CSE-4502/5717 Big Data Analytics

Written Assignment 2

PCA; K-Nearest Neighbor Classifiers; SVM; Web Data Mining; Decision Trees;
Due: End of Day (11:59 pm, EST), March 26

Requirements:

1. The written assignment will be graded based on correctness, accuracy and clarity.
2. Please prepare your answers using a **word** document, and submit the final assignment in a **pdf** file through HuskyCT.
3. Please explicitly indicate the ids of the questions you are answering for ease of grading.
4. Even if your answer is not correct, you may still get certain partial marks based on your calculation/analysis process.
5. Please refer to the syllabus for the late submission policy.

1. (Total 10 marks) **PCA**

(a) (5 marks) Consider the following matrix, representing four sample points $X \in R^2$.

$$X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$$

We want to represent the data in only one dimension, so we turn to PCA. Please compute the unit-length principal component directions of X, and state which one the PCA algorithm would choose if you request just one principal component. (Suggested steps: center the data, calculate the sample covariance matrix, calculate the eigenvectors and eigenvalues, identify the principal component; please provide the calculation process).

(b) (5 marks) Given the following 3D input data, $X \in R^3$.

$$X = \begin{bmatrix} 1 & 1 & 9 \\ 2 & 4 & 6 \\ 3 & 7 & 4 \\ 4 & 11 & 4 \\ 5 & 9 & 2 \end{bmatrix}$$

Please compute the principal component which corresponds to the largest eigenvalue. (Suggested steps: center the data, calculate the sample covariance matrix, calculate the eigenvectors and eigenvalues, identify the principal component).

Note: sample covariance

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

(a)

Answers:

We center X, yielding

$$X = \begin{bmatrix} 1 & -1 \\ -1 & 1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix}$$

Then $X^T X = \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$

(Divide by 3 if you want the sample covariance matrix. But we don't care about the magnitude.)

Its unit-length eigenvectors are $\left[\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$ with eigenvalue 16 and $\left[\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}\right]^T$ with eigenvalue 4. The former eigenvector is chosen.

(Negated versions of these vectors also get full points.)

(b)

Answers:

Center the data:

$$\begin{bmatrix} -2 & -5.4 & 4 \\ -1 & -2.4 & 1 \\ 0 & 0.6 & -1 \\ 1 & 4.6 & -1 \\ 2 & 2.6 & -3 \end{bmatrix}$$

Find the sample covariance matrix $(((x - \mu)'(x - \mu))/(n - 1))$:

$$\begin{bmatrix} 2.5 & 5.75 & -4 \\ 5.75 & 15.8 & -9.25 \\ -4 & -9.25 & 7 \end{bmatrix}$$

Solve characteristic equation to obtain the eigenvalues and eigenvectors:

Eigen values:

$$\begin{bmatrix} 0.1298 \\ 1.2415 \\ 23.9287 \end{bmatrix}$$

Eigen vectors:

$$\begin{bmatrix} 0.9138 & -0.2617 & -0.3105 \\ -0.1035 & 0.5891 & -0.8014 \\ 0.3926 & 0.7645 & 0.5112 \end{bmatrix}$$

Select the principal component, i.e., the eigenvector corresponding to the largest eigenvalue:

$$\begin{bmatrix} -0.3105 \\ -0.8014 \\ 0.5112 \end{bmatrix}.$$

(Negated versions of the vector also get full points. Unit-length normalization not required for Q1(b))

2. (Total 10 marks) **K-Nearest Neighbors**
   Use the k-nearest neighbor to classify the unknown samples. Here we consider $K = 2$ and use Euclidean distance for the similarity.
   Please provide: (1) distance measure values calculated; (2) the final inference of the water type.

Sample ID	Ca+	Mg+	Na+	Cl-	Water Type
A	0.2	0.5	0.1	0.1	Glacier Water
B	0.4	0.3	0.4	0.3	Lake Water
C	0.3	0.4	0.6	0.3	Glacier Water
D	0.2	0.6	0.2	0.1	Glacier Water
E	0.5	0.5	0.1	0	Lake Water
F	0.3	0.3	0.4	0.4	Lake Water
G	0.3	0.3	0.3	0.2	?
H	0.1	0.5	0.2	0.2	?

