KNN and Naïve Bayes Algorithm

Supervised Learning Classification

Agenda

- Proximity Measures
  - Similarity measures
  - Dissimilarity measure
- Distance Measures
  - Euclidean Distance
  - Manhattan Distance
  - Minkowski Distance
  - Chebyshev’s Distance
- KNN Algorithm

Agenda

- Basic Concepts
  - Probability
  - Bayes theorem
- Naïve Bayes Algorithm
  - Business Problem
  - Laplace smoothing method

- In the previous session, we covered the decision tree algorithm which is the rule based classifier

- In this session we shall study the K nearest Neighbours classifier and Naïve Bayes Classifier

Data matrix

- The data represented in form of a matrix is called the data matrix

- For data with m features are n observations the data matrix is given as

$$\begin{pmatrix} x_{11} & & x_{21} \\ & \cdots & x_{m1} \\ x_{12} & & x_{22} \\ & \cdots & x_{m2} \\ & & \ddots \end{pmatrix}$$

x_{1n} x_{2n}

Proximity Measures

Proximity measures

- The proximity measures find the distance between two instances
- Proximity measures include
  - Similarity measures
  - Dissimilarity measure
- Depending upon the data types, we choose the proximity measure

Similarity measures

A similarity measure for two objects, will return the value 0 if the objects are unlike, and the value 1 if the objects are alike.

A similarity matrix

$$\begin{array}{c|ccc} & x_1 & x_2 & x_n \\ \hline x_1 & 1 & & \\ x_2 & d_{12} & 1 & \\ & & \ddots & \\ x_n & d_{1n} & d_{2n} & \cdots \end{array}$$

Dissimilarity measures

- Dissimilarity measure work exactly opposite of a similarity measure

- A dissimilarity measure for two objects, will return the value 1 if the objects are unlike, and the value 0 if the objects are alike

A dissimilarity matrix

$$
\begin{array}{cccc}
 & x_1 & x_2 & x_n \\
x_1 & 0 & & \\
x_2 & d_{12} & 0 & \\
 & & \ddots & \\
x_n & d_{1n} & d_{2n} &
\end{array}
$$

Distance measures

Depending on the type of data we have different distance measures

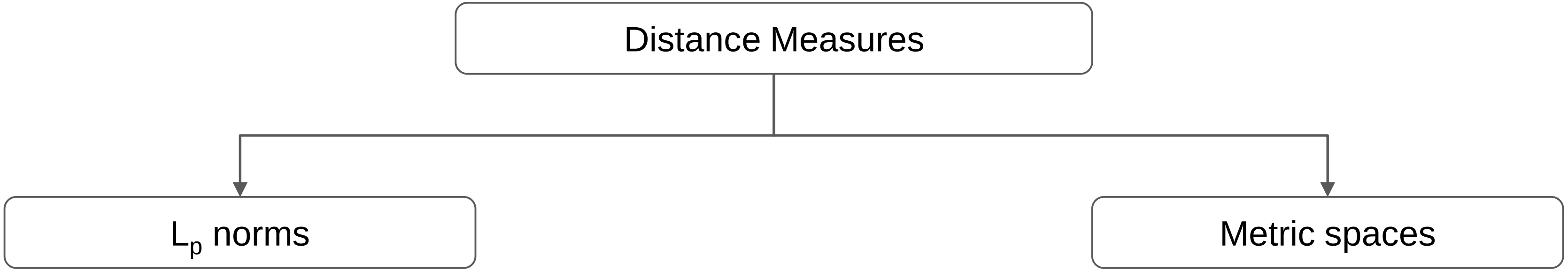
Numeric data

- Euclidean distance
- Manhattan distance
- Minkowski distance
- Chebyshev's distance

String data

- Cosine distance
- Edit distance
- Longest Common Sequence
- Hamming distance

Distance measures



Distance Measures

L_p norms

Metric spaces

The distance measures based on norm

The distance measures which satisfy the following properties:

$$d(a,b) \geq 0 \text{ (non-negativity)}$$

$$d(a,b) \equiv 0 \iff a \equiv b \text{ (positive definiteness)}$$

$$d(a,b) \equiv d(b,a) \text{ (symmetry)}$$

$$d(a,b) \leq d(a,c) + d(c,b) \text{ (triangle inequality)}$$

Euclidean distance - numeric data

- Euclidean distance is obtained for numeric data

- It is the L2 norm

- For two instances X and Y the Euclidean distance is given by

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

where x_i and y_i are the values taken by X and Y respectively

- More the distance between the two instance more the dissimilarity measure

Euclidean distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the Euclidean distance.

$$\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} = \sqrt{(5-1)^2 + (6-3)^2}$$

$$= \sqrt{16+9}$$

$$= \sqrt{25}$$

$$= 5$$

Squared euclidean distance - numeric data

- Similar to Euclidean distance is obtained for numeric data
- Just the squared value of Euclidean distance
- For two instances X and Y the squared Euclidean distance is given by

$$\sum_{i=1}^{n} (x_i - y_i)^2$$

where x_i and y_i are the values taken by X and Y respectively

- It is computationally more efficient than the Euclidean distance

Manhattan distance - numeric data

- It is the L1 norm

- For two instances X and Y, it is given by

$$\sum_{i=1}^{n} |x_i - y_i|$$

where x_i and y_i are the values taken by X and Y respectively

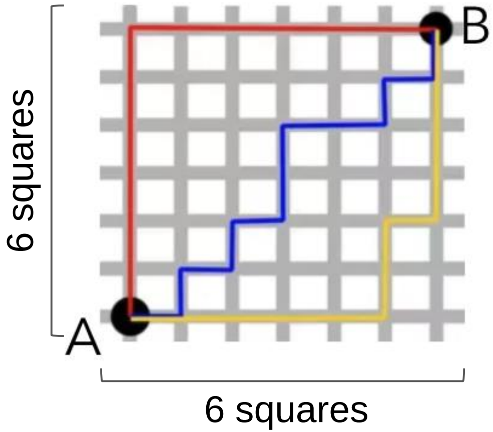
- Also known as the Taxicab distance or Snake distance or the City block distance

Manhattan distance - numeric data

- Consider points A and B

- The Manhattan distance between the points is given by the edge of the squares it crosses

