Notebook No:__________:מס׳ מחברת **I.D number:______________:מס׳ ת.ז.**

בית ספר ״אפי ארזי״ למדעי המחשב המרכז הבינתחומי

The Efi Arazi school of computer science

The Interdisciplinary Center

סמסטר ב׳ תשע״ז
Spring 2017

מבחן מועד א בלמידה ממוכנת

Machine Learning Exam A

Lecturer : Prof Ariel Shamir,
Dr. Zohar Yakhini
Time limit : 3 hours
Additional material or calculators are not allowed in use!

Answer 5 out of 6 from the following questions (each one is 20 points)

Good Luck!

מרצים: פרופ אריאל שמיר,
ד״ר זהר יכיני

משך המבחן: 3 שעות
אין להשתמש בחומר עזר ואין להשתמש במחשבונים!

יש לענות על 5 מתוך 6 השאלות הבאות
לכל השאלות משקל שווה (20 נקודות)

בהצלחה!

Question 1 (5 parts)

In $\mathbb{R}^2$ consider a hypotheses space of the interiors of "top-right facing right angled isosceles triangles" – all isosceles triangles where the two equal sides are parallel to the x and y axes and the hypotenuse is on the upper right.

Such triangles can be created by intersecting the following lines: x=a, y=b, x+y=r, where r>a and r>b.

Formally:

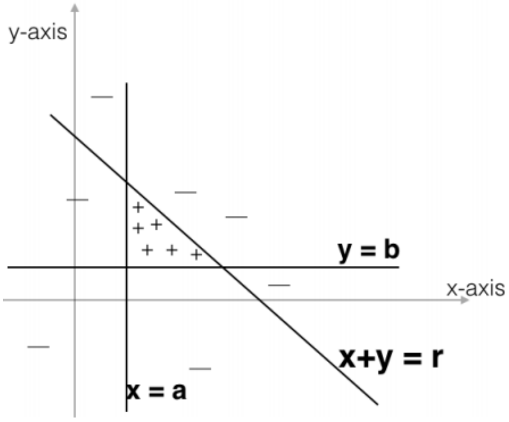
$H_\Delta = \{h_{a,b,r}: a, b, r \in \mathbb{R}, \mathrm{r} > \mathrm{a}, \mathrm{r} > \mathrm{b}\}$

where positives are inside:

$$h_{a,b,r} = \begin{cases} 1 & if\ x \geq a\ and\ y \geq b\ and\ x + y \leq r \\ -1 & otherwise \end{cases}$$

That is: $h_{a,b,r}$ assigns a positive label to any point in the interior or on the edges of the triangle.

This is a picture of a single hypothesis in this space:



Recall the three sample complexity bounds we have learned in class:

$$m \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$
$$m \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$
$$m \geq \frac{1}{\varepsilon}\left(8 \cdot VC(H)\log_2\frac{13}{\epsilon} + 4\log_2\frac{2}{\delta}\right)$$

A. Given that we only consider $1 \leq a = b \leq n$ and $1 \leq r \leq 2n$ and for integer valued a, b, r , compute the size of the hypotheses space, as a function of n

Solution

We will check the number of possible values for a, b, r when we know that $a + b \leq r$

When $a = b = 1$, r can be any integer in the range $[2, 2n]$ – 2n-1 options

When $a = b = 2$, r can be any integer in the range $[4, 2n]$ – 2n-3 options

⋮

When $a = b = \mathrm{n}$, r can be any integer in the range $[2\mathrm{n}, 2n]$ – 1 option

We got a total of $1 + 3 + \cdots + (2n - 1)$ possible values for a, b, r.

This is an arithmetic series and its sum is: n^2.

B. Let n = 100.

1. Define a consistent learning algorithm for this case. That is – find $h_c \in H_\Delta$ from any given training data that can be assumed to have been generated from a concept $c \in H_\Delta$

Solution

Find the minimum X value

Find the minimum Y value

Find the maximum X+Y value

The triangle will be:

$$a = b = floor(\min(x, y))$$
$$r = ceil(x + y)$$

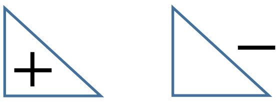
2. Show that the space H_{Δ} is PAC learnable by the algorithm you defined above (no need to provide a tight bound)
   Solution
   The algorithm is running in polynomial time
   The hypothesis space is finite and contain the target concept. Therefor we can use the first bound in order to show that the number of training instances is polynomial.

