ECE M146 Discussion 4
Introduction to Machine Learning
Instructor: Lara Dolecek
TA: Zehui (Alex) Chen

1. **Multi-class Least Squares** In this section, you will determine the parameter matrix $\mathbf{W} \in \mathbb{R}^{m \times p}$ for the Multi-class Least Squares problem.

   Given a data matrix $\mathbf{X} \in \mathbb{R}^{n \times m}$ and target matrix $\mathbf{T} \in \mathbb{R}^{n \times p}$, the sum-of-squares error function can be written as

   $$Er(\mathbf{W}) = \mathbf{Tr}\{(\mathbf{XW} - \mathbf{T})^T(\mathbf{XW} - \mathbf{T})\}$$

   where $\mathbf{Tr}$ is the trace of a matrix. You can assume that $\mathbf{X}$ has full rank.

   We will solve this problem by setting the derivative with respect to $\mathbf{W}$ to be zero and solve for $\mathbf{W}$. To do this we must first know some matrix derivative properties.

   (a) Let $\mathbf{A}, \mathbf{Z}$ be two matrices. Prove

   $$\frac{d\mathbf{Tr}(\mathbf{AZ})}{d\mathbf{Z}} = \mathbf{A}^T$$

(b) Let $\mathbf{A}, \mathbf{Z}$ be two matrices. Prove

$$\frac{d\mathbf{Tr}(\mathbf{Z}\mathbf{A}\mathbf{Z}^T)}{d\mathbf{Z}} = \mathbf{Z}\mathbf{A}^T + \mathbf{Z}\mathbf{A}$$

(c) Now, we can take the derivative of $Er(\mathbf{W})$ and set it to zero. Show that this results in

$$\mathbf{W} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{T}$$

2. In this problem, we will derive the least square solution for multi-class classification. Consider a general classification problem with K classes, with a 1-of-K binary encoding scheme (defined latter) for the target vector $t, t \in \mathbb{R}^K$. Suppose we are given a training data set $\{x_n, t_n\}, n = 1, \cdots, n$ where $x_n \in \mathbb{R}^D$. For the 1-of-K binary encoding scheme, t_n has the k-th element being 1 and all other elements being 0 if the n-th data is in class k. We can use the following linear model to describe each class:

$$y_k(x) = w_k^T x + w_{k0},$$

where $k = 1, \cdots, K$. We can conveniently group these together using vector notation so that

$$y(x) = \tilde{\mathbf{W}}^T \tilde{x},$$

where $\tilde{\mathbf{W}}$ is a matrix whose k-th column comprises the $D+1$-dimensional vector $\tilde{w} = [w_{k0}, w_k^T]^T$ and $\tilde{x}$ is the corresponding augmented input vector $[1, x^T]^T$. For each new input with feature x, we assign it to the class for which the output $y_k = \tilde{w}_k^T \tilde{x}$ is largest. Define a matrix $\mathbf{T}$ whose n-th row is the vector t_n^T and together a matrix $\tilde{\mathbf{X}}$ whose n-th row is $\tilde{x}_n^T$, the sum-of-squares error function can be written as

$$J(\tilde{\mathbf{W}}) = \frac{1}{2}\mathbf{Tr}\left\{(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})^T(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \mathbf{T})\right\}.$$

Find the closed form solution of $\tilde{\mathbf{W}}$ that minimizes the objective function $J(\tilde{\mathbf{W}})$. Hint: You many use the following two matrix derivative about trace, $\frac{\partial}{\partial Z}\mathbf{Tr}(\mathbf{AZ}) = \mathbf{A}^T$ and $\frac{\partial}{\partial Z}\mathbf{Tr}(\mathbf{Z^T A Z}) = (\mathbf{A^T} + \mathbf{A})\mathbf{Z}$.