- Here the Manhattan distance is 12

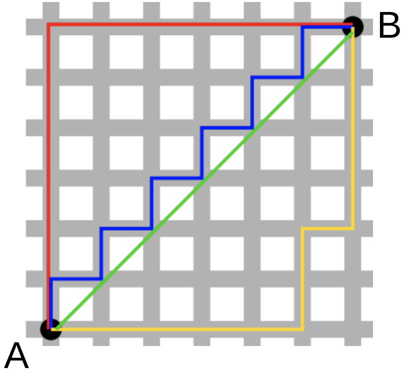


Manhattan distance - numeric data

- For the same points the Euclidean distance is the given by the shortest distance between them

- Given by the green line

- The euclidean distance is 6√2



Minkowski distance - numeric data

- It is the generalized form of the Manhattan and the Euclidean distance
- It is the Lp norm
- For two instances X and Y, it is given by

$$\sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

where x_i and y_i are the values taken by X and Y respectively and p>0

Minkowski distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the minkowski distance (take p = 4)

$$\sqrt[p]{\sum_{i=1}^{n}(x_i - y_i)^p} = \sqrt[4]{\sum_{i=1}^{n}(x_i - y_i)^4}$$

$$= \sqrt[4]{(5-1)^4 + (6-3)^4}$$

$$= \sqrt[4]{256 + 81} = \sqrt[4]{337}$$

$$= 4.28$$

Chebyshev distance - numeric data

- It is the L_∞ norm

- For two instances X and Y

$$\mathbf{max}_{i=1}^{n} \left| x_i - y_i \right|$$

where x_i and y_i are the values taken by X and Y respectively

Chebyshev distance - numeric data

Example: Consider two points (5,6) and (1, 3). Obtain the Chebyshev distance.

$$\max_{i=1}^{n} |x_i - y_i| = \max\{|5-1|, |6-3|\}$$

$$= \max\{4, 3\}$$

$$= 4$$

Summary

- K-nearest neighbours is a distance based algorithm
- Different distance metrics are used to find the points that are closer to the testing point
- Minkowski distance is the generalized form of Manhattan and Euclidean distance
- The Chebyshev distance is obtained when the $p \rightarrow \infty$ in the Minkowski distance
- In python, the KNeighborsClassifier() considers the Euclidean distance by default

K - NN algorithm

K - NN algorithm

Specifies number of nearest neighbours

NN stands for Nearest Neighbours

KNN algorithm

- The K - Nearest Neighbour (KNN) algorithm classifies the data based on the similarity measure

- K specifies the number of nearest neighbours to be considered

- Does not require the data to be trained

KNN algorithm

KNN is considered to be:

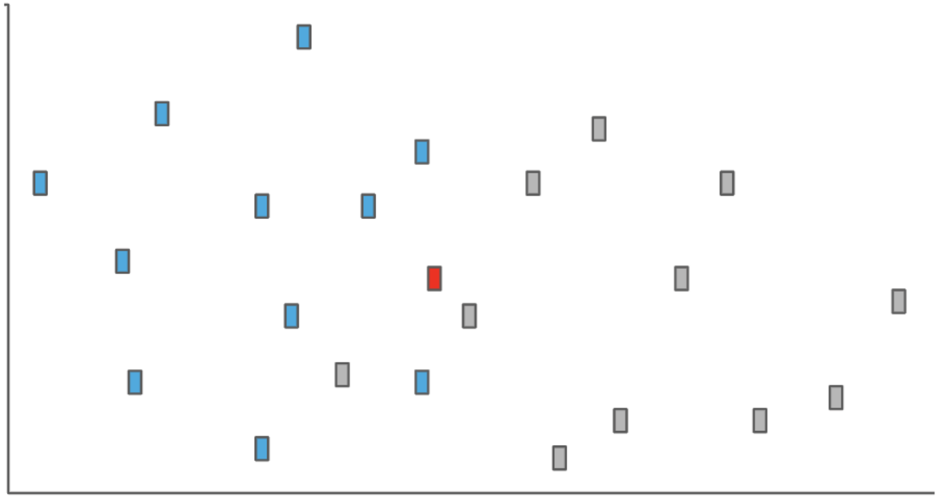
- Instance based learning algorithm: uses training instances to make predictions

- Lazy learning algorithm: does not required a model to be trained

- Non-Parametric algorithm: no assumptions are made about the functional form of the the problem being solved

KNN algorithm

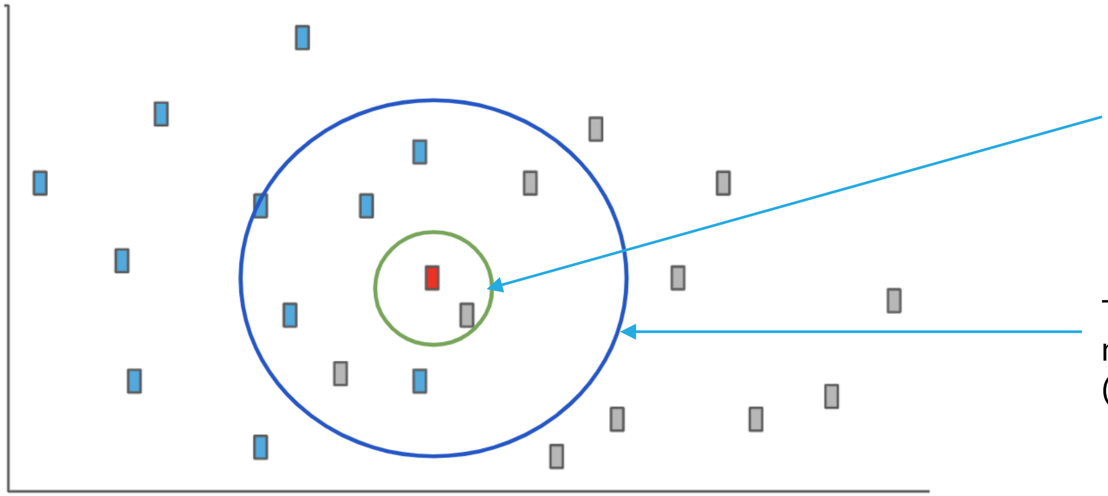
Consider two classes as shown in the figure - Class A and Class B



Which class does the red point belong to?