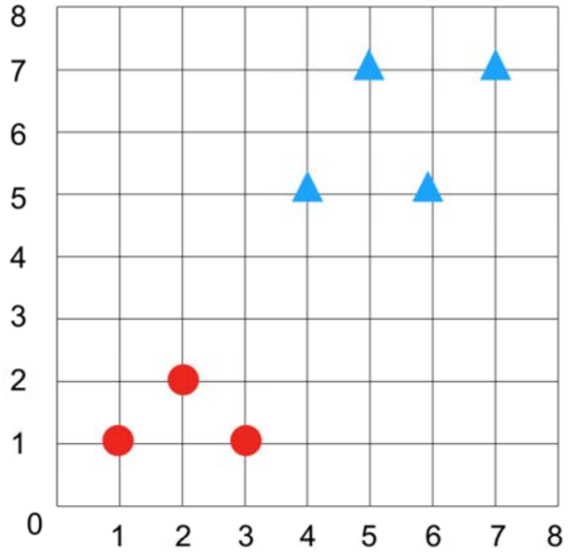
Answers

G: Lake Water
H: Glacier Water

To classify G, we have:
EuclideanDist(G, A) = 0.3162,
EuclideanDist(G, B) = 0.1732,
EuclideanDist(G, C) = 0.3317,
EuclideanDist(G, D) = 0.3464,
EuclideanDist(G, E) = 0.4000,
EuclideanDist(G, F) = 0.2236.
Therefore, we have the 2 nearest neighbors B and F, and we infer G is Lake Water.
EuclideanDist(H, A) = 0.1732,
EuclideanDist(H, B) = 0.4243,
EuclideanDist(H, C) = 0.4690,
EuclideanDist(H, D) = 0.1732,
EuclideanDist(H, E) = 0.4583,
EuclideanDist(H, F) = 0.4000.
Therefore, we have the 2 nearest neighbors A and D, and we infer H is Glacier Water.

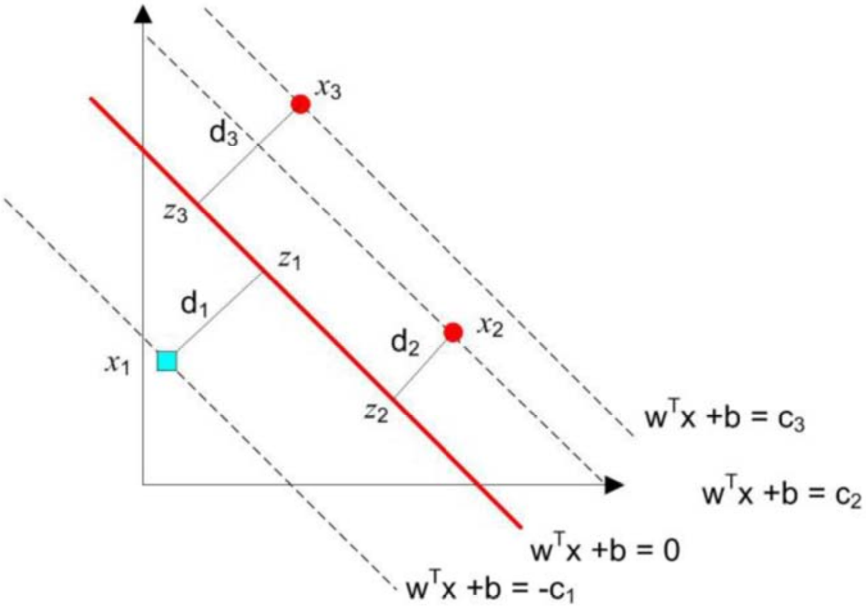
3. (Total 20 marks) **SVM**

(a) (12 marks) Consider the following training points. Circles are classified as positive examples with label + 1 and triangles are classified as negative examples with label −1.



Which points are the support vectors? If we add the sample point x = [5, 1] with label -1 (triangle) to the training set, which points are the support vectors? Recall that an SVM has the form of $y_i(w \cdot X_i + b) \geq 1$, where y_i's are the labels and X_i's are the training data points. What is the geometric relationship between the weight vector w and the decision boundary of SVM?

(b) (8 marks) Consider the 2-dimensional data shown in the following figure. There are three data points, two of them are classified as positive (red circles) and one is negative (blue square). Let the decision boundary of the linear classifier be $w^T x + b = 0$ (shown as a red line in the diagram).



The geometric margin of each data point is the **perpendicular** distance between each data point to the decision boundary (shown as d_1, d_2, and d_3, respectively). Derive an expression for the total geometric margin, $M = d_1 + d_2 + d_3$, as a function of w, c_1, c_2, and c_3. (Hint: you can use the virtual points z_1, z_2 and z_3 as helpers to derive the expression; recall the process in deriving the margin in the lecture and leverage some of the formulas for the expression)

Answers

(a)

(i) [2 2] and [4 5];

(ii) after adding, the support vectors are the points [3, 1], [4, 5] and [5, 1].

(iii) Perpendicular.

(b)

The three data points $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ satisfy the following equations:

$$\begin{aligned} w^T\mathbf{x}_1 + b &= -c_1 \\ w^T\mathbf{x}_2 + b &= c_2 \\ w^T\mathbf{x}_3 + b &= c_3 \end{aligned}$$

Similarly, the corresponding three points on the decision boundary satisfy the following equations:

$$\begin{aligned} w^T\mathbf{z}_1 + b &= 0 \\ w^T\mathbf{z}_2 + b &= 0 \\ w^T\mathbf{z}_3 + b &= 0 \end{aligned}$$

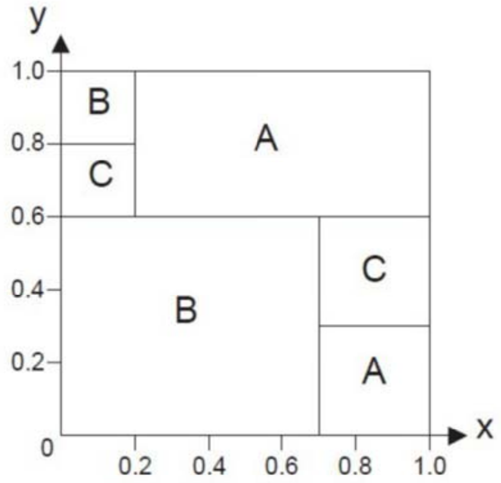
Subtracting the equations, we obtain:

$$\begin{aligned} w^T(\mathbf{z}_1 - \mathbf{x}_1) &= c_1 \implies \|w\|d_1 = c_1 \\ w^T(\mathbf{x}_2 - \mathbf{z}_2) &= c_2 \implies \|w\|d_2 = c_2 \\ w^T(\mathbf{x}_3 - \mathbf{z}_3) &= c_3 \implies \|w\|d_3 = c_3 \end{aligned}$$

Putting them together, we obtain

$$M = \frac{c_1 + c_2 + c_3}{\|w\|}.$$

4. (Total 15 marks) **Decision Tree**
   Consider a training set sampled uniformly from the the two-dimensional space shown in the following figure.



Assume that the training set size is large enough so that the probabilities can be calculated accurately based on the areas of the selected regions. The space is divided into three classes—A, B, and C. For example, we can say probability of class A is $p(A) = 0.3 \times 0.3 + 0.8 \times 0.4 = 0.41$.

In this exercise, you will build a decision tree, using the concepts of **information gain** and **entropy**, from the training set. Note that it is slightly different from the examples taught in lecture, as the dataset only has two attributes, X and Y. You might need to repeatedly use some split points in X and Y to finally classify the labels of data points. Each node in the decision tree only checks one attribute (either X or Y).

(1) (4 marks) Given above, please calculate the overall entropy for the data (Info(T)).

(2) (9 marks) Compare the entropy when the data is split at x <= 0.2, x <= 0.7, and y <= 0.6.

