# Homework #2

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**](mailto: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 – hw2\_yourID#.doc & 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:  


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

1. Derive the conditional probabilities of P(<=30 | buys\_computer), P(31…40 | buys\_computer) and P(>40 | buys\_computer)
2. Derive the conditional probabilities of P(student | buys\_computer) and P(not\_student | buys\_computer)
3. What is the probability of a customer with the following profile (>40, medium\_income, student) buying a computer from *MicroShop*?

**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?



**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:

1. k=1, distance = Euclidean, weight = uniform
2. k=3, distance = Manhattan, weight = uniform

**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.**