# Homework #1

Assign date: 2015-05-29

**Due date: 2015-06-07, 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, 3) from the python code: you should submit 2 files in your email - hw1.doc & hw1.py**
3. **20 pts per day will be deducted for late submission**

**Problem 1)** 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.

1. Assuming an even prior distribution, what are the posterior probability that an Orange was drawn from a Red, Green, or Blue box?
2. 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?

**Problem 2)** Bayes Rule (30 pts.)  
We know that in general 30% of the $20 toy cameras made by the company ***DirtCheapCam*** are defective. The QA engineers in Walmart have develop 2 separate tests to detect the bad cameras and they have the following performance.

Test-1: Sensitivity or TP (true positive) is 80% and Specificity or TN (true negative) is 80%

Test-2: Sensitivity or TP (true positive) is 90% and Specificity or TN (true negative) is 70%

1. For both tests, derive the *posterior* probability that a camera is really defective when it fails the test. (correct detection)
2. For both tests, derive the *posterior* probability that a camera is defective but it passes the test. (miss detection)
3. If a defective camera is not detected (miss detection), in average it will cost Walmart $12 to handle customer complaint and return. On the other hand, if a good camera is incorrectly rejected (false alarm) by the test, it will cost Walmart $6 in manufacturing cost. Also assume that the costs of the two tests are equal and Walmart only wants to implement one test. Considering all the costs, discuss which of the 2 tests is more preferred and why.

**Problem 3)** Linear Regression (20 pts)

Using the following table of Fat (X) vs Calories (Y) in Burgers, apply the linear regression fitting and find out the slope and intercept.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | HealthyBun | Burgerlet | YourBurg | FatBurger | GreesyJoint |
| Fat | 19 | 31 | 35 | 39 | 43 |
| Calories | 410 | 580 | 570 | 640 | 660 |

Note: You may use Python or SK-learn to calculate but please provide the written answer, not the code.

**Problem 4)** Python program (40 pts)

Take the sk-learn sample code *plot\_ols.py* discuss in the class. Make the following changes:

1. Replace **datasets.load\_diabetes()** with **datasets.make\_regression()** function, as in the other sample code *plot\_ransac.py*
2. Ask the user to input noise level instead of using the preset value (noise=10)
3. Generate 100 samples, and separate them into 2 groups:
   1. 80 points as training set
   2. 20 points as test set

Now use the training set to train the linear regression model, and then apply the model on the test set. **Output: Print out the coefficients and the residual sum of squares on the test data.**