C. Assume that a=b=0 and that r can be any positive real number. Compute the VC dimension of the resulting hypotheses space H_{Δ}

Solution

VC=1

One instance can be shattered:



* The instance need to be in the first quadrant (otherwise, it will always be outside of the triangle).

Given 2 instances in the first quadrant (if at least one of them is not in the first quadrant it will always be outside of the triangle, and therefor can't be shattered), (x_1, y_1), (x_2, y_2), we will calculate r_1, r_2 as follows:

$$r_1 = x_1 + y_1$$
$$r_2 = x_2 + y_2$$

We can't shattered those instances because if $r_1 \le r_2$ the first instance can't be negative and the second positive, and if $r_1 < r_2$ the second instance can't be negative and the first positive.

D. Consider the learning algorithm you had defined in B1 (recall that n=100 and values are integers). Give a bound on the number of samples required to learn a concept with 90% probability and an error of at most 0.05. Show the required calculation only. (provide a formula, No need to provide final answer)

Solution

We will use the first bound (finite hypothesis space that contain the target concept)

$$m \ge \frac{1}{0.05}\left(\ln 100^2 + \ln\frac{1}{0.1}\right)$$

E. Does your answer to D change if we use the space definition from C? In not – why? If yes – how would you compute a bound on the sample size here (provide a formula, No need to provide final answer)

Solution

In this case the hypothesis space is infinite, and therefore we can't use the first or the second bound, and we will use the third one

$$m \ge \frac{1}{0.05}\left(8 \cdot 1\, log_2 \frac{13}{0.05} + 4\, log_2 \frac{2}{0.1}\right)$$

Question 2 (7 parts)

You are given a set of m instances that are defined by some feature x (i.e. x_i is the value of the feature of instance i), and the values of the target function y defined on each instance as y_i. The formula to calculate the Pearson Correlation Coefficient between the feature x and the target function y is given by:

$$\rho = \frac{\sum_{i=1}^{m}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{m}(x_i - \mu_x)^2 \sum_{i=1}^{m}\left(y_i - \mu_y\right)^2}}$$

A. Explain what does the correlation formula measures? what is the meaning of the nominator and what is the meaning of the denominator in the given formula? and what are the possible values of this formula?

Solution

The given correlation formula model a linear relationship (dependence) between 2 variables. The numerator describes linear dependency between the two variables x and y. The denominator contains the standard deviation (variance root) of each of the variables. The possible values are between -1 and 1.

B. Explain the connection between correlation and dependency between x and y (as random variables)? Explain the meaning of this connection and the dependency between x and y in the following cases:

- The correlation is 1
- The correlation is -0.8
- The correlation is 0

Solution

If the variables are independent then the correlation between them will necessarily be 0. Thus, if the correlation is different than 0, then the variables cannot be independent (there is some linear relationship between them).

- If the correlation value is 1 then they are dependent and the linear relationship between the variables is the strongest possible, accurate and identical in its direction: If x grows at a constant size then y also grows at a fixed size.
- If the value is -0.8, the variables depend and the relation is not accurate but still strong and reversed: if x increase then y decrease.
- If the correlation is 0, then there is no linear relationship between the variables, but this does not necessarily mean that they are independent.

C. Explain the connection between the correlation measure and linear regression that is trying to explain y using x. In which of the three cases appearing in Section B it is worthwhile to use linear regression? Explain why. Will there be an error in the evaluation of y in each of these cases?

Solution

Because a linear regression searches for a linear function that explains the relationship between x and y, then when the correlation is greater than 0 the linear relationship is stronger and the regression is more successful (there will be less error in the estimation). Therefore, if the correlation is 1 we can find a linear function that will explain y with no error. If the correlation is -0.8, then the regression can be used to explain y but the direction will be reversed (the slope of the function will be negative) and there will be some error in the estimation. If the correlation is 0 then it is not possible to find a linear relationship between them and the error will be the largest possible and it is not advisable to use regression.

Now assume that the instances have n different features and not just one. We will indicate each feature with a lower index such as x_i is the i-th feature, and to indicate an instance from the set we will use an upper index. For example, x_i^d will mark the i-th feature of the d-th instance from the set of m instances.

