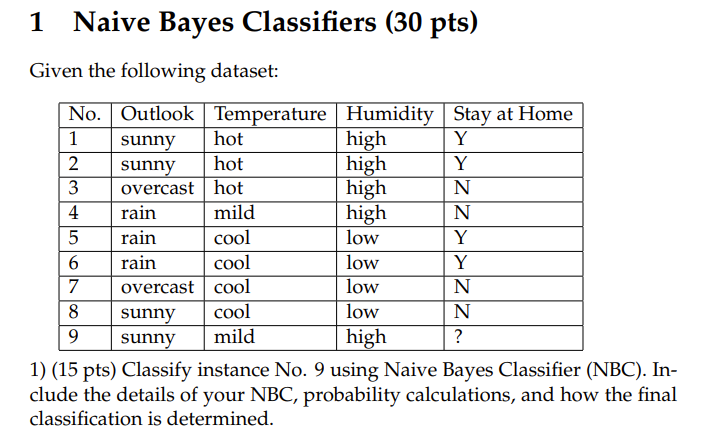
COMP631 Introduction to IR

Homework 3



P(Y) = 4/8 = 0.5

P(N) = 4/8 = 0.5

P(sunny|N) = 0.25

P(sunny|Y) = 1/2= 0.5

P(mild|N) = ¼ = 0.25

P(mild|Y) = 0

P(high|N) = ½ =0.5

P(high|Y) = ½ =0.5

P(Instnace 9|N) = P(N) \* P(sunny|N)\*P(mild|N) \*P(high|N)

= 4/8\*¼\*¼\*2/4

= 1/64 = 0.015

P(Instnace 9|Y) = P(Y) \* P(sunny|Y)\*P(mild|Y) \*P(high|Y)

= 4/8 \* 2/4 \* 0 \* 2/4

= 0

The probability of N is higher than probability of Y

Thus the answer is N

**2) (5 pts) What is the time complexity for training and testing Naive Bayes classifier, respectively?**

ANSWER

Let N = number of training examples,

D= dimensionality of the features

C = number of classes.

Then *training complexity* of Naive Bayes is O(ND), it computes frequency of every feature value for each class.

*Testing complexity* of Naive Bayes is in O(CD), it retrieve D feature values for each of the Cclasses.

3) (10 pts) After a yearly checkup for a software developer, there are both bad news and good news from the doctor. The bad news is that the developer has a test result positive for a serious disease, and the test is 98% accurate (i.e., if you have the disease, then the probability of testing positive is 0.98; if you do not have the disease, the probability of testing negative is also 0.98). The good news is that this is a rare disease, because only 1 in 20,000 people will have it. What are the chances that the developer actually has the disease?

ANSWER

P(ill) = 1/20000

P(healthy) = 1999/20000

p(positive|ill) = 0.98

p(positive|healthy)= 0.02

P(positive) = p(positive|ill) \* P(ill) + p(positive|healthy)\*P(healthy)

= 0.98\*1/20000 + 0.02\*19999/20000

=0.20048

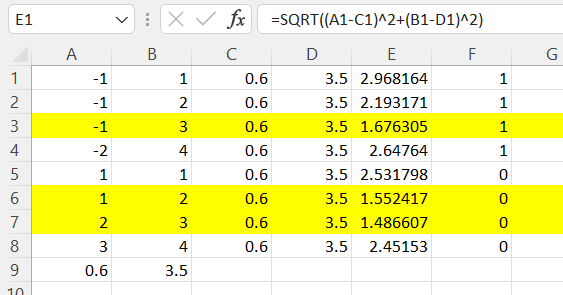
We can find P(positive|ill) using bayes theorem

P(positive|ill) = P(positive|ill)\*P(ill)/P(positive)

= (0.98\*1/20000)/0.020048 = 0.00244413

# 2. kNN Classifiers (20 pts)

## 1) (8 pts) Classify instance No. 9 using the 3-Nearest-Neighbors classifier. Use squared distance (i.e., Euclidean distance, L2 distance) as the distance measured between two instances.



