Introduction to Statistical Machine Learning
Problem set 1

Alan Joseph Bekker

Assume that there are n given training examples (x_1, y_1), . . . ,(x_n, y_n), where each input data point x_i has m real valued features, $x_i \in R^m$. The goal of regression is to learn to predict y from x. The linear regression model assumes that the output y is a linear combination of the input features x plus noise terms $\epsilon \sim \mathcal{N}(0, \sigma)$ with weights given by β as shown in equation 1.

$$Y = X\beta + \epsilon \tag{1}$$

where

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{pmatrix}, X = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1m} \\ x_{21} & x_{22} & ... & x_{2m} \\ . & & & \\ . & & & \\ x_{n1} & x_{n2} & ... & x_{nm} \end{pmatrix}, \epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ . \\ . \\ \epsilon_n \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_m \end{pmatrix}.$$

Linear regression seeks to find the parameter vector β that provides the best fit of the above regression model. One criteria to measure fitness, is to find that minimizes a given loss function $J(\beta)$. In class, we have shown that if we take the loss function to be the square-error, i.e.:

$$J(\beta) = (X\beta - Y)^T (X\beta - Y) \tag{2}$$

then,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{3}$$

Moreover, we have also shown that if we assume that ϵ_1 . . . ϵ_n are IID and sampled from the same zero mean Gaussian that is,$\epsilon \sim \mathcal{N}(0, \sigma)$, then the MSE is also the MLE estimate for $P(Y|X; \beta)$.

In this assigmnet you will need to create your own data as following:
1.Produce your own X matrix using rand(n,m) command.
2.Determine your β parameters.
3.Produce your noise vector ϵ using randn(n,1) command .
4.Compute the Y vector using equation 1.
5.Use equation 3 to solve the β parameters.
6.Compare your solution to the β parameters you determined in 2.
7.Increase the σ of your noise vector and check how this influence the estimation of the parameters.