D. Explain (including formula) what is the covariance matrix (or scatter matrix) of the features of the set of instances, and explain what is the connection between its elements and the Correlation Coefficient from A.
Solution
The covariance matrix is a matrix in which the elements within it are the covariance between each pair of features σ_{ij} outside the main diagonal and the variance of each σ_{ij} variable on the main diagonal:
$$\sigma_{ij} = \sum_{d=1}^{m}(x_i^d - \mu_i)(x_j^d - \mu_j) \quad \sigma_i = \sum_{d=1}^{m}\left(x_i^d - \mu_i\right)^2 \text{ when } \mu_i = \frac{1}{m}\sum_{d=1}^{m} x_i^d$$
$$S = \begin{pmatrix} \sigma_1 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n \end{pmatrix}$$
The formula for correlation between two variables is actually their covariance normalized (divided) by the variance of both. Therefore, the matrix contains values that express the linear relationship (unnormalized) between two variables outside the diagonal (and the variance within the diagonal).

E. Explain what is the purpose of PCA algorithm (principal component analysis)
Solution
The goal of the PCA algorithm is to transform the features space so that there will be no correlation between each pair of variables and also to sort the new axes according to their variance so that the first axis will have the largest variance and the last with the smallest variance. This transformation involves moving the origin of the axes to the center of gravity of the variables (centroid) + rotation of the axes. The new axes (the new coordinates after the transformation) are called the Principal Components. This algorithm can be used for various applications, including dimension reduction and compression - using the first set of coordinates.

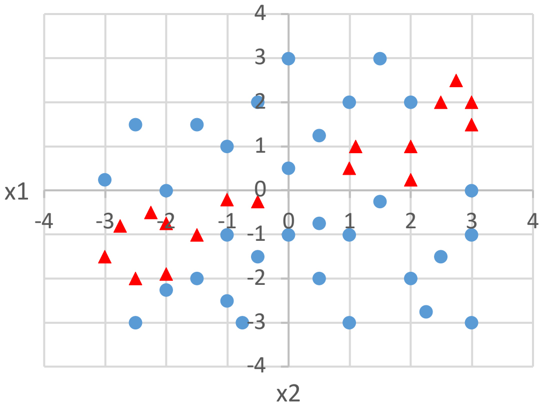
F. Explain what is the purpose of LDA algorithm (linear discriminant function)
Solution
The goal of the LDA algorithm is to find a transformation by moving and rotating the axes and then if we use dimension reduction then the classes will be best separated - their centers will be far apart and the variance of each class will be small (the instances will be centered around the center).

G. Explain what is the different of the scatter matrix in use in PCA algorithm and the one in use in LDA algorithm.
Solution
The PCA algorithm uses the scatter matrix, which is actually the covariance matrix from section D and finds eigenvectors. In this matrix, all instances are involved and there is no difference between them (because most of the use is for unsupervised learning). In contrast, in the LDA algorithm the classes are known. Thus, a covariance matrix is constructed for each class separately (according to the class instances and

average) and the within-class-scatter matrix is essentially the sum of all the covariance matrices of all classes.

Question 3 (6 parts)

Consider the following training data:



A. Will a perceptron learning algorithm find a good classifier for this training set? If yes – what are the weights for the separator? If not – explain why not.

Solution

The perceptron will not be able to find a separator for this training set because the set is not linearly separated, and therefore the algorithm will not converge.

B. We applied SVM learning to this training data. We have tried different kernels. We obtained the following classification rule (execution algorithm):

$$t(x) = \mathrm{sgn}\big(g(\mathrm{x})\big) = \mathrm{sgn}\left(\sum_{i \in SV} \alpha_i t_i (x_i \cdot x)^2\right)$$

Where

SV is the set of support vectors, α_i is the weight form the i-th support vector, t_i is the class the i-th support.

What kernel $K(x, y)$ was used here?

Solution

The kernel is $K(x, y) = (x \cdot \mathrm{y})^2$

C. Find $\varphi: \mathbb{R}^2 \to \mathbb{R}^3$, so that the kernel function from B satisfies $K(x, y) = \varphi(x) \cdot \varphi(y)$.

