Bayesian classifier & Naive Bayes

Lecture outline

1. Bayes theorem
2. Bayesian classifier
3. Naive Bayes
4. Demo

Bayes Theorem

Bayes Theorem - Definitions

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

$P(A \mid B)$	Posterior probability
$P(B \mid A)$	Likelihood
$P(A)$	Prior probability
$P(B)$	Evidence

What is the probability that there is fire given there is smoke?

$$P(\,fire \mid smoke) \;=\; \frac{P(smoke \mid fire)\,P(fire)}{P(smoke)}$$

Scenario: Medical diagnostic test

Medical diagnostic test - scenario

- Consider a human population that may or may not have cancer
  - Cancer = True or False
- Consider a medical test supposed to detect cancer, that returns positive or negative
  - Test = Positive or Negative

Problem: You test a random person from the population and he/she gets a positive test result. What is the probability that the person actually has cancer?

Question

Given a person tests positive using the diagnostic test, what is the probability the person actually has cancer?

Sensitivity

- Medical diagnostic tests are not perfect
  - You might have heard of the PCR test or the Rapid Antigen Test (RAT) for Covid-19
- The capability of a test to accurately detect the condition is referred to as the **sensitivity** of the test or the **true positive rate**

$$\frac{True\,positives}{Positive\,test\,results} \cdot 100\,\%$$

- The sensitivity is usually determined by a statistical analysis of a large number of data points
- In our medical diagnostic example, after an appropriate study, the test was found to have a sensitivity of 0.85
  - If 100 people are tested positive, only 85 will actually have cancer

Question

Given a person tests positive using the diagnostic test, what is the probability the person actually has cancer?

(a) Is it 100%

(b) Is it 85%?

(c) Is it something else?

Question - Bayes formulation

Given a person tests positive using the diagnostic test, what is the probability the person actually has cancer?

$$P(Cancer = true \mid Test = positive) = \frac{P(Test = positive \mid Cancer = true)\, P(Cancer = true)}{P(Test = positive)}$$

?

Base rate or prior probability

What's the probability of any person in a population having cancer?

- *P(Cancer = true) = ?*
- Determined using a statistical analysis
- In the event of lack of data, make a sensible assumption.

- *P(Cancer = true) = 0.0002*

Question

Given a person tests positive using the diagnostic test, what is the probability the person actually has cancer?

$$P(Cancer = true \mid Test = positive) = \frac{P(Test = positive \mid Cancer = true)\, P(Cancer = true)}{P(Test = positive)}$$

?

Evidence term

P(Test = positive) = ?

- Typically, the evidence term is difficult to reliably estimate statistically
- But we have an alternative way of calculating it, using some operations from probability theory

$P(B) = P(B|A) * P(A) + P(B|\neg A) * P(\neg A)$

P(Test=Positive) =

P(Test=Positive|Cancer=True) * P(Cancer=True) + P(Test=Positive|Cancer=False) * P(Cancer=False)

Evidence term (2)

P(Cancer=False) = 1 – P(Cancer=True)

= 1 – 0.0002

= 0.9998

We can plug in our known values as follows:

P(Test=Positive) = 0.85 * 0.0002 + P(Test=Positive|Cancer=False) * 0.9998

False positive rate?

Specificity

P(Test=Negative|Cancer=False)

How good is the test, at correctly identifying people without cancer

Also known as the **true negative rate**

$$\frac{True\ negatives}{Negative\ test\ results} \cdot 100\,\%$$

In our study, the specificity of the test was found to be 95%

P(Test=Negative | Cancer=False) = 0.95

P(Test=Positive|Cancer=False) = 1 – P(Test=Negative | Cancer=False)

= 1 – 0.95

= 0.05

Back tracking (1): base rate probability

P(Test=Positive) = 0.85 * 0.0002 + P(Test=Positive|Cancer=False) * 0.9998

P(Test=Positive|Cancer=False) = = 1 – P(Test=Negative | Cancer=False)

P(Test=Positive) = 0.85 * 0.0002 + 0.05 * 0.9998

P(Test=Positive) = 0.05016

The probability of the test returning a positive result, regardless of whether the person has cancer or not is about 5%

Back tracking (2): Posterior probability

Given a person tests positive using the diagnostic test, what is the probability the person actually has cancer?

$$P(Cancer = true \,|\, Test = positive) = \frac{P(Test = positive \,|\, Cancer = true)\, P(Cancer = true)}{P(Test = positive)}$$

P(Cancer=True | Test=Positive) = 0.85 * 0.0002 / 0.05016

P(Cancer=True | Test=Positive) = 0.00017 / 0.05016

P(Cancer=True | Test=Positive) = 0.003389154704944

If the patient is informed they have cancer with this test, then there is only 0.33% chance that they have cancer.

