FL2021.E81.CSE.514A.01 Exam 2

Anh Le

TOTAL POINTS

87 / 85

QUESTION 1

True or False 16 pts

1.1 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

- 2 pts Left blank

1.2 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

1.3 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

1.4 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

- 2 pts Left blank

1.5 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

1.6 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

1.7 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

1.8 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

QUESTION 2

Supervised or Unsupervised 10 pts

2.1 2 / 2

✓ **- 0 pts Correct**

2.2 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

2.3 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

2.4 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

2.5 2 / 2

✓ **- 0 pts Correct**

- 2 pts Incorrect

QUESTION 3

Clustering 9 pts

3.1 Single linkage 3 / 3

✓ **- 0 pts Correct**

- 1.5 pts Incorrect/missing distance

- 2 pts Mostly incorrect

- 1 pts Incorrect cluster

3.2 Complete linkage 1 / 3

- 0 pts Correct

✓ **- 2 pts Mostly incorrect**

- 1.5 pts Incorrect/missing distance

3.3 Average linkage 3 / 3

✓ **- 0 pts Correct**

- 2 pts Mostly incorrect

- 1.5 pts Incorrect/missing distance

3.4 Extra credit 0 / 0

✓ **- 0 pts No extra credit**

+ 1 pts EC for single linkage

+ 1 pts EC for complete linkage

+ 1 pts EC for average linkage

QUESTION 4

SVM kernels 5 pts

4.1 Polynomial 2 / 2

✓ **- 0 pts Correct**

4.2 RBF 3 / 3

✓ **- 0 pts Correct**

- 2 pts Mostly incorrect

- 1 pts Partly incorrect

QUESTION 5

Naive Bayes Classifier 18 pts

5.1 Citric Acid 4 / 4

✓ **- 0 pts Correct**

- 2 pts Incorrect counts

- 1 pts Incorrect counts

5.2 Residual Sugar 4 / 4

✓ **- 0 pts Correct**

5.3 Alcohol 4 / 4

✓ **- 0 pts Correct**

- 1 pts Incorrect counts

5.4 Color 3 / 3

✓ **- 0 pts Correct**

- 1 pts Incorrect counts

5.5 Quality 3 / 3

✓ **- 0 pts Correct**

- 1 pts Incorrect count

QUESTION 6

Classify 7 pts

6.1 With Q5 classifier 7 / 7

✓ **- 0 pts Correct**

- 1 pts An incorrect probability in high

- 1 pts Missing final conclusion

- 1 pts An incorrect probability in low

- 2 pts Missing probabilities for priors (Quality)