Solution

$$K(x, y) = (x \cdot \mathrm{y})^2 = (x_1 y_1 + x_2 y_2)^2 = {x_1}^2 {y_1}^2 + 2 x_1 y_1 x_2 y_2 + {x_2}^2 {y_2}^2 =$$
$$\left({x_1}^2, \sqrt{2} x_1 x_2, {x_2}^2\right) \cdot \left({y_1}^2, \sqrt{2} y_1 y_2, {y_2}^2\right) = \varphi(x) \cdot \varphi(y)$$

D. Find the weights of the linear separator in $\mathbb{R}^3$ which is equivalent to $x_2(x_2 - x_1) = 0$

Solution

$$x_2(x_2 - x_1) = {x_2}^2 - x_1 x_2 =$$
$$w \cdot \varphi(x) = (w_1, w_2, w_3) \cdot \left({x_1}^2, \sqrt{2} x_1 x_2, {x_2}^2\right) =$$
$$\left(0, -\frac{1}{\sqrt{2}}, 1\right) \cdot \left({x_1}^2, \sqrt{2} x_1 x_2, {x_2}^2\right)$$
$$\downarrow$$

$$w = \left(0, -\frac{1}{\sqrt{2}}, 1\right)$$

E. Assume that an RBF kernel was selected. That is – a kernel of the form:

$$K(x,y) = \varphi(x) \cdot \varphi(y) = exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

Prove that for every two points $x, y \in \mathbb{R}^2$ the following holds:

$$\|\varphi(x) - \varphi(y)\|^2 = 2 - 2exp\left(-\frac{1}{2}\|x-y\|^2\right)$$

Solution

$$\begin{aligned}&\|\varphi(x) - \varphi(y)\|^2\\ &= (\varphi(x)-\varphi(y))(\varphi(x)-\varphi(y))\\ &= \varphi(x)\varphi(x) + \varphi(y)\varphi(y) - 2\varphi(x)\varphi(y)\\ &= 2 - 2exp\left(-\frac{1}{2}\|x-y\|^2\right)\end{aligned}$$

F. For each of the following statements decide TRUE or FALSE and explain your answer:

1. After mapping into higher dimensional space using the RBF it is possible that a perceptron achieves better separation than in the original space.

   Solution

   True. It is possible that in the higher dimensional space the instances will be separated linearly and the results will be better (although this does not necessarily always exist).

2. After mapping into higher dimensional space using the RBF it is possible that a 1-NN classification algorithm, based on unweighted Euclidean distance, achieves better classification results than in the original space.

   Hint: use the identity proven in E.

   Solution

   False. Suppose x_i, x_j are two neighbors of the test instant x with the following property:

   $$\|x - x_i\| < \|x - x_j\|$$

   That is, x_i is closer to x than x_j.

   After mapping with RBF, we get:

   $$\begin{aligned}\|\varphi(x) - \varphi(x_i)\|^2 &= 2 - 2exp\left(-\frac{1}{2}\|x-x_i\|^2\right)\\ &< 2 - 2exp\left(-\frac{1}{2}\|x-x_j\|^2\right) = \|\varphi(x) - \varphi(x_j)\|^2\end{aligned}$$

   That is, if x_i was the nearest neighbor in the original dimension, he will be the nearest neighbor in the new dimension, so the classification will not change and the results will be the same.

   * The equalities derive from section 5 and the inequality derives from the inequality in the original space

3. The answer to F2 doesn't change if we use K-NN with weighted distances, where k>1

   Solution

   False. The distances in the higher dimensional space may be different, and therefore the neighbors' weights may be different, which can affect the classification and the results.

Question 4 (7 parts)

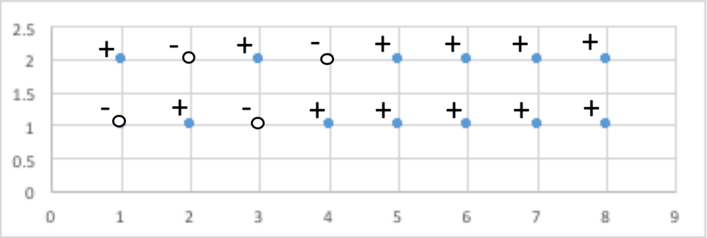
The following table and graph represent a data set containing two features x1 and x2 from two classes that we mark as + and -. This set is used as the training set for a decision tree.

We create two decision trees:

TOver Tree is the tree that is extended until the end of the algorithm with no limit and no pruning.

TUnder is the tree that contains only one root node where all the data is included.

instance	x1	x2	Value
1	1	2	+
2	2	1	+
3	3	2	+
4	4	1	+
5	5	1	+
6	5	2	+
7	6	1	+
8	6	2	+
9	7	1	+
10	7	2	+
11	8	1	+
12	8	2	+
13	1	1	-
14	2	2	-
15	3	1	-
16	4	2	-



A. Write the formula for “Goodness of Split” criterion that determines the feature that is chosen to split the decision tree and explain it.

