# CLASIFICATION

# What is classification?

We are trying to predict a variable Y that is categorical.

We say Y is categorical if Y ∈ {1,...,K}

We will mostly focus on binary responses Y ∈ {0, 1}

# Examples

**Purchase**: what brand of phone will you buy?

– Apple (Y =2) or Samsung (Y =1) or None (Y =0)

**Churn:** will a customer cancel their subscription?

– Leave (Y =1) or Stay (Y =0)

**Credit:** will a person default on a loan?

– Default (Y =1) or Pay (Y =0)

**Football**: will the footbal player score today?

– Score (Y = 1) or Doesn’t Score (Y = 0)

# Nearest Neighbour Classification

## Euclidean distance

Definition. Let the Euclidean distance between two points a = (X1a, X2a) and b = (X1b, X2b) be defined as

d(a,b)=||b−a||=

This measure of similarity yields nearest neighbor (NN) classification algorithms

## Theory of kNN

1. compute d(Xi,X∗) for each i = 1,...,n
2. sort Y1,...,Yn based on the values d(X1,X∗),...,d(Xn,X∗) so that Y(1) corresponds to the observation with the smallest distance
3. predict y∗ based on the classification rule  
   Y∗ =cknn(X∗)=V(Y(1),...,Y(k))

where V (·) returns the “majority vote” – i.e. the value that is returned most often among Y(1), . . . , Y(k)

## Example

**3rd – nearest neighbors**

**classifier**

Consider the following data set on spam email classification. The **spam** variable indicates whether the email is actually spam or not, **n\_foreign** reports the number of foreign characters used in the message, and **cap\_run\_length** counts the longest string of capital letters used in the message.

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **spam** | **N\_foreign** | **Cap\_run\_lenght** |
| 1 | Yes | 10 | 15 |
| 2 | No | 2 | 4 |
| 3 | No | 5 | 8 |
| 4 | Yes | 4 | 10 |
| 5 | no | 4 | 2 |

Suppose you observed the following data on a new email:

|  |  |  |  |
| --- | --- | --- | --- |
| **ID** | **spam** | **N\_foreign** | **Cap\_run\_lenght** |
| 6 | ? | 6 | 10 |

Find the 3rd – nearest neighbors classifier of the new email (using a majority voting function). What is the probability of predicting the new email as spam?

1. Euclidean distance

d(a,b)=||b−a||=

d(6,1)= ==6.40

d(6,2)===7.21

d(6,3)===2.23

d(6,4)= =

d(6,5)===

1. Sorting (ranking)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| X1 | X2 | y | distance | rank |
| 10 | 15 | yes | 6.40 |  |
| 2 | 4 | no | 7.21 |  |
| 5 | 8 | no | 2.23 |  |
|  |  |  |  |  |
|  |  |  |  |  |

1. Majority vote -rule ( 3-NN )

Y∗ =cknn(X∗)=V(Y(1),...,Y(k)) = V(Y(1),Y(2),Y(3))=

The probability of predicting the new email as spam is ...

## Conclusions

How many neighbours (k) to choose?

The common practice is to try more values for k, then see which model has the highest accuracy rate.