Final exam review

C S 488/508

Venue & Time

- December 7 (Tuesday)
- 1:00pm-3:00pm
- M02: Come to class
- M70: Take it online (Lockdown browser + camera)

How to take the exam

- Calculator is allowed
- Cell phone is NOT allowed
- 1 page cheat sheet (letter-size), allowing one-side or both-sided, hand-written or printed
- Plenty of blank paper
- Pencil/pen
- 2hrs (online with 15 minutes extra to accommodate technical issues)

Question types

- Short answer questions
- Multi-choice questions
- True/False questions
- Programming questions (NO)

Scope (1)

- ~~Introduction~~
- Data
  - Distance calculation
  - ~~Sampling~~
  - ~~Exploration (I will not ask you to write program; but you may be given some plots and I will ask you questions related to those plots)~~
- Classification
  - ~~Decision trees (Gini index, entropy, information gain)~~
  - KNN
  - Bayesian Classifier,
  - Logistic Regression (only T/F or multi-choice questions)
  - SVM
- ~~Classification issues & concepts~~
  - ~~Non-balanced datasets~~
  - ~~Performance measurements: Contingence matrix, accuracy, F1, Precision, recall, ROC, AUC~~

Scope (2)

- Clustering
  - K-means
  - Hierarchical
  - MST-based
  - DBSCAN
  - Spectral clustering (only T/F or multi-choice questions)
- Association rules
  - Concepts
  - Apriori algorithm, FP-Growth (graduate)
  - Sequential patterns – GSP algorithm
- Anomaly detection
  - ~~Statistical based~~
  - Approximaty based (distance-based or density based)
- Avoid false discoveries
  - Concepts (only T/F or multi-choice questions)

Exercise 1 - Naïve Bayesian classifier

Consider the following data set

a) Estimate the conditional probabilities for P(A|+), P(B|+), P(C|+), P(A|−), P(B|−), and P(C|−).

b) Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A = 0, B = 1, C = 0) using the naive Bayes approach.

c) Estimate the conditional probabilities using the m-estimate approach, with p = 1/2 and m = 4.

d) Repeat part (b) using the conditional probabilities given in part (c).

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	−
3	0	1	1	−
4	0	1	1	−
5	0	0	1	+
6	1	0	1	+
7	1	0	1	−
8	1	0	1	−
9	1	1	1	+
10	1	0	1	+

Steps to solve 1.b

$P(A=0|+) = 2/5$

$P(+|A=0,B=1,C=0)$??? $P(-|A=0,B=1,C=0)$

$P(+|A=0,B=1,C=0) = P(A=0,B=1,C=0|+)\ P(+)/P(A=0,B=1,C=0)$

proportional to $P(A=0,B=1,C=0|+)\ P(+)$

$P(-|A=0,B=1,C=0) = P(A=0,B=1,C=0|-)\ P(-)/P(A=0,B=1,C=0)$

$\Longrightarrow P(A=0,B=1,C=0|+)\ P(+)$??? $P(A=0,B=1,C=0|-)\ P(-)$

$P(A=0,B=1,C=0|+)\ P(+) = P(A=0|+)\ P(B=1|+)\ P(C=0|+)\ P(+) = 2/5 \,.\, 1/5 \,.\, 1/5 \,.\, ½$

$P(A=0,B=1,C=0|-)\ P(-) = P(A=0|-)\ P(B=1|-)\ P(C=0|-)\ P(-) \quad = 3/5 \,.\, 2/5 \,.\, 0/5 \,.\, ½$

➔ $P(A=0,B=1,C=0|+)\ P(+) > P(A=0,B=1,C=0|-)\ P(-)$

Predict this instance to be positive.

M-estimate

Problem 1.c):

$P(A=0|+) = (2+4*1/2) / (5+4) = 4/9$

$P(B=1|+) = (1+4*1/2) / (5+4) = 3/9$

…

Problem 1.d)

$P(+|A=0,B=1,C=0)$??? $P(-|A=0,B=1,C=0)$

$P(+|A=0,B=1,C=0) = P(A=0,B=1,C=0|+)\ P(+)/P(A=0,B=1,C=0)$ proportional to $P(A=0,B=1,C=0|+)\ P(+)$

$P(-|A=0,B=1,C=0) = P(A=0,B=1,C=0|-)\ P(-)/P(A=0,B=1,C=0)$

==> $P(A=0,B=1,C=0|+)\ P(+)$??? $P(A=0,B=1,C=0|-)\ P(-)$

$P(A=0,B=1,C=0|+)\ P(+) = P(A=0|+)\ P(B=1|+)\ P(C=0|+)\ P(+) = 4/9 . 3/9 . 3/9 . ½$

$P(A=0,B=1,C=0|-)\ P(-) = P(A=0|-)\ P(B=1|-)\ P(C=0|-)\ P(-) = 5/9 . 4/9 . 2/9 . ½$

➔ $P(A=0,B=1,C=0|+)\ P(+) < P(A=0,B=1,C=0|-)\ P(-)$

Predict this instance to be negative.