- 1 pts An incorrect probability in medium

6.2 Extra credit 1 / 0

- 0 pts No EC

✓ **+ 1 pts Extra Credit**

QUESTION 7

Association Rules Mining 10 pts

7.1 Itemsets 5 / 5

✓ **- 0 pts Correct**

- 1 pts One or two missing/incorrect

- 2 pts Three missing/incorrect

- 3 pts Four or five missing/incorrect

- 4 pts Missing most

7.2 Rules 5 / 5

✓ **- 0 pts Correct**

- 2 pts Incomplete due to incorrect part a

- 1 pts Missing one

- 2 pts Missing several

- 3 pts Incorrect

- 4 pts Incorrect

- 5 pts Left blank / Not rules

7.3 Extra credit 1 / 0

✓ **+ 1 pts Found rules with lift >1**

+ 2 pts Found rules with greatest lift

- 0 pts No extra credit

+ 0.5 pts Was looking for actual rules, not how to

pick them

QUESTION 8

8 Soft vs Maximal Margin Classifier 3 / 3

✓ **- 0 pts** **Correct**

- 1 pts Partially incomplete/incorrect

- 3 pts Incorrect/left blank

- 2 pts Mostly incomplete/incorrect

QUESTION 9

9 SVM vs Clustering dimensions 4 / 4

✓ **- 0 pts** **Correct**

- 2 pts Incorrect/incomplete

- 1 pts Incomplete

- 3 pts Mostly incorrect/incomplete

QUESTION 10

Data preprocessing 3 pts

10.1 Two steps 3 / 3

✓ **- 0 pts** **Correct**

- 1 pts Redundant/addresses the same problem

10.2 Extra credit 2 / 0

- 0 pts No extra credit

✓ **+ 2 pts** **Extra credit**

+ 1 pts Process, no value

CSE514 Fall 2021 Exam 2
Q&A Packet

Name: Anh Le **Student ID:** 488493

When you are finished, please check that any exam answers are written in the indicated areas of the exam before you submit

This exam is worth a total of 85 points
This exam includes extra credit opportunities, up to 8 EC points

Q1 (16 pts): True or False?

Place a single check mark in each row to indicate if the following statements are true or false.

	True	False
The hyperplane boundary from an SVM must perfectly separate samples of different classes		✓
To calculate a dot product between samples in a higher dimension, an SVM first projects the data points and then applies a kernel		✓
Naïve Bayes Classifiers assume all features are conditionally independent of each other	✓	
The predictions from a Naïve Bayes Classifier can change based on the trainer's prior beliefs	✓	
The first merge of Agglomerative Clustering is the same no matter if you implement single-linkage, complete-linkage, or average-linkage	✓	
Association Rules Mining is a supervised learning method		✓
Principal Component Analysis is a supervised learning method		✓
Feature selection can be either supervised or unsupervised	✓	

Q2 (10 pts): What kind of learning is this?

Place a single check mark in each row to indicate if the problem is supervised or unsupervised

	Supervised	Unsupervised
Predict tomorrow's rainfall by comparing today's cloud cover to past records of weather conditions before rainy days	✓	
Find sets of products that should be grouped together on shelf displays because they are often purchased together		✓
Identify potential security threats by scanning network traffic for outlier behavior		✓
Use the PCA method to extract new features that maintain at least 90% of the original data's variance		✓
Reduce data size by dropping features that correlate with the response value with correlation coefficient $\| r \| < 0.3$	✓	

Q3 (9 pts + 3EC): Single vs. Complete vs. Average-linkage clustering of Dataset 3

Refer to the clusters and cluster distances on pg3 of the reference packet

Q3a (3pts) Which clusters would you merge if implementing Single-Linkage?
Answer by giving the two cluster numbers, and the distance between them.

Answer here:

Cluster 1 & Cluster 3

Distance: 1.09

Q3b (3pts) Which clusters would you merge if implementing Complete-Linkage?
Answer by giving the two cluster numbers, and the distance between them.

Answer here:

Cluster 1 & Cluster 3

Distance: 2.30

Q3c (3pts) Which clusters would you merge if implementing Average-Linkage? Answer by giving the two cluster numbers, and the distance between them.

Answer here:
Cluster 1 & Cluster 3
Distance: 1.75

Q3d (+3EC) For each of the three linkage approaches, give one reason why (or example scenario where) it is better at creating good clusters than the others.

Single-linkage:

Complete-linkage:

Average-linkage:

Q4 (5 pts): Calculate the dot product of the following two samples in higher dimensions

Sample 1: [2, 1] Sample 2: [5, 2]

Q4a (2pts): Polynomial kernel, $d = 2, r = 1$

Answer here:
$(2 \times 5 + 1 \times 2 + 1)^2$
$= 169$

Q4b (3pts): RBF kernel, $\gamma = 0.1$

Answer here:
$e^{-0.1((2-5)^2 + (1-2)^2)}$
$= 0.368$

Q5 (18 pts): Train a Naïve Bayes Classifier on Dataset 2 for predicting Wine Quality by completing the following tables.

Refer to pg2 of the reference packet and apply a pseudo-count value of 1.

