# B529: Homework 3

*Nathan Byers*

*Wednesday, April 7, 2015*

## Question 1

In the clustering problems, the input data set contains the following six data points (1,1), (0,1), (1,0), (3,3) (4,4), (5,4). Manhattan distance is chosen to compute the distance between two data points. The k-means clustering method is used to find the clusters with k = 2 and initial centroids (3,3) and (5,4). Please describe the steps of the k-means clustering. (In each step, provide the information about cluster assignments of data points, centroids; and when the algorithm stops). (25 points)

## Answer 1

## Question 2

Suppose DNA bases in a protein-coding region follow the distribution:

| DNA base | Probability |
|---|---|
| A | $\theta$ |
| C | $\frac{1}{4}$ |
| G | $\frac{1}{2}$ |
| T | $\frac{1}{4} - \theta$ |

In an experiment, the number of observed "A" or "C" bases at the position is x, and the number of observed "G" or "T" bases at the position is y. The EM algorithm can be used to find parameter $\theta$. Describe the Expectation step and Maximization step in the EM algorithm (25 points)

## Answer 2

## Question 3

Use the ID3 method to construct a decision tree using the following data set for credit card application. (25 points)

| Age | Income | Gender | Risk |
|---|---|---|---|
| <25 | >50K | M | High |
| <25 | >50K | F | High |
| $\geq$ 25 | <50K | F | High |
| $\geq$ 25 | >50K | F | Low |

| Age | Income | Gender | Risk |
|---|---|---|---|
| $\geq 25$ | >50K | M | Low |
| <25 | <50K | M | High |

## Answer 3

## Question 4

A linear SVM is used to analyze a 2-dimensional data set, and the solution $\alpha$ to the quadratic programming program contains only three non-zero elements $\alpha_1 = 0.05$, $\alpha_2 = 0.05$, and $\alpha_3 = 0.1$. The corresponding input data points are $\mathbf{x}_1 = [0, 3]^T$, $y_1 = 1$, $\mathbf{x}_2 = [2, 4]^T$, $y_2 = 1$, $\mathbf{x}_3 = [3, -0.5]^T$, $y_3 = -1$. Compute the weight vector $\mathbf{w}$ and parameter b for the separating line with the maximum margin (15 points). Compute the margin of the separating line (10 points).