Exercise 2 - SVM

x_{i1}	x_{i2}	y_i	λ_i
0.4	0.5	1	100
0.5	0.6	-1	100
0.9	0.4	-1	0
0.7	0.9	-1	0
0.17	0.05	1	0
0.4	0.35	1	0
0.9	0.8	-1	0
0.2	0	1	0

- $y_z = sign(\mathbf{w}^\intercal \mathbf{z} + b) = sign((\sum_{i=1}^{N} \lambda_i y_i \mathbf{x}_i^\intercal)\mathbf{z} + b)$
- if $y_z = 1$, the test instance is classified as positive class
- if $y_z = -1$, the test instance is classified as negative class

- Solve λ using quadratic programming packages
- $\mathbf{w}^\intercal = (w_1, w_2)$
  - $w_1 = \sum_{i=1}^{2} \lambda_i y_i x_{i1} = 100 * 1 * 0.4 + 100 * (-1) * 0.5 = -10$
  - $w_2 = \sum_{i=1}^{2} \lambda_i y_i x_{i2} = 100 * 1 * 0.5 + 100 * (-1) * 0.6 = -10$
- $b = 1 - \mathbf{w}^\intercal \mathbf{x}_1 = 1 - ((-10) * 0.4 + (-10) * (0.5)) = 10$

Questions:

a) What are the support vectors

b) What are w1, w2?

c) Give a new instance (0.1, 0.3), what's your prediction?

Steps to solve 2.c)

$w^T z + b = (-10,-10) \binom{0.1}{0.3} + 10 = -4 + 10 = 6$

Sign is positive

Predict this instance to be positive.

Exercise 3- KNN

Id	A	B	Class label
1	0.1	0.2	1
2	0.2	0.3	-1
3	0.4	0.3	-1
4	0.8	0.9	1
5	0.7	0.6	1

Questions: Given a new instance (0.3, 0.2) what will be the predicted class labels using KNN classification algorithm if

a) K=1
b) K=3

Rough solution steps

p = (0.3,0.2), I will use Manhattan distance

p1= (0.1,0.2), dist(p,p1) = 0.2

p2= (0.2,0.3), dist(p,p2) = 0.2

p3= (0.4,0.3), dist(p,p3) = 0.2

p4= (0.8,0.9), dist(p,p4) = 1.2

p5= (0.7,0.6), dist(p,p5) = 0.8

K=1, there is a tie among p1, p2, p3. I randomly choose one, p1, then, I predict p's class label to be the same as p1, POSITIVE.

K=3, 3NN={p1,p2,p3}, we choose the majority class label, which is negative.

Exercise 4 – Clustering (k-means, DBSCAN)

Suppose that the data mining task is to cluster points (with (x, y) representing location) into three clusters, where the points are.

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$, $B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$, $C_1(1,2)$, $C_2(4,9)$

The distance function is Euclidean distance. Suppose initially we assign A_1, B_1, and C_1 as the center of each cluster, respectively. Use the *k*-means algorithm to show only

(a) The three cluster centers after the first round of execution.

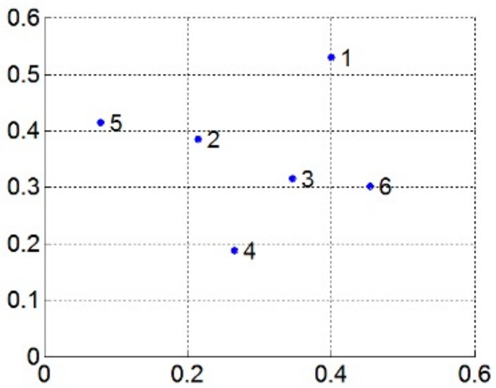
(b) The final three clusters.

(c) Apply the DBSCAN algorithm on the above data points (with parameters Eps = 1.5 and minPts = 3), indicate the final clusters.

Exercise 5 - Hierarchical, MST-based clustering

Given the dataset below

- a) show the clustering steps using min/max strategy
- b) construct MST step by step



Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Exercise 6 – association rules

Consider the following dataset

a) Compute the support for itemsets {*e*}, {*b*, *d*}, and {*b*, *d*, *e*}

b) Use the results in part (*a*) to compute the confidence for the association rules {*b*, *d*} → {e} and {*e*} → {*b*, *d*}. Is confidence a symmetric measure?

Transaction ID	Items Bought
0001	$\{a, d, e\}$
0024	$\{a, b, c, e\}$
0012	$\{a, b, d, e\}$
0031	$\{a, c, d, e\}$
0015	$\{b, c, e\}$
0022	$\{b, d, e\}$
0029	$\{c, d\}$
0040	$\{a, b, c\}$
0033	$\{a, d, e\}$
0038	$\{a, b, e\}$

Exercise 7 – sequential patterns

Consider the following dataset and min_sup=50%,

a) What is the support of sequential pattern <{1}{2}>?

b) Find the frequent sequential patterns in the form of <{x}{y}> where x and y represent one item.

c) If F2={<{1,2}>, <{2,3}>, <{2,4}>, <{3}{5}>, <{1}{2}>, <{2}{2}>}, what will be C3 (candidate length-3 patterns)?

Object	Timestamp	Events
A	1	1,2,4
A	2	2,3
A	3	5
B	1	1,2
B	2	2,3,4
C	1	1, 2
C	2	2,3,4
C	3	2,4,5
D	1	2
D	2	3, 4
D	3	4, 5
E	1	1, 3
E	2	2, 4, 5

Exercise 8 – anomaly detection

Consider a data set containing a single cluster with the points { (1, 1), (0, 0), (2, 2.1), (3, 3.1), (4, 4), (5.1, 5) }.

a) Which point does a 1-NN algorithm set as the highest outlier score with the Euclidean metric?

b) Which point does a 1-NN algorithm set as the lowest outlier score with the Euclidean metric?