Introduction to Machine Learning
Homework 2: Multiple Linear Regression

Prof. Sundeep Rangan

1. An online retailer like Amazon wants to determine which products to promote based on reviews. They only want to promote products that are likely to sell. For each product, they have past sales as well as reviews. The reviews have both a numeric score (from 1 to 5) and text.

   (a) To formulate this as a machine learning problem, suggest a target variable that the online retailer could use.

   (b) For the predictors of the target variable, a data scientist suggests to combine the numeric score with frequency of occurrence of words that convey judgement like "bad", "good", and "doesn't work." Describe a possible linear model for this relation.

   (c) Now, suppose that some reviews have a numeric score from 1 to 5 and others have a score from 1 to 10. How would change your features?

   (d) Now suppose the reviews have either (a) a score from 1 to 5; (b) a rating that is simply good or bad; or (c) no numeric rating at all. How would you change your features?

   (e) For the frequency of occurrence of a word such as "good", which variable would you suggest to use as a predictor: (a) total number of reviews with the word "good"; or (b) fraction of reviews with the word "good"?

2. Suppose we are given data:

x_{i1}	0	0	1	1
x_{i2}	0	1	0	1
y_i	1	4	3	7

   (a) Write an equation for a linear model for y in terms of x_1 and x_2.

   (b) Given the data compute the least-squares estimate for the parameters in the model.

3. An automobile engineer wants to model the relation between the accelerator control and the velocity of the car. The relation may not be simple since there is a lag in depressing the accelerator and the car actually accelerating. To determine the relation, the engineers measures the acceleration control input x_k and velocity of the car y_k at time instants $k = 0, 1, \ldots, T-1$. The measurements are made at some sampling rate, say once every 10 ms. The engineer then wants to fit a model of the form

$$y_k = \sum_{j=1}^{M} a_j y_{k-j} + \sum_{j=0}^{N} b_j x_{k-j} + \epsilon_k, \tag{1}$$

for coefficients a_j and b_j. In engineering this relation is called a *linear filter* and it statistics it is called an *auto-regressive moving average (ARMA)* model.

(a) Describe a vector $\boldsymbol{\beta}$ with the unknown parameters. How many unknown parameters are there?

(b) Describe the matrix $\mathbf{A}$ and target vector $\mathbf{y}$ so that we can rewrite the model (1) in matrix form,

$$\mathbf{y} = \mathbf{A}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Your matrix $\mathbf{A}$ will have entries of y_k and x_k in it.

(c) (Graduate students only) Show that, for $T \gg N$ and $T \gg M$, the coefficients of $(1/T)\mathbf{A}^\mathsf{T}\mathbf{A}$ and $(1/T)\mathbf{A}^\mathsf{T}\mathbf{y}$ can be approximately computed from the so-called auto-correlation functions

$$R_{xy}(\ell) = \frac{1}{T}\sum_{k=0}^{T-1} x_k y_{k+\ell}, \quad R_{yy}(\ell) = \frac{1}{T}\sum_{k=0}^{T-1} y_k y_{k+\ell}.$$

In the sum, we take $x_k = 0$ or y_k whenever $k < 0$ or $k \geq T$.

4. In audio processing, one often wants to find tonal sounds in segments of the recordings. This can be formulated as follows: We are given samples of an audio segment, x_k, $k = 0, \ldots, N-1$, and wish to fit a model of the form,

$$x_k \approx \sum_{\ell=1}^{L} a_\ell \cos(\Omega_\ell k) + b_\ell \sin(\Omega_\ell k), \tag{2}$$

where L are a number of tones present in the audio segment; Ω_ℓ are the tonal frequencies and a_ℓ and b_ℓ are the coefficients.

(a) Show that if the frequencies Ω_ℓ are given, we can solve for the coefficients a_ℓ and b_ℓ using linear regression. Specifically, rewrite the model (2) as $\mathbf{x} \approx \mathbf{A}\boldsymbol{\beta}$ for appropriate $\mathbf{x}$, $\mathbf{A}$ and $\boldsymbol{\beta}$. Then describe exactly how we obtain the coefficients a_ℓ and b_ℓ from this model.

(b) Now suppose the frequencies Ω_ℓ were not known. If we had to solve for the parameters a_ℓ, b_ℓ and Ω_ℓ, would the problem be a linear regression problem?