KNN algorithm

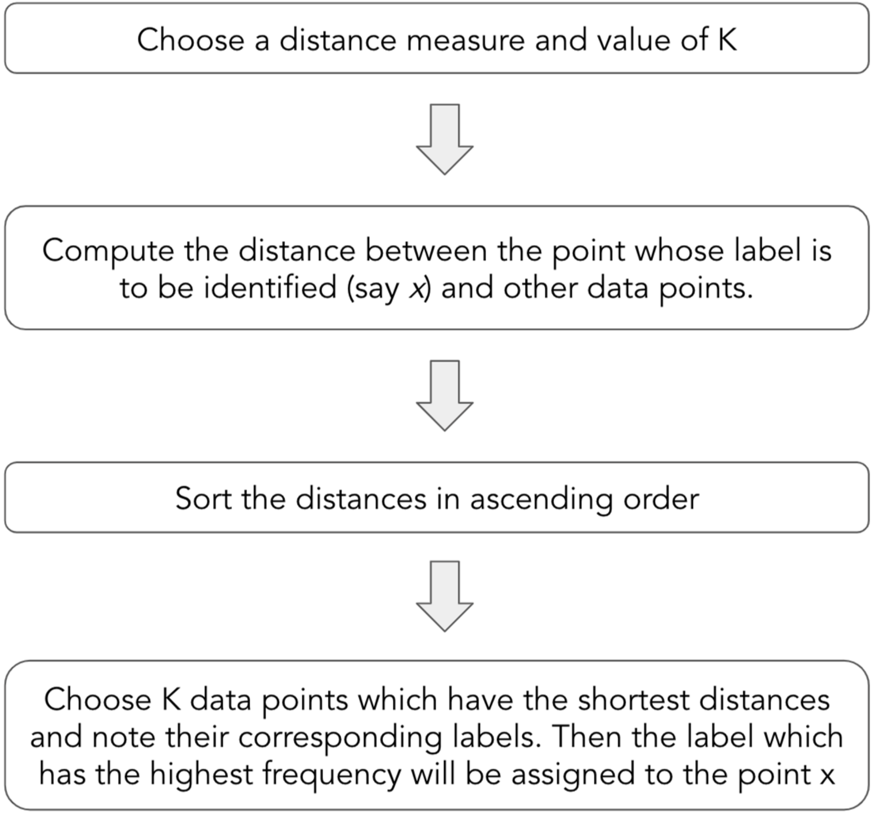
An easy approach is to label it as the class which has the highest frequency around it



The highest frequency of the nearest class is Class B (K = 1)

The highest frequency of the nearest classes is Class A (K = 7)

KNN algorithm - Procedure



Choose a distance measure and value of K

Compute the distance between the point whose label is to be identified (say *x*) and other data points.

Sort the distances in ascending order

Choose K data points which have the shortest distances and note their corresponding labels. Then the label which has the highest frequency will be assigned to the point x

Data Scaling

Consider a data with 5 features, of which 4 features have small range (say between 0 to 25) and the fifth feature ranges from -100 to 2000. The classification would be majorly based on the feature with high range. Since it's contribution to the distance measure would be high with very little or no effect of the other variables thus we scale the data.

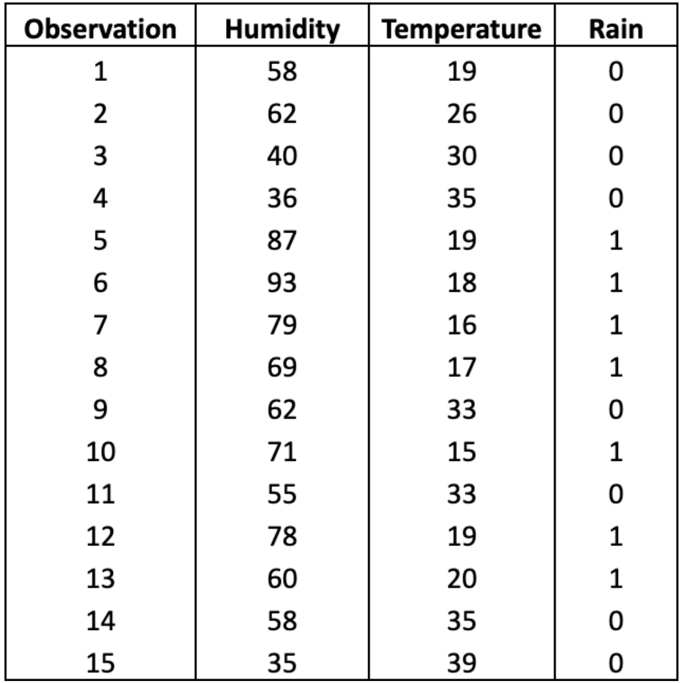
Note: In most of the scenarios, min-max normalization works better. But, we can also perform standardization. Whether to normalize or standardize the data is completely experimental.

Example

- Humidity: (Independent variable) the percentage of humidity in the atmosphere

- Temperature: (Independent variable) the temperature average temperature during precipitation

- Rain: (Target variable) indicates whether it rained or not; takes value 1 if rained and value 0 otherwise



Observation	Humidity	Temperature	Rain
1	58	19	0
2	62	26	0
3	40	30	0
4	36	35	0
5	87	19	1
6	93	18	1
7	79	16	1
8	69	17	1
9	62	33	0
10	71	15	1
11	55	33	0
12	78	19	1
13	60	20	1
14	58	35	0
15	35	39	0

Would it rain if Humidity = 84 and Temperature = 37?

