K-Nearest Neighbour (KNN)

October 26, 2017

Outline

1 Introduction

2 KNN

3 KNN example

Parametric model

- We select a hypothesis space and adjust a fixed set of parameters with the training data.
- We assume that the parameters summarize the training and we can forget about it. This methods are called parametric models.
- When we have a small amount of data it makes sense to have a small set of parameters and to constraint the complexity of the model.

Non-Parametric model

- If data shows that the hypothesis has to be complex, we can try to adjust to that complexity.
- A non parametric model is one that can not be characterized by a fixed set of parameters
- A family of non parametric models is Instance Based Learning

Instance Based Learning

- Instance based learning is based on the memorization of the dataset
- The number of parameters is unbounded and grows with the size of the data
- There is not a model associated to the learned concepts
- The classification is obtained by looking into the memorized examples
- The cost of the learning process is 0, all the cost is in the computation of the prediction
- This kind learning is also known as lazy learning

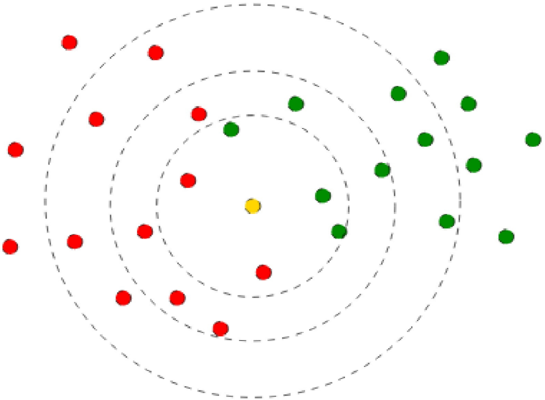
KNN

- KNN uses the local neighborhood to obtain a prediction
- The K memorized examples more similar to the one that is being classified are retrieved
- A distance function is needed to compare the examples similarity
  - Euclidean distance $d(x_j, x_k) = \sqrt{\sum_i (x_{j,i} - x_{k,i})^2}$
  - Mahnattan distance $d(x_j, x_k) = \sum_i |x_{j,i} - x_{k,i}|$
- This means that if we change the distance function, we change how examples are classified

KNN Algorithm

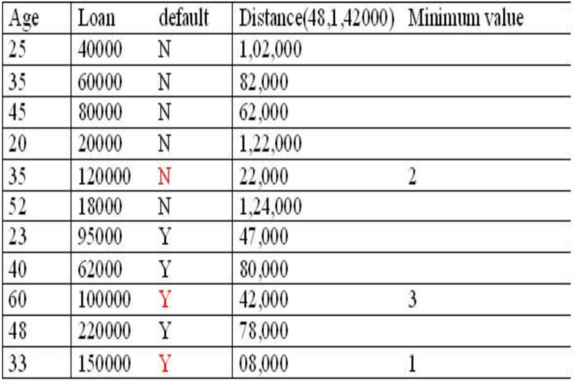
- Training: Store all the examples
- Prediction: $h(x_{new})$
  - Let be $x_1, x_2, x_3, ..., x_k$ the k more similar examples to x_{new}
  - $h(x_{new})=$ combine predictions $x_1, x_2, x_3, ..., x_k$
- The parameters of the algorithm are the number k of neighbours and the procedure for combining the predictions of the k examples
- The value of k has to be adjusted
  - Can overfit (k too low)
  - Can underfit (k too high)

KNN prediction



KNN Example

Test case Input- Age=48,loan=1,42000



Age	Loan	default	Distance(48,1,42000)	Minimum value
25	40000	N	1,02,000	
35	60000	N	82,000	
45	80000	N	62,000	
20	20000	N	1,22,000	
35	120000	N	22,000	2
52	18000	N	1,24,000	
23	95000	Y	47,000	
40	62000	Y	80,000	
60	100000	Y	42,000	3
48	220000	Y	78,000	
33	150000	Y	08,000	1

$$\sqrt{(48-25)^2+(1,42,000-40000)^2} = 102000.002$$

With K=3, there are two Default=Y and one Default=N out of three closest neighbours. The prediction for the Test case Input- Age=48,loan=1,42000 is Y

THANK YOU