Mid-term Exam Sample Questions & Solutions

Release date: 2015-07-03

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

- 2 hrs.
- Calculator
- close book

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[-3:-1]
```

[4, 5]

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

```
names = ['Adam', 'Bravo', 'Charlie', 'Delta']
print names[-1][-2]
```

't'

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

```
counter = 0
list = [6, 2, 4]
list1 = [1, 3, 5, 7, 9]
def doStuff(list):
    global counter
    for i in list
        counter += i
     counter /= (len(list)+1)

doStuff(list1)
print counter
```

int

1+3+5+7+9 / 6 = 25/6 → 4.

integer

37/9 = 4

(if list+list1 → 6, 24, 1, 35, ...)

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 ls[0] == 'Alice':
        sum += 1
    if ls[1] == 'Bob':
        sum += 10

print sum
```

only ref

also affect names1

not affect names1

Answers:

a) [4, 5]
b) t
c) 4
d) 12

Problem 2) Bayes Rule (10 pts.)
You are given three boxes, one Red, one Blue and one Green.
• The Red box contains 1 Orange, 4 Apples and 5 Cherries.
• The Green box contains 3 Orange, 4 Apples and 3 Cherries.
• The Blue box contains 7 Orange, 2 Apples and 1 Cherries.

Assume a prior distribution of [Red Green Blue] = [0.2 0.5 0.3]. What are probabilities that a Cherry was drawn from a Red, Green, and Blue box?

Solution:

given a Cherry

	Orange	Apple	Cherry	Sum
Red	1	4	5	10
Green	3	4	3	10
Blue	7	2	1	10
Sum	11	10	9	

P(Cherry) = P(Cherry | Red) * P(Red) + P(Cherry | Green) * P(Green) + P(Cherry | Blue) * P(Blue) = 5/10 * 1/5 + 3/10 * ½ + 1/10 * 3/10 = 1/10 + 3/20 + 3/100 = (10+15+3)/100 = 28/100 = 7/25

P(Red | Cherry) = P(Cherry | Red) * P(Red) / P(Cherry) = 5/10 * 1/5 / (7/25) = 25/70 = 5/14

P(Green | Cherry) = P(Cherry | Green) * P(Green) / P(Cherry) = 3/10 * 1/2 / (7/25) = 15/28

P (Blue | Cherry) = P(Cherry | Blue) * P(Blue) / P(Cherry) = 1/10 * 3/10 / (7/25) = 3/28

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	medium	yes	yes
<=30	high	no	no
31…40	high	no	yes
>40	medium	no	yes
>40	low	yes	yes
>40	low	yes	no
31...40	medium	no	yes
31…40	low	yes	yes
<=30	medium	no	no
<=30	low	yes	yes
>40	high	yes	yes
<=30	medium	no	no
<=30	medium	yes	yes
31…40	medium	no	yes
31…40	high	yes	yes
>40	medium	no	no
> 40	low	no	no
31…40	low	yes	no
>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(<=30 | buys_computer), P(31…40 | buys_computer) and P(not_student | buys_computer)
b) What is the probability of a customer with the following profile (>40, medium_income, student) buying a computer from *MicroShop*?

Answers:

a) P(buys_computer) = 12/20
P(<=30 | buys_computer) = 3/12
P(31…40 | buys_computer) = 5/12
P(>40 | buys_computer) = 4/12
P(student | buys_computer) = 7/12
P(not_student | buys_computer) = 5/12
b) P(medium_income | buys_computer) = 5/12
P(>40, medium_income, student | buys_computer) = 4/12 * 5/12 * 7/12 = 35/432 = 0.081
P(>40, medium_income, student | not_buys_computer) = 3/8 * 3/8 * 2/8 = ~~9/128 = 0.07~~
P(buys_computer | >40, medium_income, student) = 0.08/(0.07+0.08) = 8/15

Problem 4) Decision Tree (20 pts.)
We are given the following training data set from *MicroShop* customers. Assuming Maximum Information gain is employed as the criterion to select features, which feature will be used to split the first level and what is the resulting information gain using the best split? Information is defined as:

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i)$$

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

Answers:

Calculate the information gains as follows.

p(buy)	0.33			
p(not_buy)	0.67			
info(D)	0.92			
p(<=30)	0.47		p(high)	0.27
p(buy|<=30)	0.57		p(buy|high)	0.50
p(not_buy|<=30)	0.43		p(not_buy|high)	0.50
info(D(<=30))	0.99		info(D(high))	1.00
p(>40)	0.27		p(low)	0.27

p(buy|>40)	0.75		p(buy|low)	0.75
p(not_buy|>40)	0.25		p(not_buy|low)	0.25
info(D(>40))	0.81		info(D(low))	0.81
p(31..40)	0.27		p(medium)	0.47
p(buy|31..40)	1.00		p(buy|medium)	0.71
p(not_buy|31..40)	0.00		p(not_buy|lmedium)	0.29
info(D(31..40))	0.00		info(D(medium))	0.86
info(age)	0.68		info(income)	0.89
info_gain(age)	0.24		info_gain(income)	0.03

Therefore, **use age for the first split for the decision tree**,

and the resulting **information_gain = 0.24**

Euclidean distance = $\sqrt{(x_1-x_2)^2+(y_1-y_2)^2}$

Manhattan distance = $|x_1-x_2|+|y_1-y_2|$

	76,180	65,200	72,165	65,150	60,140	68,158
68,155	25	48	14.	8	23.	3
72,160	24.	47				

Problem 5) k-Nearest Neighbors (10 pts.)

We have the following training data for height/weight vs gender:

Height	76	65	72	65	60	68
Weight	180	200	165	150	140	158
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=155) and (h=72, w=160). Predict the gender for both cases using the kNN classifier with the following properties: k=1, distance = Manhattan, weight = uniform

if k=3, uniform

use 3 point, and vote (2/3) wins

Answers:

(h=68, w=155) is closest to (68,158) => gender=1
(h=72, w=160) is closest to (72,165) => gender=0

Problem 6) K-means clustering (10 pts.)

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

ID	1	2	3	4
Height	76	73	62	60
Weight	180	200	150	140

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

a) Initialize with ID=1 and ID=2 ← should start with at least data
b) Assume Euclidean distance

What are the final groupings and the cluster means?

Answer:

step-0: [([76, 180], [1, 3, 4]), ([73, 200], [2])]

step-1: [([66.0, 156.66666666666666], [1, 2]), ([73.0, 200.0], [3, 4])]

step-2: **[([74.5, 190.0], [1, 2]), ([61.0, 145.0], [3, 4])]**

Problem 7) Model Evaluation (10 pts.)
We are given the following classification results for a cancer screening test. Please derive the evaluation parameters:

a) Accuracy
b) Specificity
c) Precision
d) Recall
e) F1

Actual / Prediction	test = positive	test = negative	Total
cancer = yes	**150**	**210**	360
cancer = no	**80**	**9560**	9640
Total	230	9770	10000

Answer:

a) Accuracy = (150+9560) / 10000 = 0.971
b) Specificity = 9560 / 9640 = 0.992
c) Precision = 150 / 230 = 0.652
d) Recall = 150 / 360 = 0.417
e) F1 = 2*P*R / (P+R) = 0.508