Citric Acid (4pts):

	P(x | Low Quality)	P(x | Medium Quality)	P(x | High Quality)
Low	3 + 1 = 4; 4/8 = 50%	1 + 1 = 2; 2/5 = 40%	0+1 = 1 ; 1/6
Medium	1 + 1 = 2; 2/8 = 25%	1 + 1 = 2; 2/5 = 40%	2+1 = 3; 3/6 = 50%
High	1 + 1 = 2; 2/8 = 25%	0 + 1 = 1; 1/5 = 20%	1+1 = 2; 2/6 = 1/3
Total	8	5	3+3 = 6

Residual Sugar (4pts):

	P(x | Low Quality)	P(x | Medium Quality)	P(x | High Quality)
Low	1+1 = 2 ; 2/8 = 25%	1 + 1 = 2; 2/5 = 40%	3 + 1 = 4; 4/6 = 66.7%
Medium	3+1 = 4; 4/8 = 50%	1 + 1 = 2; 2/5 = 40%	0 + 1 = 1; 1/6 = 16.7%
High	1+1 = 2; 2/8 = 50%	0 + 1 = 1; 1/5 = 20%	0 + 1 = 1; 1/6 = 16.7%
Total	8	5	6

Alcohol (4pts):

	P(x | Low Quality)	P(x | Medium Quality)	P(x | High Quality)
Low	3 + 1 = 4; 4/8 = 50%	0+1 = 1 ; 1/5 = 20%	0 + 1 = 1; 1/6 = 16.7%
Medium	1 + 1 = 2; 2/8 = 25%	2+1 = 3 ; 3/5 = 60%	0 + 1 = 1; 1/6 = 16.7%
High	1 + 1 = 2; 2/8 = 25%	0+1 = 1; 1/5 = 20%	3 + 1 = 4; 4/6 = 66.7%
Total	8	5	6

Color (3pts):

	P(x | Low Quality)	P(x | Medium Quality)	P(x | High Quality)
Red	3 + 1 = 4; 4/7 = 57.1%	1 + 1 = 2; 2/4 = 50%	1+1 = 2 ; 2/5 = 40%
White	2 + 1 = 3; 3/7 = 42.9%	1 + 1 = 2; 2/4 = 50%	2+1 = 3 ; 3/5 = 60%
Total	7	4	5

Quality (3pts):

P(x | Low Quality)	5/10 = 50%
P(x | Medium Quality)	2/10 = 20%
P(x | High Quality)	3/10 = 30%
Total	1

Q6 (7pts + 1EC) Classify the following sample:

ID	Citric Acid	Residual Sugar	Alcohol	Color	Quality
11	Medium	High	Low	White	-

Q6a (7pts) Use the Naïve Bayes Classifier from Q7 to classify the sample:
Answer by calculating the proportional probabilities of each quality value

Answer here:

$P(11 \mid \text{Low Quality}) = \frac{1}{2} \times \frac{2}{8} \times \frac{2}{8} \times \frac{4}{8} \times \frac{3}{7} = 6.7 \times 10^{-3}$

$P(11 \mid \text{Medium Quality}) = \frac{1}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{2} = 1.6 \times 10^{-3}$

$P(11 \mid \text{High Quality}) = \frac{3}{10} \times \frac{1}{2} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} = 2.5 \times 10^{-3}$

→ Sample 11 is Low quality

Q6b (+1EC) A wine expert tells you that in his experience:

Probability (low-quality wine) = 25%
Probability (medium-quality wine) = 70%
Probability (high-quality wine) = 5%

Re-classify the sample based on his beliefs.

Answer here:

$P(11 \mid \text{Low Quality}) = \frac{1}{4} \times \frac{2}{8} \times \frac{2}{8} \times \frac{4}{8} \times \frac{3}{7} = 3.35 \times 10^{-3}$

$P(11 \mid \text{Medium Quality}) = \frac{7}{10} \times \frac{2}{5} \times \frac{1}{5} \times \frac{1}{5} \times \frac{1}{2} = 5.6 \times 10^{-3}$

$P(11 \mid \text{High Quality}) = \frac{5}{100} \times \frac{1}{2} \times \frac{1}{6} \times \frac{1}{6} \times \frac{3}{5} = 4.17 \times 10^{-4}$

→ Sample 11 is Medium Quality

Q7 (10 pts + 2EC) Association Rules Mining from Dataset 1.