Connecting Bayes Theorem with Binary Classification

Confusion Matrix (two-class scenario)

	Positive class	**Negative class**
Positive prediction	True Positive (TP)	False Positive (FP)
Negative prediction	False Negative (FN)	True Negative (TN)

True Positive Rate (TPR) = TP / (TP + FN) = Sensitivity

False Positive Rate (FPR) = FP / (FP + TN)

True Negative Rate (TNR) = TN / (TN + FP) = Specificity

False Negative Rate (FNR) = FN / (FN + TP)

Precision or Positive Predictive Value (PPV)

$$PPV = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- P(A|B) = PPV

$\boxed{\text{Precision takes the same value as Posterior!}}$

Can you figure out the connection from the other terms given in Bayes Theorem to the terms given in the confusion matrix?

Bayes Optimal Classifier

Bayes theorem for classification

Classification: assign a label to a given input

This can be framed as calculating the conditional probability of a class label given a data sample

$$P(class \mid data) = \frac{P(data \mid class) \cdot P(class)}{P(data)}$$

The likelihood term tends to be difficult to estimate. It typically requires a very large number of examples to effectively determine the probability distribution p(data|class)

Maximum *A Posteriori* (MAP) Estimation

What is the class that maximizes P(class | data)?

$$P(class \mid data) = \frac{\boxed{P(data \mid class) \cdot P(class)}}{P(data)}$$

- Since we are only interested in in the MAP estimate, we can simplify the optimization
- We can ignore the denominator P(data), because it is a constant over all the classes
- The class that maximizes the above is the same as the class that maximizes:

$$\boxed{P(data \mid class) * P(class)}$$

* It's also common to maximize the log of the above expression, because log is a monotonic function

Bayes classifier: Toy example (1)

Given:

- Features: $X = (X_1, X_2, \ldots, X_n)$
- Labels: $Y = (Y_1, Y_2, \ldots, Y_n)$

Find the value of Y for which the following posterior probability is maximum

$P(Y=y \mid X = (x_1, x_2, \ldots, x_m))$

In other words, what is the class label Y,

for a new data point $X = (x_1, x_2, \ldots, x_m)$?

Bayes classifier: Toy example (2)

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

- $X_i \in \{0,1,2\}$
- $Y_i \in \{0,1\}$

Estimate Y, given X = (0,2)

That is, find the value y that maximizes the posterior:

$P(Y=y \mid X = (0,2))$

Since y can be either 0 or 1, we calculate the posterior corresponding to both cases

$P(Y=0 \mid X = (0,2))$ and $P(Y=1 \mid X = (0,2))$

Bayes classifier: Toy example (3)

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Estimate Y, given X = (0,2)

P(Y=y | X = (0,2)) = ?

(a) Posterior for Y=0: P(X=(0,2) | Y=0) P (Y=0)

P(Y=0) = 6/10, P(X=(0,2) | Y=0) = 0

=> (a) = 0

(b) Posterior for Y=1: P(X=(0,2) | Y=1) P (Y=1)

P(Y=1) = 4/10

P(X=(0,2) | Y=1) = 1/4

=> (b) = ¼ * 4/10 = 0.1

=> Y=1 maximizes the posterior for X=(0,2)

We have a problem!

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Estimate Y, given X = (0,2)

$P(Y=y \mid X = (0,2)) = ?$

(a) $P(Y=0 \mid X = (0,2)) \propto P(X=(0,2) \mid Y=0)\ P(Y=0)$

$P(Y=0) = 6/10$, $P(X=(0,2) \mid Y=0) = 0$

=> (a) = 0

(b) $P(Y=1 \mid X = (0,2)) \propto P(X=(0,2) \mid Y=1)\ P(Y=1)$

$P(Y=1) = 4/10$

$P(X=(0,2) \mid Y=1) = 1/4$

=> (b) = ¼ * 4/10 = 0.1

=> Y=1 maximizes the posterior for X=(0,2)

The likelihood term becomes zero for all the combinations that are NOT directly observed before. This is common for a larger number of features X

It's difficult to compare multiple values of zero!

Solution: Naive Bayes Classifier

Solution: Naive Bayes

Consider X_1 and X_2 are independent (a naive assumption?)