(3) (2 marks) Based on your answer in part (b), which attribute split condition do you think should be used as the root of the decision tree.

Answers:

(1) For overall data, p(A) = 0.32 + 0.09 = 0.41, p(B) = 0.42 + 0.04 = 0.46, and p(C) = 0.04 + 0.09 = 0.13. Therefore, the overall entropy is

$$-0.41 \log 2\ 0.41 - 0.46 \log 2\ 0.46 - 0.13 \log 2\ 0.13 = 1.4254.$$

(2) We check each split point:

i. Split at $x = 0.2$:
For the child node $x \leq 0.2$, $p(A) = 0$, $p(B) = 0.8$, and $p(C) = 0.2$. Its entropy is $-0.8 \log_2 0.8 - 0.2 \log_2 0.2 = 0.7219$. For the child node $x > 0.2$, $p(A) = 0.41/0.80 = 0.5125$, $p(B) = 0.3/0.8 = 0.3750$, and $p(C) = 0.09/0.80 = 0.1125$. Its entropy is $-0.5125 \log_2 0.5125 - 0.375 \log_2 0.375 - 0.1125 \log_2 0.1125 = 1.3795$. Therefore, the average entropy for the children is $0.2 \times 0.7219 + 0.8 \times 1.3795 = 1.2480$.

ii. Split at $x = 0.7$:
For the child node $x \leq 0.7$, $p(A) = 0.2/0.7 = 0.2857$, $p(B) = 0.46/0.7 = 0.6571$, and $p(C) = 0.04/0.7 = 0.0571$. Its entropy is $-0.2857 \log_2 0.2857 - 0.6571 \log_2 0.6571 - 0.0571 \log_2 0.0571 = 1.1503$. For the child node $x > 0.7$, $p(A) = 0.7$, $p(B) = 0$, and $p(C) = 0.3$. Its entropy is $-0.7 \log_2 0.7 - 0.3 \log_2 0.3 = 0.8813$. Therefore the average entropy for the children is $0.7 \times 1.1503 + 0.3 \times 0.8813 = 1.0696$.

iii. Split at $y = 0.6$.
For the child node $y \leq 0.6$, $p(A) = 0.09/0.6 = 0.15$, $p(B) = 0.42/0.6 = 0.7$, and $p(C) = 0.09/0.6 = 0.15$. Its entropy is $-0.15 \log_2 0.15 - 0.7 \log_2 0.7 - 0.15 \log_2 0.15 = 1.1813$. For the child node $y > 0.6$, $p(A) = 0.32/0.4 = 0.8$ and $p(B) = p(C) = 0.04/0.4 = 0.1$. Its entropy is $-0.8 \log_2 0.8 - 0.1 \log_2 0.1 - 0.1 \log_2 0.1 = 0.9219$. Therefore, the average entropy for the children is $0.6 \times 1.1813 + 0.4 \times 0.9219 = 1.0776$.

(3) Comparing their entropy values, the split at x = 0.7 has the highest gain.

5. (Total 25 marks) **Decision Tree**

The following shows a history of customers with their incomes, ages and an attribute called "Have_iPhone" indicating whether they have an iPhone. We also indicate whether they will buy an iPad or not in the last column.

ID	Income	Age	Have_iPhone	Buy_iPad
1	high	young	Yes	Yes
2	high	old	Yes	Yes
3	medium	young	No	Yes
4	high	old	No	Yes
5	medium	young	No	No
6	medium	young	No	No
7	medium	old	No	No
8	medium	old	No	No

We want to train a **CART decision tree** classifier to predict whether a new customer will buy an iPad or

not. We define the value of attribute Buy_iPad is the label of a record.

(i) (20 marks) Please find a CART decision tree according to the above example. In the decision tree, whenever we process a node containing at most **3** records, we **stop** to process this node for splitting.

(ii) (5 marks) Consider a new young customer whose income is medium and he has an iPhone. Please predict whether this new customer will buy an iPad or not.

Answers

(i) Info(T) = $1 - 0.5^2 - 0.5^2 = 0.5$

For attribute Income

Info(T) = $1 - 0.5^2 - 0.5^2 = 0.5$

Info(T_{high})= $1-1^2 - 0^2 = 0$

Info(T_{medium})= $1 - (1/5)^2 - (4/5)^2 = 0.32$

Info(Income, T) = 3/8 Info(T_{high}) + 5/8 Info(T_{medium}) = 0.2

Gain(Income, T)= Info(T)-Info(Income, T)=0.3

For attribute Age,

Info(T_{young})= $1 - 0.5^2 - 0.5^2 = 0.5$

Info(T_{old})=$1- 0.5^2 - 0.5^2$ =0.5

Info(Age, T)= 1/2Info(T_{young}) +1/2Info(T_{old}) = 0.5

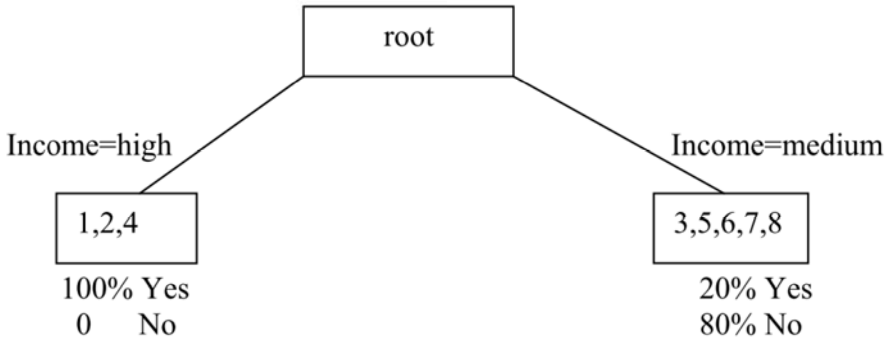
Gain(Age, T)= Info(T)-Info(Age, T)=0

For attribute Have-iPhone,

Info(T_{yes})=$1 - 1^2 - 0^2$=0

Info(T_{no})= 1-$(1/3)^2$ – $(2/3)^2$ = 0.4444
Info(Have-iPhone, T)= 1/4 Info(T_{yes}) + 3/4 Info(T_{no}) = 0.3333
Gain(Have-iPhone, T)= Info(T)-Info(Have-iPhone, T)= 0.1667

We choose attribute Income for Splitting:

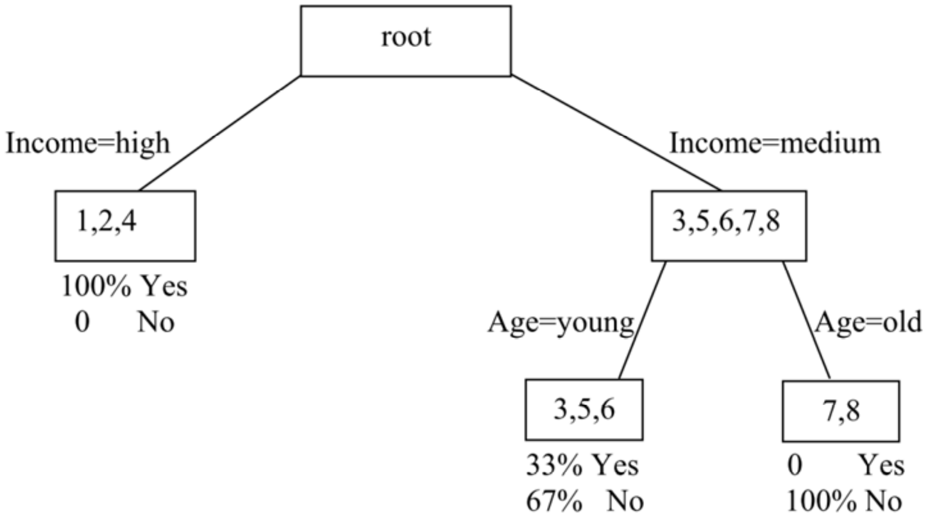


Consider the node for Income = medium:
Info(T) = 1-(1/5)2-(4/5)2 = 0.32

For attribute Age,
Info(T_{young})= $1-(1/3)^2-(2/3)^2=0.4444$
Info(T_{old})=1− 1^2 – 0^2 = 0
Info(Age, T)= 3/5Info(T_{young}) +2/5Info(T_{old}) = 0.26664
Gain(Age, T)= Info(T)-Info(Age, T)=0.05336

For attribute Have-iPhone,
Info(T_{yes})= undefined
Info(T_{no})= 1-$(1/5)^2$ – $(4/5)^2$ = 0.32
Info(Have-iPhone, T)= 0 × Info(T_{yes}) + 1 × Info(T_{no}) = 0.32
Gain(Have-iPhone, T)= Info(T)-Info(Have-iPhone, T)= 0

We choose attribute Age for Splitting:



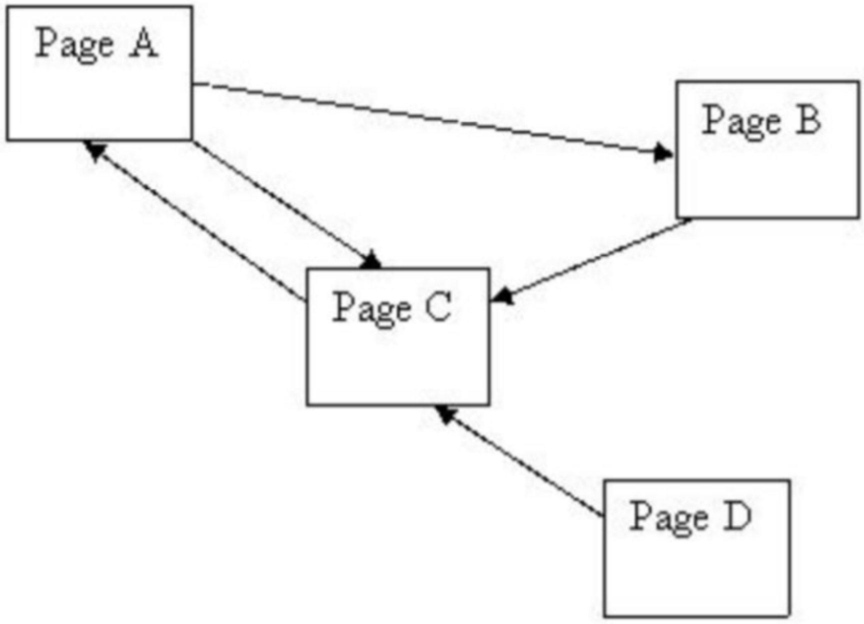
(ii) It is likely that he will not buy an iPad.

6. (Total 20 marks) **Page Rank**

   Given the 4 web pages and their mutual links, find their **Page Ranks** according to our lecture notes. Suppose we initialize all ranks as 1.

   (1) (Total 10 marks) Please provide the convergence ranks based on $r = Mr$. Note that for the stochastic matrix, please use the alphabetical order (from A to D) to list the web pages in the matrix. You can write a program to calculate but you are not required to submit the program. You can show me at which iteration the rank starts to converge.

   (2) (Total 10 marks) Please provide the convergence ranks based on $r = 0.8Mr + c$, where $c = [0.2 \;\; 0.2 \;\; 0.2 \;\; 0.2]^T$. You can write a program to calculate but you are not required to submit the program. You can show me at which iteration the rank starts to converge.



Answer

(1) Stochastic matrix is given by

$$\begin{matrix} 0 & 0 & 1 & 0 \\ 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{matrix}$$

Based on $r = Mr$, we initialize all the page rank as $[1 \;\; 1 \;\; 1 \;\; 1]^T$. The final convergence rank (after 32 iterations) is

$$\begin{bmatrix} 1.6 \\ 0.8 \\ 1.6 \\ 0 \end{bmatrix}$$

(2) Based on r = Mr + c, we finally get (after around 20 iterations)

$$\begin{bmatrix}1.4528\\0.7811\\1.5660\\0.2000\end{bmatrix}$$