Mid-term Exam Solutions

Student Name: **Student ID:**

Exam date: 2015-07-10, 6-8pm

Close Books & Notes, calculators only

Problem 1) Python programming (20 pts.)

a) What is the output of the following Python code?

```
from numpy import array
a = array([0,1,2,3,4,5,6])
print a[1:-1:2]
```

b) What is the output of the following Python code?

```
foo = {1:'1', 2:'2', 3:'3'}
del foo[1]
foo[1] = '10'
del foo[2]
print len(foo)
```

c) What is the output of the following Python code?

```
list = [1, 2, 3, 4, 5]
list1 = [2, 4, 6, 8]
def doStuff(list, counter=0):
    for i in list
        counter += i
     counter /= (len(list)+1)
     return counter
counter = 5
counter = doStuff(list1+list)
print counter
```

d) What is the output of the following Python code?

```
names1 = ['Adam', 'Bravo', 'Charlie', 'Delta']
names2 = names1
names3 = names1[:]

names2[0] = 'Alice'
names3[1] = 'Bob'

sum = 0
for ls in (names1, names2, names3):
    if 'Charlie' in ls:
        sum += 5
    if 'Bob' in ls:
        sum += 10

print sum
```

Answers:

a) [1, 3, 5]
b) 2
c) 3
d) 20

Problem 2) Bayes Rule (10 pts.)
You are given 2 boxes, one Red and one Green.
• The Red box contains 6 Orange, 3 Apples and 6 Cherries.
• The Green box contains 4 Orange, 7 Apples and 4 Cherries.

Assume a prior distribution of [Red Green] = [0.4 0.6] for the box. If we have drawn a cherry, what are probabilities that it comes from a Red or Green box, respectively?

	Orange	Apple	Cherry	Sum
Red	6	3	6	15
Green	4	7	4	15
Sum	10	10	10	

Answers:

P(Cherry) = P(Cherry | Red)*P(Red) + P(Cherry | Green)*P(Green) = 4/10*6/15 + 6/10*4/15 = 48/150 = 4/25

P(Red | Cherry) = P(Cherry | Red) * P(Red) / P(Cherry) = 4/10*6/15 / (4/25) = **0.5**

P(Green | Cherry) = P(Cherry | Green) * P(Green) / P(Cherry) = 6/10*4/15 / (4/25) = **0.5**

Problem 3) Naïve Bayes Classifier (20 pts.)

You are given the following customers summary table from *MicroShop* as training data:

age	income	student	buys computer
<=30	high	no	no
<=30	high	no	no
>40	low	yes	no
<=30	medium	no	no
<=30	medium	no	no
>40	medium	no	no
> 40	low	no	no
31…40	low	yes	no
<=30	medium	yes	yes
31…40	high	no	yes
>40	medium	no	yes
>40	high	yes	yes
31...40	medium	no	yes
31…40	low	yes	yes
<=30	low	yes	yes
>40	high	yes	yes
<=30	medium	yes	yes
31…40	medium	no	yes
31…40	high	yes	yes
>40	high	no	yes

Please answer the following questions by using Naïve Bayes classifier and Laplace smoothing if necessary.

a) Derive the conditional probabilities of P(high_income | buys_computer), P(<=30 | buys_computer), P(student | buys_computer),
b) Using Naïve Bayes classifier, what is the probability of a customer with the following profile P(<=30, high_income, student) buying a computer from *MicroShop*?

Answers:

a) **P(<=30 | buys_computer) = 3/12**
P(high | buys_computer) = 5/12
P(student | buys_computer) = 7/12

P(buys_computer) = 12/20
P(31…40 | buys_computer) = 5/12
P(>40 | buys_computer) = 4/12
P(medium | buys_computer) = 5/12
P(low | buys_computer) = 2/12
P(not_student | buys_computer) = 5/12
P(student | not_buys_computer) = 2/8
P(high | not_buys_computer) = 2/8
P(<=30 | not_buys_computer) = 4/8

b) Likelihood(buys_computer | <=30, high_income, student) = 12/20 * 3/12 * 5/12 * 7/12 = 0.03646
Likelihood (not_buys_computer | <=30, high_income, student) = 8/20 * 2/8 * 2/8 * 4/8 = 0.0125
P(buys_computer | <=30, high_income, student) = 0.03646/(0.03646+0.0125) = 0.7447

Problem 4) k-Nearest Neighbors (20 pts.)
We have the following training data for height/weight vs gender:

Height	70	65	66	60	58	62
Weight	175	170	168	150	155	160
Gender	0	0	0	1	1	1

Gender=0 is male and Gender=1 is female. Now we have 2 test cases: (h=68, w=160) and (h=60, w=162). Predict the gender for both cases using the kNN classifier with the following properties: k=3, distance = Manhattan, weight = uniform

Answers:

Distance for (h=68, w=160):

Height	70	65	66	60	58	62
Weight	175	170	168	150	155	160
Gender	0	0	0	1	1	1
h=68, w=160	17	13	10	18	15	6

(h=68, w=160) is closest to (62,160) (66, 180) (65, 170) => gender=0

Distance for (h=60, w=162):

Height	70	65	66	60	58	62
Weight	175	170	168	150	155	160
Gender	0	0	0	1	1	1
h=60, w=162	23	13	12	12	9	4

(h=60, w=162) is closest to (62,160) (58,155) (60,150) => gender=1

Problem 5) K-means clustering (10 pts.)

You are given the following 10 data points of height & weight:

ID	1	2	3	4
Height	78	72	60	64
Weight	174	172	156	164

Apply k-means algorithms to get 2 clusters with the following parameters:

a) Initialize with ID=3 and ID=4
b) Euclidean distance

What are the final groupings and the cluster means?

Hint: In computing Euclidean distance to determine groupings, it is not necessary to take the square root, since the square root operation preserves the order of positive real numbers.

Answer:

step- 0: [([60.0, 156.0], [3]), ([64.0, 164.0], [1, 2, 4])]

step- 1: [([60.0, 156.0], [3, 4]), ([71.3, 170.0], [1, 2])]

step- 2: [([62.0, 160.0], [3, 4]), ([75.0, 173.0], [1, 2])]

Final: 2 clusters ([62.0, 160.0], [3,4]) and ([75.0, 173.0], [1, 2])]

Problem 6) Dendrogram (20 pts.)

You are given the 5 cities: Antioch (A), Berkeley (B), Cupertino (C), Danville (D), Evergreen (E), and the distances between any 2 cities as follows:

Distances	Antioch (A)	Berkeley (B)	Cupertino (C)	Danville (D)	Evergreen (E)
Antioch (A)	0	45	92	35	84
Berkeley (B)	45	0	64	25	52
Cupertino (C)	92	64	0	75	17
Danville (D)	35	25	75	0	57
Evergreen (E)	84	52	17	57	0

Assume single linkage for cluster distance, construct the resulting dendrogram to cluster the 5 cities based on their distances. Please clearly mark the resulting distance at each level.

Answer: