# Homework #3

Assign date: 2015-06-21

**Due date: 2015-07-01, 6pm**

Submission:

1. **Please submit your results in email to the grader: 130301039@svuca.edu Chi Zhang**
2. **Please separate the written answers from the python code: you should submit 2 files in your email – cs596-29-hw2_yourID#.doc & cs596-29-hw2_yourID#.py**
3. **30 pts per day will be deducted for late submission**

**Problem 1)** K-means clustering (20 pts.)

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

| ID | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Height | 66 | 73 | 72 | 70 | 74 | 68 | 65 | 64 | 63 | 67 |
| Weight | 170 | 210 | 165 | 180 | 185 | 155 | 150 | 120 | 125 | 140 |

Manually apply k-means algorithms to get 2 clusters. Please produce the center and grouping step by step, using the following parameters:

a) Initialize with ID=1 and ID=2
b) Assume Euclidean distance

**Problem 2)** Agglomerative clustering (20 pts.)

Using the same 10 data points of height & weight as in Problem-1, apply agglomerative clustering manually. Produce the distance matrix and the resulting dendogram. Use the following assumptions:

a) Euclidean distance
b) Single linkage for cluster dissimilarity

**Problem 3)** Model Evaluation (20 pts.)
We are given the following classification results for a cancer screening test. Please derive the evaluation parameters:

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

| Actual / Prediction | test = positive | test = negative | Total |
|---|---|---|---|
| cancer = yes | **90** | **210** | 300 |
| cancer = no | **140** | **9560** | 9700 |
| Total | 230 | 9770 | 10000 |

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

Take the sk-learn sample code *cluster_kmeans.py* discussed in the class that is used for clustering iris dataset. Make the following changes:

*1.* In addition to using PCA to reduce features from 4 to 2, also evaluate using the original feature pairs (1,2) and (3, 4).
*2.* For all the 3 clustering settings (original, (1,2), (3, 4)), calculate the clustering quality CQ = IE / EV as defined in the class:

$$IV = \sum_{C} \sum_{x \in C} d(x, c)$$ and $$EV = \frac{1}{N} \sum_{i} \sum_{j} \delta(C(x_i) \neq C(x_j)) d(x_i, x_j)$$

**Output:**

**3 plots of k-means clustering results for PCA, features (1,2), and features (3,4).**

**3 CQ values for PCA, features (1,2), and features (3,4).**