Refer to pg2 of the reference packet.

Q7a (5 pts): Use the Apriori Principle to find all itemsets with Support >= 0.5

Support ({Bread}) = 9/10
Support ({Milk}) = 7/10
Support ({Sugar}) = 4/10
Support ({Eggs}) = 6/10
Support ({Shampoo}) = 6/10
Support ({Conditioner}) = 2/10

Answer here:
Support ({Bread, Milk}) = 7/10
Support ({Bread, Eggs}) = 5/10
Support ({Bread, Shampoo}) = 5/10
Support ({Milk, Shampoo}) = 5/10
Support ({Bread, Milk, Shampoo}) = 5/10
Support ({Bread}) = 9/10
Support ({Milk}) = 7/10
Support ({Eggs}) = 6/10
Support ({Shampoo}) = 6/10

Q7b (5 pts): Use the Apriori Principle to find all rules with Confidence >= 0.8
Only create rules for itemsets that passed the Support threshold.

Support ({Shampoo, Eggs}) = 3/10

Answer here:
Confidence ({Milk} → {Bread}) = 1
Confidence ({Eggs} → {Bread}) = 5/6 = 0.83
Confidence ({Shampoo} → {Bread}) = 5/6 = 0.83
Confidence ({Shampoo} → {Milk}) = 5/6 = 0.83
Confidence ({Shampoo} → {Bread, Milk}) = 5/6 = 0.83
Confidence ({Bread, Shampoo} → {Milk}) = 1
Confidence ({Milk, Shampoo} → {Bread}) = 1

Q7c (+2 EC): Which Association Rules are the most informative?

Answer here:
Lift ({Milk} → {Bread}) = $\frac{1}{0.9} > 1$
Lift ({Shampoo} → {Milk}) = $\frac{0.83}{0.7} > 1$
Lift ({Shampoo} → {Bread, Milk}) = $\frac{0.83}{0.7} > 1$

Q8 (3pts): Explain what makes a Soft Margin Classifier different from a general Maximal Margin Classifier

Answer here:
Maximal Margin Classifier defines a "hard" margin that must perfectly separate samples of different classes where misclassification is not allowed. Meanwhile, a Soft Margin Classifier allows misclassification to get some samples wrong in exchange for a bigger margin on the dataset

Q9 (4pts): Data in higher dimensions tend to be sparser than data in lower dimensions, with each sample further away from all other samples simply because there are more features. More features means more chances to find differences and/or add distance.

Explain why SVM works better by projecting data into higher dimensions while clustering works better in low dimensions.

Answer here:
Since SVM is a Soft Margin Classifier, misclassification in higher dimensions is inevitable. Moreover, using an SVM for higher dimension is better because the data is more spread, sparser, which makes it easier to find a bigger margin with misclassification allowed. Meanwhile, clustering helps separate samples into groups so if the datapoints are closer to each other, grouping might be easier & faster

Q10 (3pts + 2EC): You are given Dataset 4 on page 4 of the reference packet. In this dataset, '?' is used as a placeholder for missing values.

Q10a (3pts) Suggest two preprocessing steps before feeding this data into a logistic regression model.

Answer here:

1, Column 4 and 5 are 100% same so I would drop 1 of them

2, Column 13 has the lowest variance among all features (only 2s) so I might consider dropping it as well

Q10b (+2EC) Impute the missing value in the first column. Explain your process.

Answer here:

I would impute the missing value using 1-Nearest Neighbor method since there are only 24 samples in this dataset Looking through the dataset, I believe "?" can be replaced by 'a'