Naive Bayes

Nipun Batra

IIT Gandhinagar

August 30, 2025

Table of Contents

Introduction to Bayesian Classification

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

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

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

$$\mathbb{P}(S, G, R) = \mathbb{P}(G|S, R)\mathbb{P}(S|R)\mathbb{P}(R)$$

Bayesian Networks

Example

Known Random variables

- $\mathbb{P}(T)$
- $\mathbb{P}(E)$
- $\mathbb{P}(A|T, E)$
- $\mathbb{P}(R|E)$

Question

Given, the above, calculate

$$\mathbb{P}(A|T)$$

Solution

$$
\begin{aligned}
\mathbb{P}(A|T) &= \frac{\mathbb{P}(A, T)}{\mathbb{P}(A)} \\
&= \frac{\mathbb{P}(A, T, E) + \mathbb{P}(A, T, \bar{E})}{\mathbb{P}(A, T, E) + \mathbb{P}(A, T, \bar{E}) + \mathbb{P}(A, \bar{T}, E) + \mathbb{P}(A, \bar{T}, \bar{E})}
\end{aligned}
$$

Medical Diagnosis

You tested positive for a disease.
Well, the test is only 99% accurate.

- $\mathbb{P}(\mathit{Test} = +ve | \mathit{Disease} = \mathit{True}) = 0.99$
- $\mathbb{P}(\mathit{Test} = -ve | \mathit{Disease} = \mathit{False}) = 0.99$

Also, the disease is a rare one. Only one in 10,000 has it.

Problem

- $\mathbb{P}(T|D) = 0.99$
- $\mathbb{P}(\bar{T}|\bar{D}) = 0.99$
- $\mathbb{P}(T|\bar{D}) = 0.01$
- $\mathbb{P}(\bar{T}|D) = 0.01$
- $\mathbb{P}(D) = 10^{-4}$
- $\mathbb{P}(\bar{D}) = 1 - 10^{-4}$

Given the above, calculate $\mathbb{P}(D|T)$. Given the result is positive, what is the probability that someone has the disease

Problem

$$
\begin{aligned}
\mathbb{P}(D|T) &= \frac{\mathbb{P}(T|D)\mathbb{P}(D)}{\mathbb{P}(T)} \\
&= \frac{\mathbb{P}(T|D)\mathbb{P}(D)}{\mathbb{P}(T|D)\mathbb{P}(D) + \mathbb{P}(T|\bar{D})\mathbb{P}(\bar{D})}
\end{aligned} \tag{1}
$$

SPAM EMAIL CLASSIFICATION

From the emails construct a vector X. The vector has ones if the word is present, and zeros is the word is absent

Naive Bayes

- Classification model
- Scalable
- Generative

$$\mathbb{P}(x_1, x_2, x_3, \ldots, x_n|y) = \mathbb{P}(x_1|y)\mathbb{P}(x_2|y)\ldots\mathbb{P}(x_N|y)$$

Quick Question

Why is Naive Bayes model called Naive?

Frame Title

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

What do we need to predict?

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

Spam Mail Classification

Probability of x_i being a spam email

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

Similarly,

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

Spam Mail classification

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

Similarly,

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

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

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

$\mathbb{P}(y = 0) = 0.5$
Similarly, if y=1

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

$\mathbb{P}(y = 1) = 0.5$

Spam Classification

Given, test email 0,0,1, classify it. Using naive bayes rule, we can do the following,

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

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'

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

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 = 13.5 lbs, feet = 8 inches, classify if it's male or female.

Classify the Person

Given height $= 6$ft, weight $= 13.5$ lbs, feet $= 8$ inches, classify if it's male or female.
It is female!

Practice and Review

Pop Quiz: Naive Bayes Concepts

1. What does the "naive" assumption in Naive Bayes refer to?

Pop Quiz: Naive Bayes Concepts

1. What does the "naive" assumption in Naive Bayes refer to?
2. Why is Naive Bayes particularly effective for text classification?

Pop Quiz: Naive Bayes Concepts

1. What does the ”naive” assumption in Naive Bayes refer to?
2. Why is Naive Bayes particularly effective for text classification?
3. What happens when a feature value appears in test data but not in training data?

Pop Quiz: Naive Bayes Concepts

1. What does the "naive" assumption in Naive Bayes refer to?
2. Why is Naive Bayes particularly effective for text classification?
3. What happens when a feature value appears in test data but not in training data?
4. Compare Naive Bayes with logistic regression - when would you choose each?

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems
- **Good with Small Data**: Works well with limited training examples

Key Takeaways

- **Probabilistic Foundation**: Based on Bayes' theorem and conditional independence
- **Naive Assumption**: Features are conditionally independent given the class
- **Efficient Training**: Simple parameter estimation from training data
- **Handles Multiple Classes**: Naturally extends to multi-class problems
- **Good with Small Data**: Works well with limited training examples
- **Interpretable**: Probabilistic outputs provide confidence measures