The 3 nearest neighbors are

d(3,9) = 1.68, 3 is classified 1

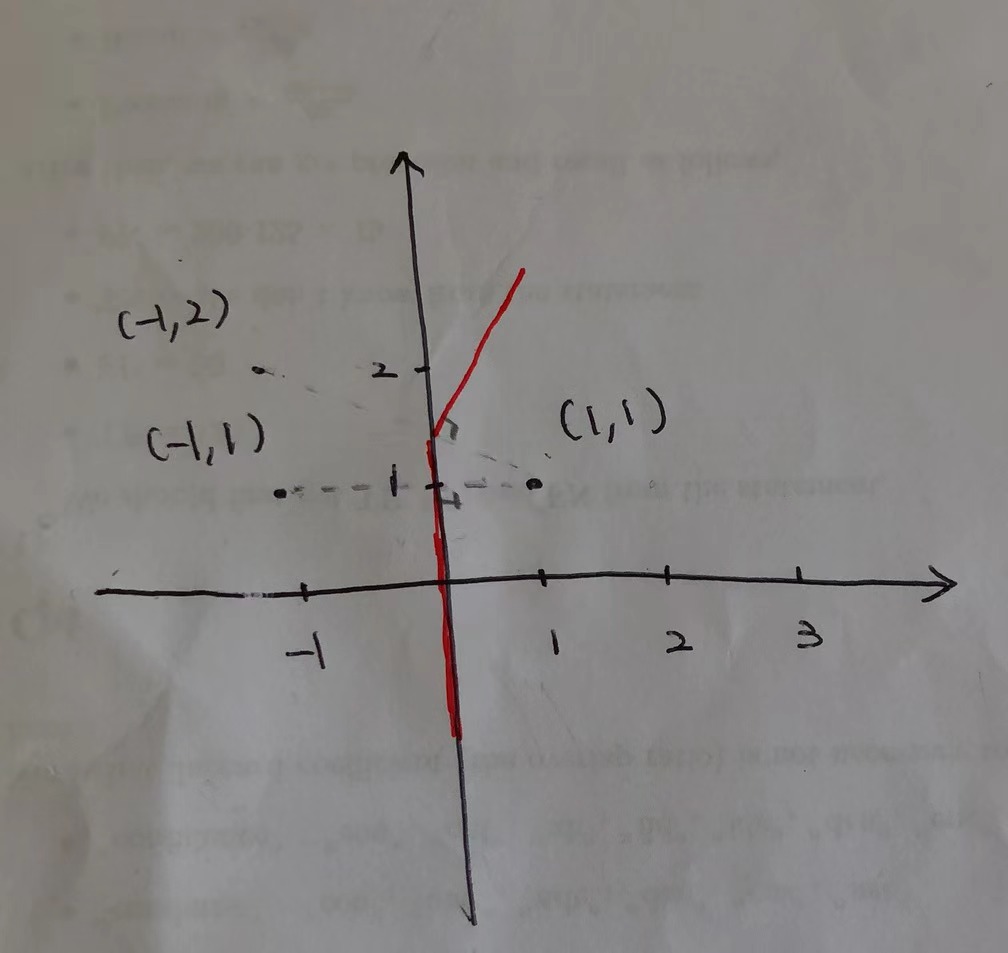
d(6,9) = 1.55, 6 is classified 0

d(7,9) = 1.48, 7 is classified 0

Thus 9 is predicted to be classified as 0

## 2) (6 pts) Draw the decision boundary of 1-Nearest-Neighbor classifier. Only consider the instance NO. 1, 2, 5 in the table above as training data.

ANSWER



## 3) (6 pts) Explain at least two disadvantages of kNN classifiers.

* Computational complexity — As your training data increases, the speed at which calculations are made rapidly decreases
* Poor performance on imbalanced data — When the majority of the data the model is being trained on represents 1 label then that label will have a high likelihood of being predicted
* The optimal value of K — If chosen incorrectly, the model will be under or overfitted to the data

# 3 Clustering (35 pts)

## 1) (6 pts) In Figure 1, there are two clusters colored in red and blue respectively, obtained by using k-means clustering method with k = 2. Suppose now we have a new sample (colored in green) to be assigned to one cluster. By using the squared distance to the cluster centroid, which cluster will the sample be assigned to? Do you think this assignment is reasonable? Explain. (Assume the figure is precisely drawn to reflect the position of data points.)

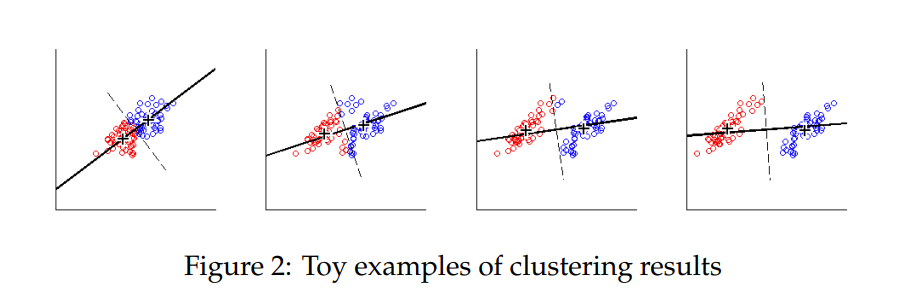
ANSWER

It should be assigned to the blue cluster because the green data point is closer to the centroid of blue clusters than the centroid of red clusters

However, i think it should be assigned to the red cluster because it closer to the data point in the red cluster than the points on the blue cluster

## 2) (8 pts) In Figure 2, by using k-means algorithm with squared distance measure, samples are clustered into two groups colored in red and blue respectively. The centroid for each cluster is shown in “+”. The decision boundary is shown in the dotted line. Explain the problem of k-means algorithm observed from the figure above.

Explain how can we improve the traditional k-means algorithm to solve the problem. (You may want to combine the phenomenon in Problem 4.1. Illustrating the high level idea is enough.)



K means Assumes clusters are spherical in vector space in this case which is circle but this is not how the data aligned in this case..

K-means based on Euclidean distance assumes that the data of each data cluster has the same prior probability and presents a spherical distribution, but this distribution is not common in real life. In the face of non-convex data distribution shape, we can introduce kernel function to optimize.

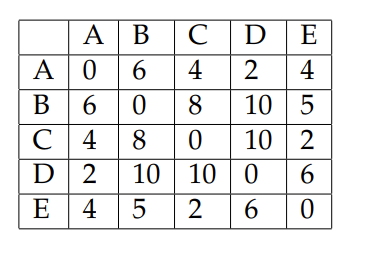
At this time, the algorithm is also called the kernel K-means algorithm, which is a kind of kernel clustering method. The main idea of ​​the kernel clustering method is to map the data points in the input space to the high-level feature space through a nonlinear mapping and perform clustering in the new feature space. The nonlinear mapping increases the probability that the data points are linearly separable so that when the classical clustering algorithm fails, more accurate clustering results can be achieved by introducing a kernel function.

## 3) (6 pts) Explain the disadvantages of the flat clustering algorithms.

ANSWER

1. Hard to choose the number of clusters, bad choice of a number of clusters causes bad results
2. bad initial partitions may cause bad final clusters
3. Choice of initial seeds has a strong impact on final results
4. Return unstructured clusterings and results are not predictable

## 4) (15 pts) Given five data points (A, B, C, D, E) and their pairwise distances in the table below, using the single link (MIN), please work out the dendrogram for clustering with necessary details.



ANSWER