Example

- Let us choose the K = 5 and used the Euclidean distance

- Compute the Euclidean distance between the new data for each instance

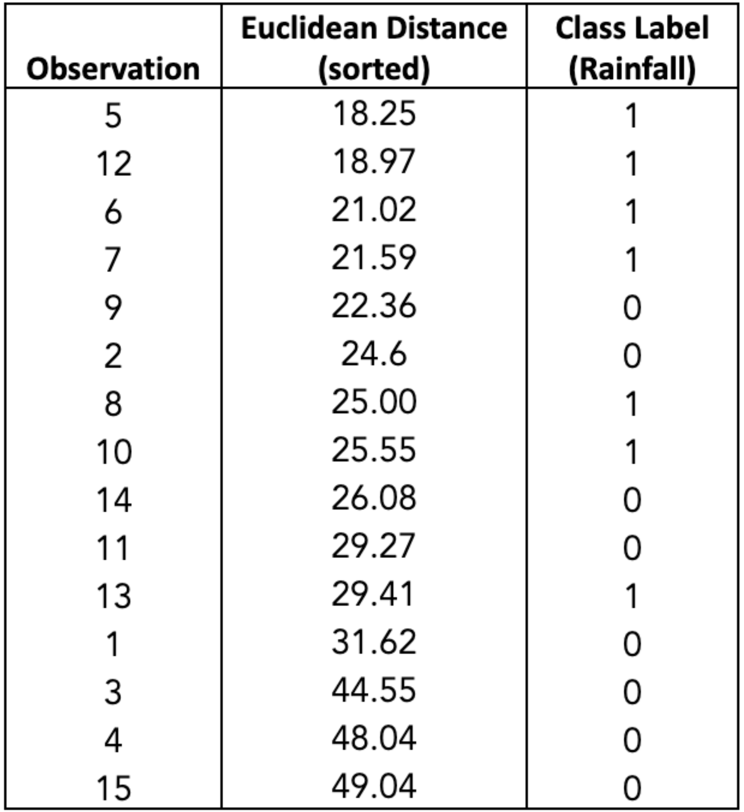
- For example, consider the first observation Humidity = 58 and Temperature = 19, the Euclidean distance is

$$[(58 - 84)^2+(19 -34)^2]^{1/2} = 31.623$$

Humidity	Temperature	Rainfall
84	34	?

Example

Computed the Euclidean distances for each instance with the new data and sort the data in ascending order with respect to the Euclidean distance

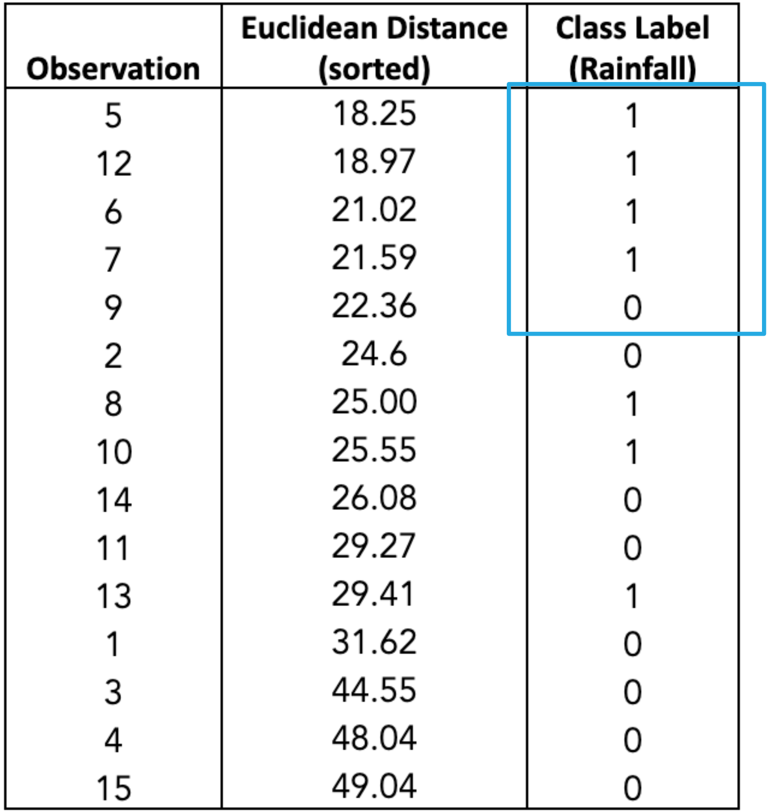


Observation	Euclidean Distance (sorted)	Class Label (Rainfall)
5	18.25	1
12	18.97	1
6	21.02	1
7	21.59	1
9	22.36	0
2	24.6	0
8	25.00	1
10	25.55	1
14	26.08	0
11	29.27	0
13	29.41	1
1	31.62	0
3	44.55	0
4	48.04	0
15	49.04	0

Example

- Since K = 5 consider the class labels of first five observations

- 1 appears 4 times and 0 appears 1 time



Observation	Euclidean Distance (sorted)	Class Label (Rainfall)
5	18.25	1
12	18.97	1
6	21.02	1
7	21.59	1
9	22.36	0
2	24.6	0
8	25.00	1
10	25.55	1
14	26.08	0
11	29.27	0
13	29.41	1
1	31.62	0
3	44.55	0
4	48.04	0
15	49.04	0

Example

- 1 appears 4 times and 0 appears 1 time

- Thus for our new instance the class label is 1 (using max voting)

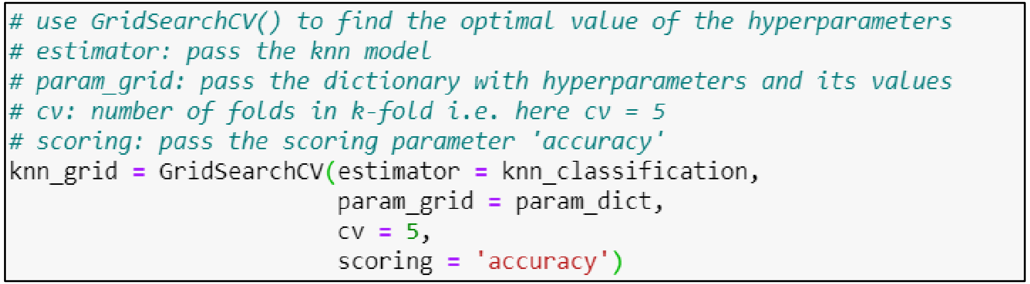
Humidity	Temperature	Rainfall
84	34	1

- Implies that for Humidity = 84 and Temperature = 34, it will rain

The value of K

- The value of K should be chosen appropriately since a large K value may reduce the variance due to the noisy data, but increase bias resulting to ignorance of smaller patterns, whereas a small K may overfit the data

- In order to avoid a tie consider the odd value of K



```
# use GridSearchCV() to find the optimal value of the hyperparameters
# estimator: pass the knn model
# param_grid: pass the dictionary with hyperparameters and its values
# cv: number of folds in k-fold i.e. here cv = 5
# scoring: pass the scoring parameter 'accuracy'
knn_grid = GridSearchCV(estimator = knn_classification,
                        param_grid = param_dict,
                        cv = 5,
                        scoring = 'accuracy')
```

did you know?