B. To build the TOver tree, assume that we are using the GiniIndex as the function that measures how homogeneous a set is:

$$GiniIndex(S) = 1 - \sum_{i=1}^{c} (p_i)^2 = 1 - \sum_{i=1}^{c} \left(\frac{|S_i|}{|S|}\right)^2$$

Explain (including formula) how this measure is used to determine the Goodness of Split.

C. Explain how the first split in TOver tree is determined? How many Goodness of Split calculations must be done? (there is no need to calculate the numbers or result, just explain how many and which calculation should be done to determine the first split).

D. Assume that the first split in TOver tree is defined by the formula: x1<4.5. How many leaves would the tree have at the end of the learning?

E. Can there be a situation where you stop building a decision tree before all leaves are homogeneous (contain instances of only one class)? Explain why and if so, how is the classification of a new instance that reaches such a leaf determined?

F. What would be the total error of TOver tree using leave-one-out? Explain!

G. What would be the total error of TUnder tree using leave-one-out? Explain!

Solution

Goodness of split allows for assessing the splitting a node according to a feature. Assume that we have a function $\phi(S)$ that measures the in-homogeneity in a set of samples S. Then goodness of split can be defined as the difference between $\phi(S)$ in the original node and the weighted average of $\phi(Si)$ in the new nodes. If the total in-homogeneity goes down we get a positive number. We therefore select the feature and the threshold that yields the biggest difference.

$$\Delta\emptyset(S,A)=\emptyset(S)-\sum_{v\in Values(A)}\frac{|S_v|}{|S|}\emptyset(S_v)$$

Here we use the Gini Index as the function $\phi(S)$. We split according to the associated goodness of split function, the Gini Gain:

$$GiniGain(S,A)=GiniIndex(S)-\sum_{v\in Values(A)}\frac{|S_v|}{|S|}GiniIndex(S_v)$$

To find the best split in TOver we need to check all split cutoffs for each one of the two features. That is, for x2 we check splitting according to x2<1.5 and for x1 we need to check all options x1<1.5, x1<2.5 , … , x1<7.5. In total – 8 options. In each such test we compute the Gini Gain as above.

The tree continues to grow until we reach full classification. Therefore, all leaves will be perfectly homogeneous. The 8 samples on the right will be in one leaf. Each of the other 8 samples will be in their own leaves. In total 9 leaves.

It is possible that we will stop splitting a tree before all leaves are homogeneous. This will be done to avoid overfitting. For example – if we have a test set and we measure test error before every split to see if the test error improves. To use such a tree, with leaves that may not be perfectly homogeneous, for classification, we take decision by a majority vote in the leaf.

In TOver all 8 samples on the right will be correctly classified. However – when leaving any of the other samples out it will be classified according to the closest neighbor (which still remains in the training and defines the leaf to which the left out sample ends up going) and therefore erroneously classified. In total 8 errors.

TUnder classifies everything as +. Therefore, all positive left out samples will be correctly classified and all negative erroneously classified. A total of only 4 errors.

Question 5 (6 parts)

We want to cluster to k groups a set S of training examples. Below is a pseudo code of an algorithm that is called k-medoids and is similar to k-means:

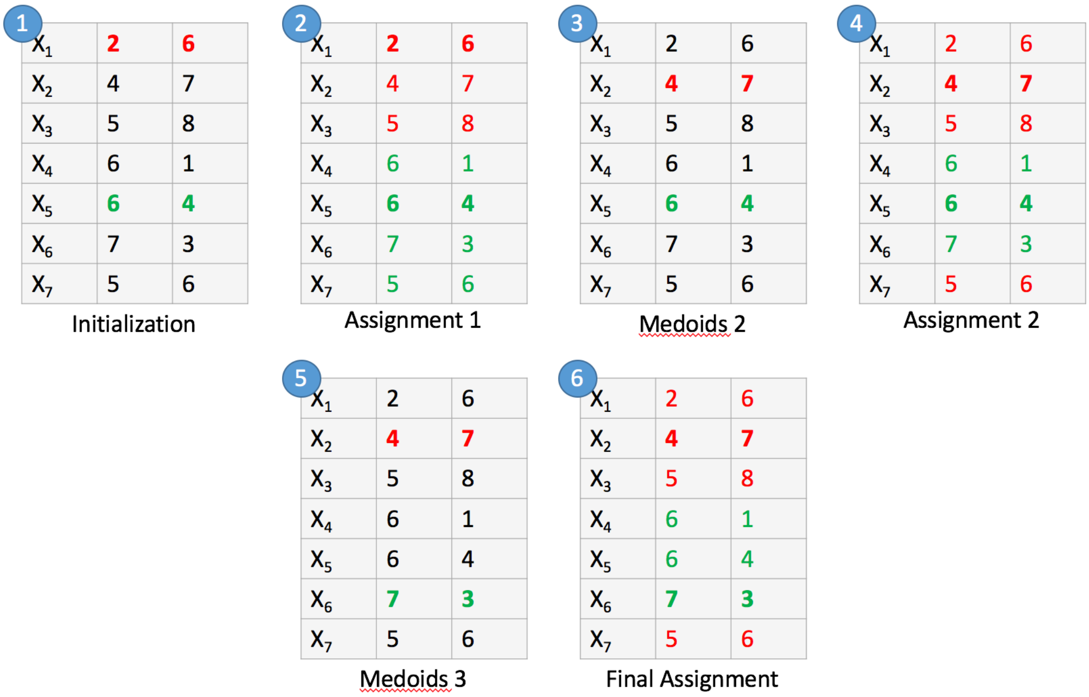
```
Initialize $c_1, \ldots, c_k$ by randomly selecting k elements from S
Loop:
    Assign all n samples to their closest $c_i$ and create k clusters $S_1, \ldots, S_k$
    For each cluster $S_i$ $(1 \le i \le k)$ define a new $c_i$:
            choose $c_i \in S_i$ whose distance to all other members in $S_i$ is the smallest
Until no change in $c_1, \ldots, c_k$
Return $c_1, \ldots, c_k$
```

A. Assume that the set S has 7 instances with 2 features as given in the table. Execute the k-medoids algorithm to cluster the set into two groups (i.e. k=2) when you initialize the execution with p_1 and p_5 as the initial centers. This means that instead of random initialization, $c_1=p_1$ and $c_2=p_5$ (hint: first plot the instances on a 2D Euclidean plane).

Solution:

Below are the stages of the process. Centroids are in boldface and clusters are colored red and green.

instance	x	y
p_1	**2**	**6**
p_2	4	7
p_3	5	8
p_4	6	1
p_5	**6**	**4**
p_6	7	3
p_7	5	6



1

X_1	**2**	**6**
X_2	4	7
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Initialization

2

X_1	**2**	**6**
X_2	4	7
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Assignment 1

3

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Medoids 2

4

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	**6**	**4**
X_6	7	3
X_7	5	6

Assignment 2

5

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	6	4
X_6	**7**	**3**
X_7	5	6

Medoids 3

6

X_1	2	6
X_2	**4**	**7**
X_3	5	8
X_4	6	1
X_5	6	4
X_6	**7**	**3**
X_7	5	6

Final Assignment

B. What is the main difference between k-means and k-medoids?
Solution:
The main difference is that in k-means the cluster representative is the geometric center of the cluster member points and doesn't have to be a data point. In k-medioids the algorithm insists on centroids that are actual datapoints. As a result the optimization criterion is also different, as detailed below.

C. Explain, using a formula, what function does k-means minimize?
Solution:
k-means minimizes, over all possible selection of any centroids μ_i , $i = 1 \dots k$, the total distances of all points to their cluster centroids – the geographic mean of the cluster member points:

$$\mathrm{J}_{\{\mathrm{means}\}} = \sum_{i=1}^{k} \sum_{x \in S_i} d(x, \mu_i)$$

D. How would you change the function from C so that it would fit the k-medoids algorithm?
Solution:
k-medioids minimizes, over all possible selection of datapoints centroids c_i , $i = 1 \dots k$, the total distances of all points to their cluster centroids:

$$\mathrm{J}_{\{\mathrm{meds}\}} = \sum_{i=1}^{k} \sum_{x \in S_i} d(x, c_i)$$

E. Assume that we execute both algorithms on the same set. Should we expect that the value of the function that k-medoid would minimize be smaller, greater or the same as the value of the function that k-means minimizes? Explain why!
Solution:
The value attained by the k-medioids process is likely to be larger than that attained by k-means. The inner sum, for k-means, is optimized by the selection of the centroid. In k medioids this selection is further constrained and therefore the value for using a medioid as a centroid is larger.

