Data Mining and Decision Support Final (Practice)

1. Assume we have a dataset (shown in (1)) consisting of observations of what some customers have purchased in single trips to the supermarket (0 and 1 indicate, respectively, that the item was or was not purchased in the corresponding trip). Find the following probabilities given this data: $P(\text{checkout time} = \text{morning})$, $P(\text{beets} = 1|\text{checkout time} = \text{afternoon})$. Find the *sample variance* of the potatoes variable. Find the correlation between paying with cash and purchasing beets. Finally, use a Naive Bayes classifier to find the most likely checkout time for a customer who only purchases beets and juice and pays with a card.

$$\begin{bmatrix} \text{checkout time} & \text{beets} & \text{potatoes} & \text{juice} & \text{paid with cash} \\ \text{morning} & 1 & 0 & 0 & 0 \\ \text{morning} & 1 & 1 & 0 & 0 \\ \text{afternoon} & 0 & 0 & 1 & 0 \\ \text{evening} & 1 & 1 & 1 & 0 \\ \text{afternoon} & 0 & 0 & 1 & 0 \\ \text{afternoon} & 1 & 1 & 1 & 1 \\ \text{morning} & 1 & 0 & 0 & 1 \\ \text{evening} & 1 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

2. For the matrices given below solve for $\mathbf{x}$ in $\mathbf{Ax} = \mathbf{b}$ in the least-squares sense.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \\ 1 & 1 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix} \tag{2}$$

   Also, show how to solve this problem using gradient descent (hint: $\frac{d}{d\mathbf{A}}||\mathbf{Ax} - \mathbf{b}||^2 = 2\mathbf{A}^T(\mathbf{Ax} - \mathbf{b})$ and compute two iterations with learning parameter $\alpha = 0.1$. Provide the body of a loop in python for performing this calculation assuming the matrices are stored as `numpy` arrays.

3. In the figure below, draw the decision boundary for the linear classifier defined by $w_1x_1 + w_2x_2 > 1$. Assuming you are free to set the threshold th then what is the best error rate we can obtain with a classifier of the form $w_1x_1 + w_2x_2 >$ th. Assume that the vector $\mathbf{w} = [w_1, w_2] = [1, -0.2]$ and that x_1 and x_2 are the coordinate axes. Also, give the formula for linear discriminant analysis in terms of the data, what are the within-class covariance matrices for plusses and boxes?