25	132.5
Variance (weight)	1.22×10^{2}	5.5×10^{2}
Mean (Foot)	11.25	7.5
Variance (Foot)	9.7×10^{-1}	1.67

Classify the Person

- Given height = 6ft, weight = 130 lbs, feet = 8 units, classify if it's male or female.

Classify the Person

- Given height = 6ft, weight = 130 lbs, feet = 8 units, classify if it's male or female.
- $P(F|6ft, 130lbs, 8units) = \dfrac{P(6ft|F)P(130lbs|F)P(8units|F)P(F)}{P(130lbs, 8units, 6ft)}$

Classify the Person

- Given height = 6ft, weight = 130 lbs, feet = 8 units, classify if it's male or female.
- $P(F|6ft, 130lbs, 8units) = \dfrac{P(6ft|F)P(130lbs|F)P(8units|F)P(F)}{P(130lbs, 8units, 6ft)}$
- $P(130lbs|F) = \frac{1}{\sqrt{2\pi \times 550}} \times \exp \frac{-(132.5-130)^2}{2\times 550} = .0167$

Classify the Person

- Given height = 6ft, weight = 130 lbs, feet = 8 units, classify if it's male or female.
- $P(F|6ft, 130lbs, 8units) = \dfrac{P(6ft|F)P(130lbs|F)P(8units|F)P(F)}{P(130lbs, 8units, 6ft)}$
- $P(130lbs|F) = \frac{1}{\sqrt{2\pi \times 550}} \times \exp \frac{-(132.5-130)^2}{2\times 550} = .0167$
- Finally, we get probability of female given data is greater than the probability of class being male given data.