Weighted KNN

- Selecting an apt K is challenging. To overcome this, weighted KNN is used
- Weights are assigned to each instance
- Generally, the weights are the inverse of the distance
- The weights are higher for the points which are nearer to the new instance
- The weights are lower for the points which are away from the new instance

KNN

Advantages

- Easy to implement
- No training required
- New data can be added at any time
- Effective if training data is large

Disadvantages

- To chose apt value for K
- Computational expensive
- Can not tell which features gives the best result

KNN - Applications

- Image classification
- Handwriting recognition
- Predict credit rating of customers
- Replace missing values

Naïve Bayes

Business problem: label the email as spam or ham

It can be helpful if an algorithm can label received emails as important (ham) emails or junk (spam) emails for a user.

Such a model can ease the effort of the user by directly showing them important emails and filter out the junk.

Visiting Basics

Probability

- Probability is how likely an event is to occur

$$P = \frac{\text{No. of ways an event can occur}}{\text{Total possible events}}$$

- The probability of an event always lies in between 0 and 1

- 0 indicates impossibility of the event and 1 indicates a certain event

Probability

Question:

There are 40 candidates in a team with equal calibre. Out of which 25 are men and 15 are women. A person is randomly chosen to be the team leader. What is the probability that the person is a woman?

Probability

Solution:

Number of ways event can occur: 15

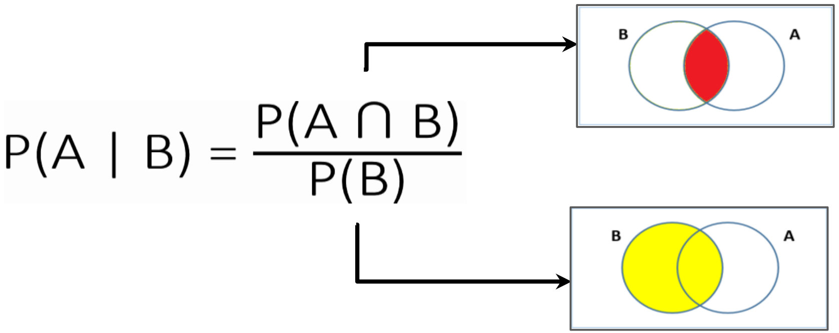
Total number of outcomes: 40

Therefore the probability: 15/40 = 0.375

Conditional probability

- The conditional probability of an event A given B is the probability that the event A will occur given that an event B has already occurred

- Denoted by P(A|B)



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

Conditional probability

Question:

A pair of fair dice is rolled. If the sum of numbers that appear is 6, find the probability that one of the dice shows 2?

Conditional probability

Solution:

Let A: the event of getting the sum as 6

The ways A can occur: {(1,5), (2,4), (3,3), (4,2), (5,1)}

Let B: the event that number 2 appears on the dice

The ways B can occur: {(1,2), (2,1), (2,2), (2,3), (2,4), (2,5), (2,6), (3,2), (4,2), (5,2), (6,2)}

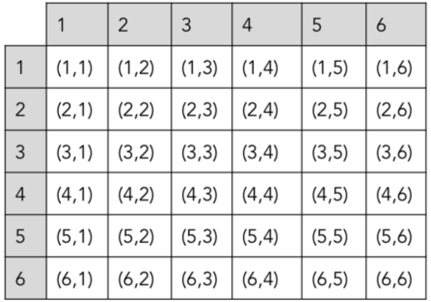
Thus, the event that sum of the die is 6 and number 2 appears on the dice is A ∩ B.

A ∩ B = {(2,4), (4,2)}

The total number of samples is 36.

First Dice

Second Dice



	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

Conditional probability

Solution:

The required probability is

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

$$P(\text{getting a number as 2} \mid \text{getting the sum as 6}) = \frac{\text{P (getting the sum as 6 and a number as 2)}}{\text{P(getting the sum as 6)}}$$

$$P(\text{getting a number as 2} \mid \text{getting the sum as 6}) = \frac{\frac{2}{36}}{\frac{5}{36}} = \frac{2}{5} = 0.4$$

Multiplication theorem

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

$$\Rightarrow P(A \cap B) = P(A \mid B). P(B)$$

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

$$\Rightarrow P(A \cap B) = P(B \mid A). P(A)$$

Thus, $P(A \cap B) = P(A \mid B). P(B) = P(B \mid A). P(A)$

Bayes theorem

- Conditional probability is the likelihood of an event given that another event has occurred

- Bayes theorem provides a way to update the probability based on the new information

- It is completely based on the conditional probability

- Known as Bayes’ Rule or Bayes law

Bayes theorem - formula

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

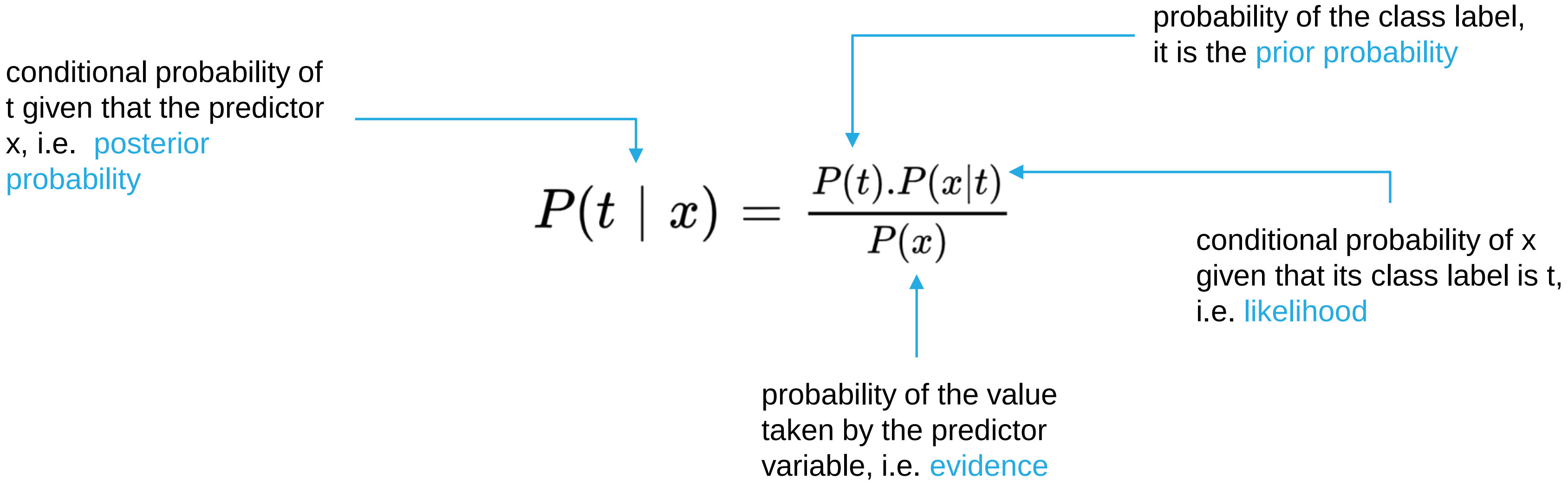
$$P(B \mid A) = \frac{P(B).P(A|B)}{P(A)}$$

