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

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

   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

$$\begin{bmatrix} \frac{\partial Tr(AZ)}{\partial z_{11}} & \frac{\partial Tr(AZ)}{\partial z_{12}} & \cdots \\ \frac{\partial Tr(AZ)}{\partial z_{21}} & \ddots & \\ \vdots & & \ddots \end{bmatrix} \Leftarrow \qquad \frac{d\mathbf{Tr}(\mathbf{AZ})}{d\mathbf{Z}} = \mathbf{A}^T$$

$$\frac{\partial\, Tr(AZ)}{\partial z_{ij}} = \frac{\partial \sum_{i'} \sum_{j'} z_{i'j'} a_{j'i'}}{\partial z_{ij}} = a_{ji}$$

$$\frac{\partial\, Tr(AZ)}{\partial Z} = A^T$$

Linear Regression (Least Square Problem)

$$x_1 \cdots x_n \in R^m \qquad \hat{t}_i = w^T x_i$$

$$t_1 \cdots t_n \in R$$

$$\min_{w} \sum_{i=1}^{n} (w^T x_i - t_i)^2$$

$$\|Xw - t\|^2 \qquad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

Multi-class Least Square Regression

$$x_1 \cdots x_n \in R^m$$

$$\begin{bmatrix} t_{11} & \cdots & t_{n1} \\ t_{12} & \cdots & t_{n2} \\ \vdots & & \vdots \\ t_{1\phi} & \cdots & t_{n\phi} \end{bmatrix} \begin{matrix} \in R \\ \in R \\ \\ \in R \end{matrix}$$

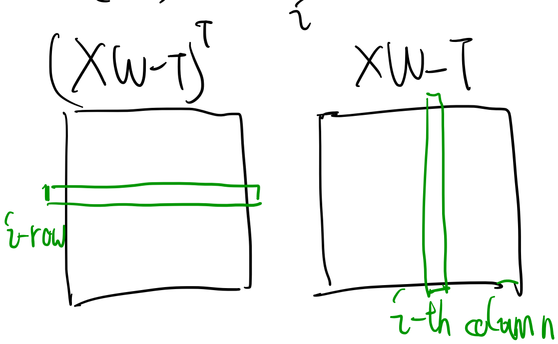
$$\hat{t}_{ij} = w_j^T x_i$$

$$\min_{w_1, w_2 \ldots w_\phi} \sum_{j=1}^{\phi} \sum_{i=1}^{n} (w_j^T x_i - t_{ij})^2$$

$\Downarrow$ matrix-vector form

$$Tr(A) = \sum_i A_{ii}$$

$$Tr\{(XW - T)^T (XW - T)\} \qquad p \times n, \quad n \times \phi$$



$$\begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} [w_1\ w_2 \cdots w_\phi] \qquad \begin{bmatrix} t_{11} & \cdots & t_{1\phi} \\ t_{21} & \cdots & t_{2\phi} \\ \vdots & & \\ t_{n1} & \cdots & t_{n\phi} \end{bmatrix}_{n \times \phi}$$

$$\begin{bmatrix} x_1^T w_1 & x_1^T w_2 & \cdots & x_1^T w_\phi \\ \vdots & & & \\ x_n^T w_1 & x_n^T w_2 & \cdots & \end{bmatrix}_{n \times \phi}$$

$$Tr\{(XW - T)^T (XW - T)\} = \sum_{j=1}^{\phi} \sum_{i=1}^{n} (w_j^T x_i - t_{ij})^2$$

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

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

$$Tr(ZAZ^T) = \sum_{i'} \sum_{j'} \sum_{k'} z_{i'j'} a_{j'k'} z_{i'k'}$$

① $j' = k'$ ② $j' \neq k'$

$$= \sum_{i'} \sum_{j'} z_{i'j'}^2 a_{j'j'} + \sum_{i'} \sum_{j'} \sum_{k': k' \neq j'} z_{i'j'} a_{j'k'} z_{i'k'}$$

$i'=i$ $j'=j$ ① $i'=i$, $j'=j$ ① $i'=i$, $k'=j$

$$\frac{\partial Tr(ZAZ^T)}{\partial z_{ij}} = 2 z_{ij} a_{jj} + \sum_{k': k' \neq j} a_{jk'} z_{ik'} + \sum_{k': k' \neq j} z_{ik'} a_{k'j}$$

$$= \sum_{k'} z_{ik'} (a_{jk'} + a_{k'j})$$

$$\frac{d\, Tr(ZAZ^T)}{dZ} = ZA^T + ZA \qquad \frac{d\, Tr(Z^TAZ)}{dZ} = \frac{d\, Tr(ZAZ^T)}{d\, Z^T} = A.Z + A^T.Z$$

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

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

$$\nabla_W L = \nabla_W Tr\{(XW-T)^T(XW-T)\}$$

$$= \nabla_W \{ Tr\{W^TX^TXW\} - Tr\{W^TX^TT\} - Tr\{T^TXW\} + Tr\{T^TT\}$$

$$= 2X^TXW - 2X^TT$$

$$2X^TXW - 2X^TT = 0$$

$$\Downarrow$$

$$W = (X^TX)^{-1}X^TT$$

$[w_1\, w_2 \dots w_p]$

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

$x_1\ x_2 \cdots x_n \in R^m$

$y_i \in \{1, 2, \cdots, \phi\}$ ϕ-class classification

$x_i : [t_{i1}\ t_{i2} \cdots t_{i\phi}]^T$

$t_{ij} = \begin{cases} 1 & \text{if data } i \text{ is in class } j \\ 0, & \text{otherwise} \end{cases}$

$W = (X^TX)^{-1}X^TT$

$\phi = K$

$n \times \phi$

$\hat{T} = X_{test}W = \begin{bmatrix} \hat{t}_{11} & \hat{t}_{12} & \cdots & \hat{t}_{1\phi} \\ \vdots & & & \\ \vdots & & & \\ \vdots & & & \end{bmatrix}$

$x_{1,test}$ is in $\arg\max_j \hat{t}_{1j}$ - class