Math 10A
Worksheet, Discussion #34; Thursday, 8/2/2018
Instructor name: Roy Zhao

1 Linear Regression

1.1 Concepts

1. Often when given data points, we want to find the line of best fit through them. To them, we want to approximate them with a line $y = ax + b$. We represent this as a solution where we want to solve for a, b. In matrix vector form and data points (x_i, y_i), this is represented as

$$A\vec{x} = \vec{b} \to \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

Often, we cannot find a perfect fit (if not all the points lie on the same line). So we want to find the error. One way to find the error is to take the least square error or $E = \sum(y_i - (ax_i + b))^2$, the sum of the squares of the error. The choice of a, b that minimizes this is

$$\begin{pmatrix} a \\ b \end{pmatrix} = (A^T A)^{-1} A^T \vec{b}.$$

Written out, we have

$$a = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x - \bar{x})^2}, b = \bar{y} - a\bar{x},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the average of the x values and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the average of the y values.

1.2 Problems

2. True False The matrix $A^T A$ will always be square.

3. Consider the set of points $\{(-2, -1), (1, 1), (3, 2)\}$. Calculate the line of best fit.

4. Find the line of best fit and the error of the fit of the points $\{(-1, 2), (0, -1), (1, 1), (3, 2)\}$ and use it to estimate the value at 2.

5. Consider the set of points $\{(-2, -1), (1, 1), (3, 2)\}$. Calculate the square error if we estimate it using the line $y = x$. Then calculate the square error if we use the line $y = 0$. Which is a better approximation?

6. The number of people applying to Berkeley is given in the following table:

Year	2011	2012	2013	2014	2015	2016	2017
Applicants(in 1000s)	53	62	68	74	79	83	85

Predict how many people applied this year (2018).