Where, A and B are events

- P(A | B) the likelihood of event A occurring given that B is true
- P(B | A) the likelihood of event B occurring given that A true
- P(A), P(B): The independent probabilities of A and B

Bayes theorem - formula

For the naïve bayes classification the formula is as



conditional probability of t given that the predictor x, i.e. posterior probability

probability of the class label, it is the prior probability

$$P(t \mid x) = \frac{P(t).P(x|t)}{P(x)}$$

conditional probability of x given that its class label is t, i.e. likelihood

probability of the value taken by the predictor variable, i.e. evidence

Posterior Probability

In context with a classification problem, the posterior probability is the conditional probability of a class label taking value t given that the predictor takes value x

Example:

Consider the example of labelling an email as spam or ham. The conditional probability that it is a spam message given the word appears in it, i.e. P(spam | word) is the posterior probability

Prior probability

Prior probability is the probability of an event computed from the data at hand

Example:

Consider the example of labelling an email as spam or ham. The probability the email is spam, i.e. P(spam) is the prior probability
Likewise P(ham) is also a prior probability

Likelihood

In context with a classification problem, the Likelihood is the conditional probability of a predictor taking value x given that its class label is t

Example:

Consider the example of labelling an email as spam or ham. The conditional probability that the word appears in a spam, i.e. P(word | spam) is the likelihood

Evidence

- It is the probability that the predictor takes value x
- Also known as marginal probability

Example:

Consider the example of labelling an email as spam or ham. The probability that the word appears in a message, i.e. P(word) is the evidence

Naïve bayes classification

- A naïve bayes classifier uses the Bayes’ theorem for classification

- It is an eager learning algorithm. Since it does not wait for test data to learn, it can classify the new instance faster

Assumptions

Assumption 1: The predictors are independent of each other.

Example:

Consider the example of labelling an email as spam or ham.
The probability of the word *Good* appearing in the email is independent of the *Money*.

Thus $P(Good \cap Money) = P(Good) \, . \, P(Money)$ … since events are independent

Assumptions

Assumption 2: All the predictors have an equal effect on the outcome.

Example:

Consider the example of labelling an email as spam or ham. The appearance of a particular word in the email does not have more importance in deciding whether it is a spam or ham

Eg: The word *Friendship* does not have more importance to say whether it’s a spam/ham email.

Bayes theorem - classification problem

We have the Bayes theorem as

$$P(t \mid x) = \frac{P(t).P(x|t)}{P(x)}$$

For X=(x_1, x_2, …, x_n), applying the chain rule, we have

$$P(t \mid x_1, x_2, ..., x_n) = \frac{P(t).P(x_1|t).P(x_2|t)...P(x_n|t)}{P(x_1)P(x_2)....P(x_n)}$$

Since the denominator does not change for the values taken by the predictor as assumed in the second assumption. The denominator can be removed.

Bayes theorem - classification problem

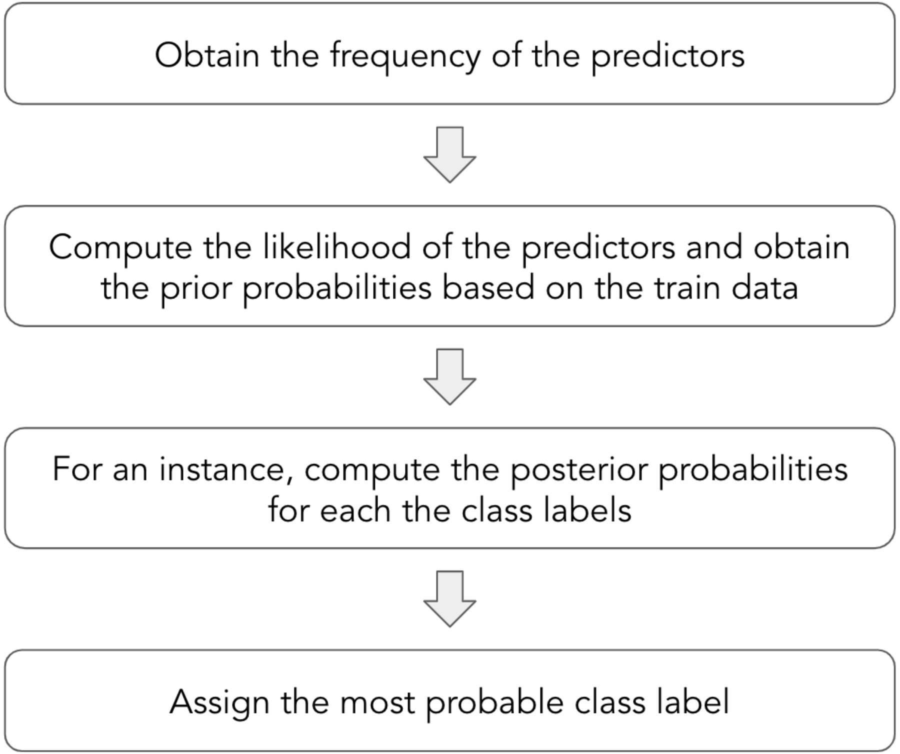
We get

$$P(t \mid x_1, x_2, ..., x_n) \propto P(t). P(x_1 \mid t). P(x_2 \mid t)...P(x_n \mid t)$$

For convenience, write it as

$$P(t \mid X) = P(t). P(x_1 \mid t). P(x_2 \mid t)...P(x_n \mid t)$$

Naïve Bayes: Procedure



Obtain the frequency of the predictors

Compute the likelihood of the predictors and obtain the prior probabilities based on the train data

