Homework #2 Solutions

Assign date: 2015-06-12

Due date: 2015-06-25, 6pm

Submission:

1. **Please submit your results in email to the grader: 130301039@svuca.edu Chi Zhang**
2. **Please separate the written answers (Problems 1, 2) from the python code: you should submit 2 files in your email – cs596-29-hw2_yourID#.doc & cs596-29-hw2_yourID#.py**
3. **20 pts per day will be deducted for late submission**

Problem 1) Naïve Bayes Classifier (30 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(>40 | buys_computer)
b) Derive the conditional probabilities of P(student | buys_computer) and P(not_student | buys_computer)
c) 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
b) P(student | buys_computer) = 7/12
   P(not_student | buys_computer) = 5/12
c) 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 2) 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?

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

Problem 3) k-Nearest Neighbors (10 pts.)
We have the following training data for height/weight vs gender:

Height	66	73	72	70	74	68	65	64	63	67	68	66
Weight	170	210	165	180	185	155	150	120	125	140	165	130
Gender	0	0	0	0	0	0	1	1	1	1	1	1

Gender=0 is male and Gender=1 is female. Now we have 2 test cases: (h=69, w=155) and (h=72,w=160). Predict the gender for both cases using the kNN classifier with the following properties:

a) k=1, distance = Euclidean, weight = uniform
b) k=3, distance = Manhattan, weight = uniform

Answers:

a) (h=69, w=155) is closest to (68,155) => gender=0
(h=72, w=160) is closest to (72,165) => gender=0

b) (h=69, w=155) is closest to (68,155) (65, 150) (68,165) => gender=1

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

Problem 4) Python program (40 pts)

Take the sk-learn sample code *regr_dtree.py* discussed in the class and inputs.csv data. Make the following changes:

1. Replace the linear regression classifier with K-nearest neighbor and set k=5.
2. Remove the "gender" (2nd) feature from the data

Now use the training set to train the decision tree and the k-NN classifiers, and then apply the model on the test set. **Output: Print out the classifier scores and the comparison of prediction vs. ground truth vectors.**