Homework #1 Solution Corrected

Assign date: 2015-05-29

Due date: 2015-06-11, 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, 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.

a) Assuming an even prior distribution, what are the posterior probability that an Orange was drawn from a Red, Green, or Blue box?
b) 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:

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

a) P(Orange) = P(Orange | Red) * P(Red) + P(Orange | Green) * P(Green) + P(Orange | Blue) * P(Blue) = 1/10 * 1/3 + 3/10 * 1/3 + 7/10 * 1/3 = 11/30

P(Red | Orange) = P(Orange | Red) * P(Red) / P(Orange) = 1/10 * 1/3 / (11/30) = 1/11

P(Green | Orange) = P(Orange | Green) * P(Green) / P(Orange) = 3/10 * 1/3 / (11/30) = 3/11

P (Blue | Orange) = P(Orange | Blue) * P(Blue) / P(Orange) = 7/10 * 1/3 / (11/30) = 7/11

b) 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 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 developed 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%

a) For both tests, derive the *posterior* probability that a camera is really defective when it fails the test. (correct detection)
b) For both tests, derive the *posterior* probability that a camera is defective but it passes the test. (miss detection)
c) 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.

Solution:

NG: no good part (defective) ~NG: good part (non-defective)

d: detected (positive in the test) ~d: not detective (negative in the test)

Prior		
P(NG) =	0.3	
P(~NG) =	0.7	
cost(~NG|d) =	6	
cost(NG|~d) =	12	
	Test-1	Test-2
TP=P(d|NG) =	0.8	0.9
TN=P(~d|~NG) =	0.8	0.7

FN=P(~d|NG) =	0.2	0.1
FP=P(d|~NG) =	0.2	0.3
Post P(NG|d) =	0.631579	0.5625
Post P(~NG|d) =	0.368421	0.4375
Post P(NG|~d) =	0.096774	0.057692
Post P(~NG|~d) =	0.903226	0.942308

a) Test-1 P(NG|d) = 0.632, Test-2 P(NG|d) = 0.563
b) Test-1 P(NG|~d) = 0.097, Test-2 P(NG|~d) = 0.058
c) Test-1 is more preferred, because its average cost is less than test-2.

Cost =\$12 × *P* (*NG, ~d*) + \$6 × *P* (*~NG, d*)
=\$12 × *P* (*~d | NG*) *P* (*NG*) + \$6 × *P* (*d | ~NG*) *P* (*~NG*)
=\$12 × (1 − *TP*) × 30% + \$6 × (1 − *TN*) × 70%
For both tests, the costs are
Cost-test-1 = \$1.56
Cost-test-2 = \$1.62

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

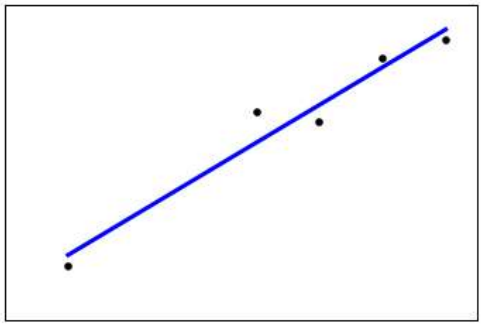
Note: You may use Python or SK-learn to calculate but please provide the written answer, not the code.

Solution:

From sk-learn

Linear Regression:

[10.4245283] 223.82073881



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:
    a. 80 points as training set
    b. 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.**