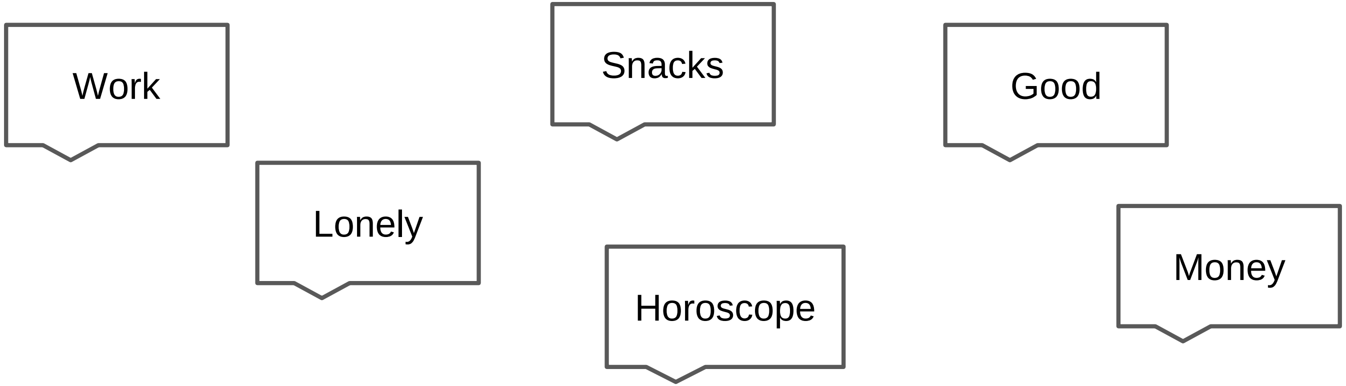
For an instance, compute the posterior probabilities for each the class labels

Assign the most probable class label

Business problem: label the email as spam or ham

- We shall consider the problem of labelling the received emails as spam or ham

- Choose a few words you find in emails



Work

Snacks

Good

Lonely

Horoscope

Money

Spam-ham example

1 Consider the frequency of these words used in spam and ham emails as shown below

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7

Spam-ham example

2

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7
Total	50	40

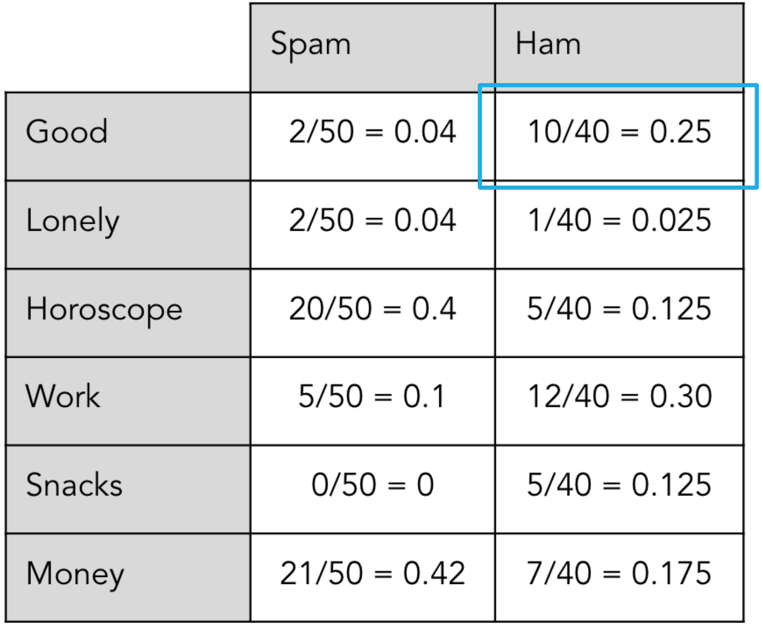
Obtain the likelihoods

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Likelihood

The probability that the word *Good* appears in a spam email, ie P(*Good* | Ham) is 0.25.

This is the Likelihood.



	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Spam-ham example

3 Obtain the prior probability

From the data we have 15% of the emails are spam and the remaining are ham

Thus the prior probabilities are

P(Spam) = 0.15 and P(Ham) = 0.85

Spam-ham example

4

Consider the word sequence *Good Work*, does it belong to a spam message?

Our instance is *Good Work*.

For our instance, compute the posterior probabilities for each the class labels - spam or ham

Spam-ham example

4 Compute the posterior probabilities for each the class labels - Spam or Ham

For Spam,

P(Spam| *Good*, *Work*) = P(Spam) . P(*Good* | Spam). P(*Work* | Spam) = (0.15) . (0.04) . (0.1)
= 0.0006

For Ham,

P(Ham| *Good*, *Work*) = P(Ham) . P(*Good* | Ham). P (*Work* | Ham)
= (0.85) . (0.25) . (0.30)
= 0.063

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Spam-ham example

5 Assign the most probable class label

For Spam,

P(Spam| *Good*, *Work*) = 0.0006

For Ham,

P(Ham| *Good*, *Work*) = 0.063

Since 0.063 > 0.0006, we assign the class label as Ham to the instance *Good Work*.

For Spam,

P(Spam| *Good*, *Work*) = 0.0006

For Ham,

P(Ham| *Good*, *Work*) = 0.063

$$\frac{\text{P(Spaml Good, Work)}}{\text{P(Spaml Good, Work)}+\text{P(Haml Good, Work)}} = \frac{0.0006}{0.0006+0.063} = 0.009$$

$$\frac{\text{P(Haml Good, Work)}}{\text{P(Spaml Good, Work)}+\text{P(Haml Good, Work)}} = \frac{0.063}{0.0006+0.063} = 0.991$$

The sum is 1

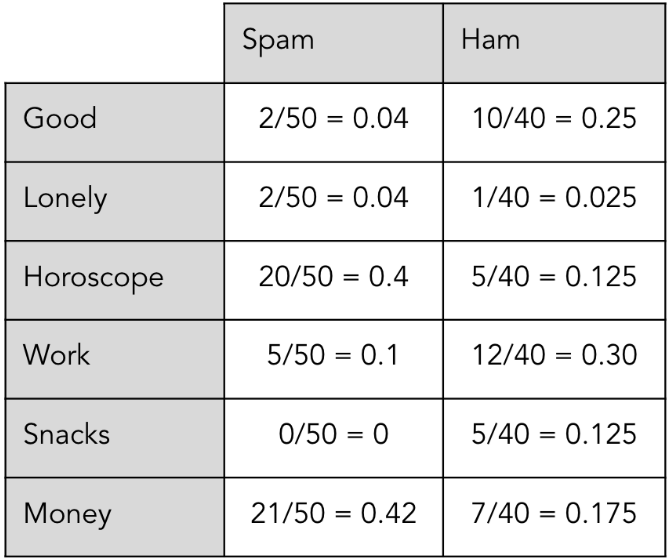
Obtain the class label

Question:

With help of the previous data. Label the email containing word Horoscope 1 time, Money 2 times and Snack 1 time

The prior probabilities are:

P(Spam) = 0.15 and P(Ham) = 0.85



	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*)

= P(Ham) . P (*Horoscope* | Ham) . P(*Money* | Ham). P(*Money* | Ham) . P(*Snack* | Ham)

= (0.85) . (0.125) . (0.175) . (0.175) . (0.125)

= 0.0004

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.4) . (0.42) . (0.42) . (0.00)

… Here is a problem

= 0.0

	Spam	Ham
Good	2/50 = 0.04	10/40 = 0.25
Lonely	2/50 = 0.04	1/40 = 0.025
Horoscope	20/50 = 0.4	5/40 = 0.125
Work	5/50 = 0.1	12/40 = 0.30
Snacks	0/50 = 0	5/40 = 0.125
Money	21/50 = 0.42	7/40 = 0.175

Obtain the class label

Solution:

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.4) . (0.42) . (0.42) . (0.00)

= 0.000

… No matter which other word(s) is seen along with Snack, the email will never be classified as Spam. Since the frequency for Snack is 0.

Laplace smoothing method

- To solve the zero probability problem, we use the Laplace smoothing method
- Add α to every count so; the count is never zero
- α > 0. Generally, α = 1
- Consider the α for the divisor as well

Obtain the class label

Solution:

	Spam	Ham
Good	2	10
Lonely	2	1
Horoscope	20	5
Work	5	12
Snacks	0	5
Money	21	7
Total	50	40

Add α = 1, to each count

	Spam	Ham
Good	3	11
Lonely	3	2
Horoscope	21	6
Work	6	13
Snacks	1	6
Money	22	8
Total	56	46

Obtain the class label

Solution:

	Spam	Ham
Good	3	11
Lonely	3	2
Horoscope	21	6
Work	6	13
Snacks	1	6
Money	22	8
Total	56	46

Obtain the new likelihoods

	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	21/56 = 0.37	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	1/56 = 0.02	6/46 = 0.13
Money	22/56= 0.40	8/46 = 0.18

Obtain the class label

Solution:

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*)

= P(Ham) . P (*Horoscope* | Ham) . P(*Money* | Ham). P(*Money* | Ham) . P(*Snack* | Ham)

= (0.85) . (0.13) . (0.18) . (0.18) . (0.13)

= 0.0004

	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	21/56 = 0.37	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	1/56 = 0.02	6/46 = 0.13
Money	22/56= 0.40	8/46 = 0.18

Obtain the class label

Solution:

For Spam,

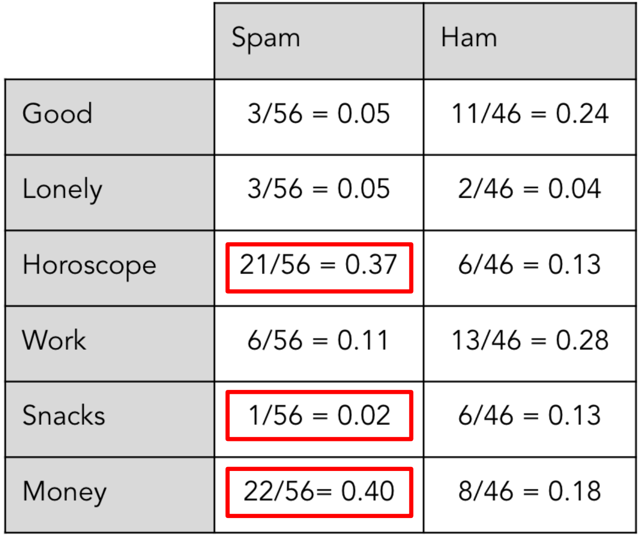
P(Spam| *Horoscope, Money, Money, Snack*)

= P(Spam) . P (*Horoscope* | Spam) . P(*Money* | Spam). P(*Money* | Spam) . P(*Snack* | Spam)

= (0.15) . (0.37) . (0.4) . (0.4) . (0.02)

= 0.0017

… The problem is solved using the Laplace smoothing method



	Spam	Ham
Good	3/56 = 0.05	11/46 = 0.24
Lonely	3/56 = 0.05	2/46 = 0.04
Horoscope	21/56 = 0.37	6/46 = 0.13
Work	6/56 = 0.11	13/46 = 0.28
Snacks	1/56 = 0.02	6/46 = 0.13
Money	22/56= 0.40	8/46 = 0.18

Obtain the class label

Assign the most probable class label

For Spam,

P(Spam| *Horoscope, Money, Money, Snack*) = 0.0017

For Ham,

P(Ham| *Horoscope, Money, Money, Snack*) = 0.0004

Since 0.0017 > 0.0004, we assign the class label as Spam to the instance *Horoscope, Money, Money, Snack*.

Naïve Bayes Classifier available in the scikit learn library

- Gaussian Naïve Bayes:
  - It is used when predictors are continuous
  - Assumes that the predictors follow normal distribution
  - The Gaussian Naïve Bayes Classifier used the normal distribution for classification

- Multinomial Naïve Bayes
  - Used for document classification problem - classify whether a document is a sport, history, or science article
  - The predictors are the frequency of the words present in the article

- Bernoulli Naïve Bayes
  - This is similar to the multinomial naive bayes, but the predictors are binary valued (boolean)

Applications of Naïve Bayes

- Spam Filtering
- Sentiment Analysis
- Recommendation System

Naïve Bayes: advantages

- Easy to implement in the case of text analytics problems
- Used for multiple class prediction problems
- Performs better for categorical data than numeric data

Naïve Bayes: disadvantages

- Fails to find relationship among features

- May not perform when the data has more number of predictor

- The assumption of independence among features may not always hold good

Thank You