=> $P(X=(0,2) \mid Y=1) = P(X_1=0 \mid Y=1) * P(X_2=2 \mid Y=1)$

$P(X=(0,2) \mid Y=0) = P(X_1=0 \mid Y=0) * P(X_2=2 \mid Y=0)$

Solution: Naive Bayes

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Consider X_1 and X_2 are independent

=> $P(X=(0,2) \mid Y=1) = P(X_1=0 \mid Y=1) * P(X_2=2 \mid Y=1)$ = ¾ *

$P(X=(0,2) \mid Y=0) = P(X_1=0 \mid Y=0) * P(X_2=2 \mid Y=0)$

Solution: Naive Bayes

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Consider X_1 and X_2 are independent

=> $P(X=(0,2) \mid Y=1) = P(X_1=0 \mid Y=1) * P(X_2=2 \mid Y=1)$ $\boxed{= ¾ * 2/4}$

$P(X=(0,2) \mid Y=0) = P(X_1=0 \mid Y=0) * P(X_2=2 \mid Y=0)$

Solution: Naive Bayes

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Consider X_1 and X_2 are independent

=> $P(X=(0,2) \mid Y=1) = P(X_1=0 \mid Y=1) * P(X_2=2 \mid Y=1) = 3/4 * 2/4$

$P(X=(0,2) \mid Y=0) = P(X_1=0 \mid Y=0) * P(X_2=2 \mid Y=0)$ $\boxed{= 1/6 * 1/6}$

Solution: Naive Bayes

X_1	X_2	Y
0	0	0
0	1	1
1	2	1
0	0	1
2	2	0
1	1	0
0	2	1
2	0	0
2	1	0
1	0	0

Consider X_1 and X_2 are independent

=> $P(X=(0,2) \mid Y=1) = P(X_1=0 \mid Y=1) * P(X_2=2 \mid Y=1) = 0.375$

$P(X=(0,2) \mid Y=0) = P(X_1=0 \mid Y=0) * P(X_2=2 \mid Y=0) = 0.0278$

Using Maximum *A Posteriori* (MAP) estimation:
For X=(0,2) => Y = 1

Naive Bayes and MAP recap

$$P(Y \mid X) = \frac{P(X \mid Y) \cdot P(Y)}{P(X)}$$

To find the value of Y that maximizes the above posterior P(Y|X), we can find the Y that maximizes the numerator P(X|Y)*P(Y)

$X = [x_1, x_2, \ldots, x_n]$

$P(X|Y)*P(Y) = P(X=[x_1, x_2, \ldots, x_n]|Y)*P(Y)$

With the conditional independence of X, this becomes

$P(X=x_1|Y)*P(X=x_2|Y)*...*P(X=x_n|Y)*P(Y) = \prod_{i=1}^{n} P(X = x_i \mid Y) \cdot P(Y)$

The “Zero Frequency / Probability” problem

$P(X=x_1|Y)*P(X=x_2|Y)*...*P(X=x_n|Y)*P(Y) = \prod_{i=1}^{n} P(X = x_i \,|\, Y) \cdot P(Y)$

- What if one of the $P(X=x_i|Y)$ terms is not observed?
- Then for that value of X=xi, the conditional probability becomes zero.
- **That makes the whole product zero**. Which makes the MAP estimation process useless
- To avoid it we should use a **smoothing technique**.
  - E.g. **Laplace smoothing/correction**

Laplace Smoothing

Also known as **Laplace Correction** and **Additive Smoothing**

- A small-sample correction will be added to every probability estimate in the likelihood term
  - Therefore, no part of the term will be zero

Original way of calculating $P(x_i|Y)$

$$P(x_i \,|\, Y = y) = \frac{n^y_{x_i} + \alpha}{N + \alpha K}$$

Laplace smoothing

$n^y_{x_i}$ - Number of times feature x_i is observed with label y

α - Smoothing parameter

K - Number of features (dimensions)

N - Total number of observations with Y=y

Smoothing parameter α

- A hyper parameter that can be tuned
- Typically set to 1
- Otherwise use an *elbow plot* or *cross validation* to determine a suitable value

Where does Naive Bayes fit in?

- Classifier
  - To find labels for data points (features)
- Supervised learning based
  - Needs examples

Disadvantages/Features of Naive Bayes

- Does not accurately capture the interdependencies among features, which might cause problems if there is a significant level of interdependencies
  - This is when the “naive assumption” i.e. conditional independence doesn’t hold
- Most of the problems caused are because of the independence assumptions because it is not realistic in many real world situations
- Although the classification results are usually reliable, the posterior probability might not be

Advantages/Features of Naive Bayes

- Easy and fast to build
- Fast and efficient in deployment
- When the conditional independence assumption holds, NB performs better than most classifiers especially when there is only limited amount of data
- Performs better with categorical input variables than with numerical variables
  - For numerical variables, often a normal distribution is assumed. This might not hold strongly.
- Suitable for multi-class classification tasks
- Suitable for real-time classification tasks

Application domains of Naive Bayes

- Some recommendation systems use collaborative filtering and Naive Bayes
- NLP tasks such as
    - Text classification
    - Spam filtering
    - Sentiment analysis

Notable cases of Naive Bayes

Differences arise because of the assumptions on how the likelihood P(X|Y) is distributed

Attribute/Feature Type	**NB Adaptation (python sklearn module)**
Multivariate bernoulli distributed	BernoulliNB
Real (continuous valued)	GaussianNB
Categorical	CategoricalNB
…	…