F. What serious problem can you find in the k-medoid algorithm as it is presented above in the pseudo code?
Solution:
The problem is that the algorithm doesn't necessarily converge.

Question 6 (5 parts)

Consider a diagnostic test that consists of measuring two quantitative features x1 and x2. We know, based on long-term measurement history, that the class-conditional distribution of values for these features are given by (D and H denote the two classed D = disease and H = healthy):

For the first feature

$$f(x_1|D) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$$

$$f(x_1|H) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(x-3)^2}{2}\right)$$

and, for the second feature

$$f(x_2|D) = \begin{cases} 1 & 0 \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

$$f(x_2|H) = \begin{cases} e & 1-\frac{1}{e} \le x_2 \le 1 \\ 0 & \text{Otherwise} \end{cases}$$

It is recommended (but not mandatory) that you schematically sketch these distributions to facilitate your answers.

A. Explain (in the general context) the difference between the ML (maximum likelihood) based classification and MAP (maximum aposteriori) based classification approaches.

Solution:

ML classifies by maximizing $P(x|A_i)$ while MAP classifies by maximizing $P(x|A_i)P(A_i)$. Thereby, MAP takes into account the class priors, when available.

B. What would the ML prediction be in the following (separate) two cases:

In general, the ML prediction is given by the maximum of p(x|D) and p(x|H).

1. We have measured, for a certain patient, the value $x_2 = 0.25$

   In this case $p(x2|D) = f(x2 = 0.25|D) = 1$ and p(x2|H) = f(x2=0.25|H) = 0 which means that the ML prediction is D.

2. We have measured, for a different patient, the value $x_1 = 1$

   In this case p(x1|D) = f(x1=1|D) = 1/2PI * exp(-1/2) and p(x1|H) = f(x1=1|H) = 1/2PI * exp(-2) which means that the ML prediction is D.

C. Assuming a prior P(H), clearly state the Naïve Bayes MAP formula for this diagnostic test.

Solution:

The naïve Bayes MAP approach will select the class C (either D or H) that maximizes the aposteriori probability of the observed data. It therefore maximizes:

P(C|x1,x2)= P(C)p(x1,x2|C) = P(C)p(x1|C)p(x2|C)

What is the minimal prior P(H) for which the MAP prediction, in Case B2 above, would be H?

Solution:
We want P(H)p(x1|H) >= P(D)p(x1|D)
So, setting p = P(H), we get:

$$p \geq \frac{f(x_1|D)}{f(x_1|D)+f(x_1|H)} = \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1-3)^2}{2}\right)} = $$

$$= \frac{e^{-\frac{1}{2}}}{e^{-\frac{1}{2}}+e^{-2}}$$

D. Assume that for B2 we also measured $x_2 = 0.95$.
In this case what is the minimal prior P(H) for which the MAP prediction would be H?

Solution:
Again, we want P(H)p(x1|H)p(x2|H) >= P(D)p(x1|D)p(x2|D)
So, in this case, setting p = P(H), we get:

$$p \geq \frac{f(x_1|D)f(x_2|D)}{f(x_1|D)f(x_2|D)+f(x_1|H)f(x_2|H)} =$$

$$= \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1^2}{2}\right)+\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(1-3)^2}{2}\right)e} = \frac{e^{-\frac{1}{2}}}{e^{-\frac{1}{2}}+e^{-1}}$$

E. For a third case we measured $x_1 = 8$ and $x_2 = 0.95(1-\frac{1}{e})$. Assume that P(H) = 0.9. What is the MAP prediction in this case? Do you think that this represents a bias or a shortcoming of the classification approach? How would you remedy this problem?

Solution:

$$p(H|x) = P(H)f(x_1|H)f(x_2|H)$$
$$p(D|x) = P(D)f(x_1|D)f(x_2|D)$$

As $f(x_2|H) = 0$ and all other terms are positive we will predict D.

This is not reasonable as the x_1 observation is strongly indicative of H and x_2 observation is very close to the boundary.
To remedy this, we could artificially modify $f(x_2|H)$ to be:

$$f(x_2|H) = \begin{cases} e & 1 - \frac{1}{e} \leq x_2 \leq 1 \\ \varepsilon & \text{Otherwise} \end{cases}$$

For some small $\varepsilon > 0$.