Case 1: Gaussian NB

When the attributes are real valued (continuous) variables

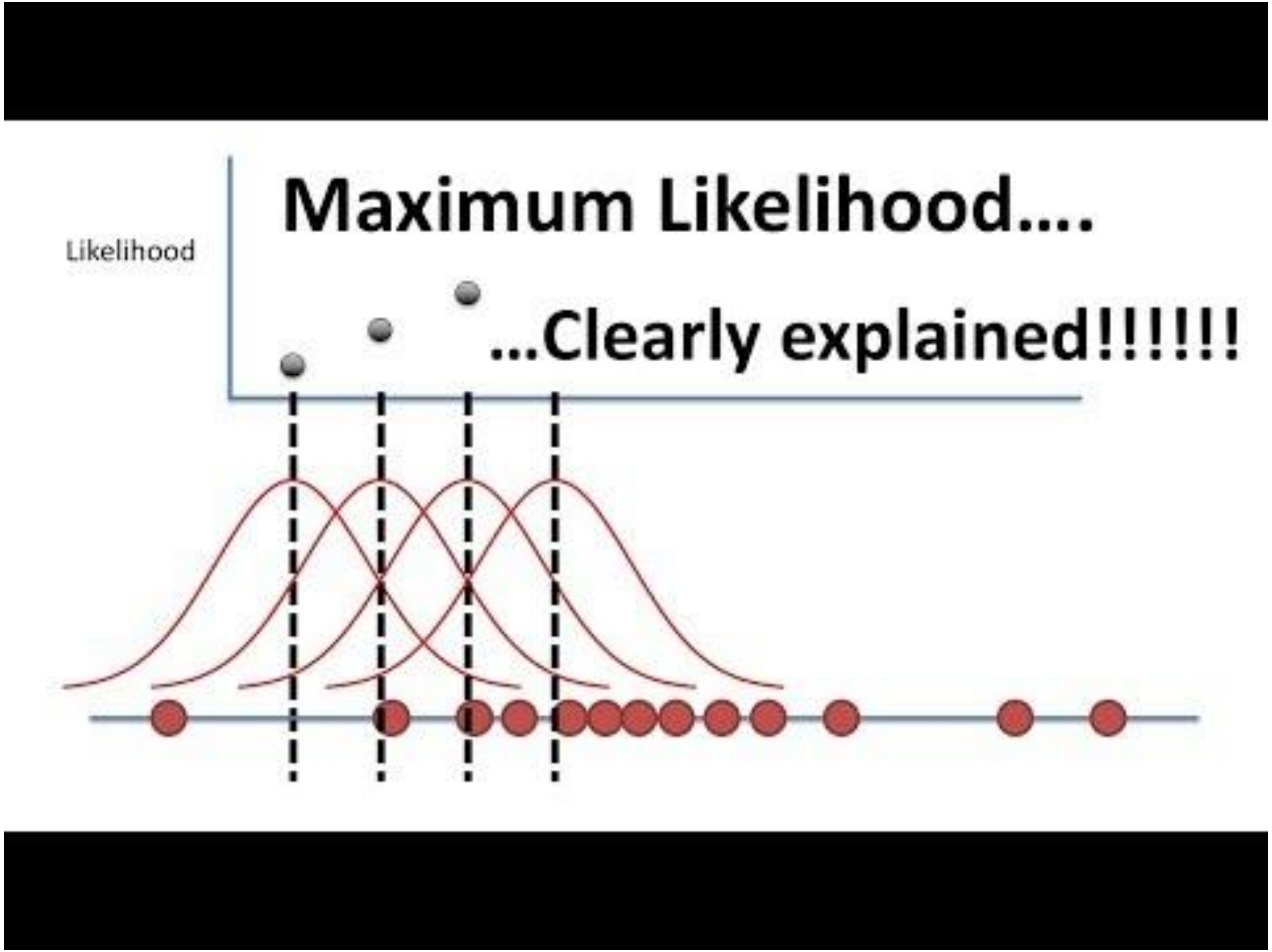
E.g. Age, height, …

.. and it’s possible to assume they are normally distributed according to:

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

The parameters μ (mean) and σ (standard deviation) are estimated from the data using maximum likelihood estimation (MLE)

Primer on Maximum Likelihood Estimation (MLE)



Case 2: Categorical NB

When each feature has its own categorical distribution

The probability of category in feature given class is estimated as:

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

$N_{tic} = |\{j \in J \mid x_{ij} = t, y_j = c\}|$ Is the number of times category t appears in the samples x_i, which belongs to class c. Alpha is a smoothing parameter and n_i is the number of available categories in feature i.

Naive Bayes Demo


simplilearn
NAIVE BAYES
CLASSIFIER