CSCI 567: Discussion 3

Linear Algebra II

- Practice Problems 3&4
- Intuition for Least Squares Solution
- Quadratic Form
- Eigenvalues & Eigenvectors
- Practice Problems 1&2

Q3: Given $f(\omega) = \|X\omega - y\|_2^2 + \omega^T M \omega$, $(X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^n, M \in \mathbb{R}^{d \times d}, \text{P.D.})$

Find $\omega^* = \arg\min_{\omega \in \mathbb{R}^d} f(\omega)$

$$
\begin{aligned}
f(\omega) &= \|X\omega - y\|_2^2 + \omega^T M \omega \\
&= (X\omega - y)^T (X\omega - y) + \omega^T M \omega \\
&= (X\omega)^T X\omega + y^T y - y^T X\omega - (X\omega)^T y + \omega^T M \omega \\
&= \omega^T X^T X \omega + y^T y - 2 y^T X \omega + \omega^T M \omega \\
&= \omega^T \underbrace{(X^T X + M)}_{A} \omega - 2 \underbrace{y^T X}_{b^T} \omega + y^T y \\
&= f_1(\omega) - 2 f_2(\omega) + y^T y
\end{aligned}
$$

$A = X^T X + M$,
$b = X^T y$ — (I)

$f_1(\omega) = \omega^T A \omega$. $f_2(\omega) = b^T \omega$.

$$\boxed{\frac{\partial f(\omega)}{\partial \omega} = \boxed{\frac{\partial f_1(\omega)}{\partial \omega}} - 2 \boxed{\frac{\partial f_2(\omega)}{\partial \omega}} = 0.}$$ — (II).

$$b^T w = \sum_{i=1}^{d} w_i b_i$$

1. $\frac{\partial f_2(w)}{\partial w}$:

$$\frac{\partial f_2(w)}{\partial w_k} = b_k$$

$$\Rightarrow \frac{\partial f_2(w)}{\partial w} = \begin{bmatrix} \frac{\partial f_2(w)}{\partial w_1} \\ \vdots \\ \frac{\partial f_2(w)}{\partial w_d} \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_d \end{bmatrix} = \boxed{b} \quad \text{— (iii)}$$

2. $\frac{\partial f_1(w)}{\partial w}$: TWO WAYS:

$$v = Aw$$

$$w^T A w = \sum_i w_i v_i$$

$$v_i = a_i^T w = \sum_j A_{ij} w_j$$

a) $$f_1(w) = \sum_{i=1}^{d} \sum_{j=1}^{d} w_i A_{ij} w_j$$

$$= \sum_{i=1}^{d} w_i^2 A_{ii} + \sum_{i=1, i \neq j}^{d} \sum_{j=1}^{d} w_i A_{ij} w_j$$

$$\frac{\partial f_1(w)}{\partial w_k} = 2 w_k A_{kk} + \sum_{i=1, i \neq k}^{d} w_i A_{ik} + \sum_{j=1, j \neq k}^{d} A_{kj} w_j$$

$$= \sum_{i=1}^{d} A_{ik} w_i + \sum_{j=1}^{d} A_{kj} w_j$$

$$= \underset{\text{column vector}}{(a^k)^T} w + \underset{\text{row vector}}{a_k^T} w$$

$$\Rightarrow \boxed{\frac{\partial f_1(w)}{\partial w}} = \underbrace{\begin{bmatrix} — (a^1)^T — \\ \vdots \\ — (a^d)^T — \end{bmatrix}}_{A^T} w + \underbrace{\begin{bmatrix} — a_1^T — \\ \vdots \\ — a_d^T — \end{bmatrix}}_{A} w$$

$$= \boxed{(A^T + A) w} \quad \text{— (iv)}$$

b) $f_1(w) = g^T(w)\, w \,, \quad g(w) = A^T w.$

$$\boxed{\frac{\partial f_1(w)}{\partial w}} = \left(\boxed{\frac{\partial g(w)}{\partial w}}\right)^T w + \left(\boxed{\frac{\partial w}{\partial w}}\right)^T g(w).$$

$$\boxed{\frac{\partial g(w)}{\partial w}} = \begin{bmatrix} \frac{\partial g(w)}{\partial w_1} & \cdots & \frac{\partial g(w)}{\partial w_d} \end{bmatrix} = \begin{bmatrix} - \left(\frac{\partial g_1(w)}{\partial w}\right)^T - \\ \vdots \\ - \left(\frac{\partial g_d(w)}{\partial w}\right)^T - \end{bmatrix}$$

$$g_k(w) = (a^k)^T w \quad \Rightarrow \frac{\partial g_k(w)}{\partial w} = a^k$$

$$\boxed{\frac{\partial g(w)}{\partial w}} = \begin{bmatrix} - a^{1T} - \\ \vdots \\ - a^{dT} - \end{bmatrix} = \boxed{A^T}$$

$$\boxed{\frac{\partial w}{\partial w}} = \begin{bmatrix} \frac{\partial w_1}{\partial w_1} & \frac{\partial w_1}{\partial w_2} & \cdots & \frac{\partial w_1}{\partial w_d} \\ \frac{\partial w_2}{\partial w_1} & \frac{\partial w_2}{\partial w_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \\ \frac{\partial w_d}{\partial w_1} & \cdots & \cdots & \frac{\partial w_d}{\partial w_d} \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \\ 0 & \cdots & \cdots & 1 \end{bmatrix} = \boxed{I_{d \times d}}$$

$$\boxed{\frac{\partial f_1(w)}{\partial w}} = A w + I^T A^T w = \boxed{(A + A^T)\, w} \quad \text{— (iv)}$$

$$\boxed{\frac{\partial f(w)}{\partial w} = \boxed{\frac{\partial f_1(w)}{\partial w}} - 2\boxed{\frac{\partial f_2(w)}{\partial w}} = 0.}$$

Using (III), (IV) in (II):

$$\frac{\partial f(w)}{\partial w} = \boxed{(A + A^T)\, w} - 2\boxed{b} = 0.$$

$$\Rightarrow \quad w^* = \left(\frac{A + A^T}{2}\right)^{-1} b \qquad \text{(when } A + A^T \text{ is invertible)}$$

Using (II):

$$A = X^T X + M, \quad b = X^T y$$

$$\Rightarrow \quad A + A^T = 2X^T X + M + M^T$$

$$\Rightarrow \quad \boxed{w^* = \left[X^T X + \frac{1}{2}(M + M^T)\right]^{-1} (X^T y).} \qquad \text{(Ans.)}$$

Special case: $M = \lambda I_{d \times d}$

$$\Rightarrow \quad w^* = (X^T X + \lambda I)^{-1} (X^T y).$$

$\lambda = 0$

$$\Rightarrow \quad w^* = (X^T X)^{-1} X^T y.$$

SHORT-ANSWER QUESTION. *The following questions use linear algebra and calculus in ML formulations. They particularly test your knowledge of gradients of multivariate functions.*

Q3 Consider the following optimization problem:

$$\boldsymbol{w}_* = \arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \boldsymbol{w}^T\boldsymbol{M}\boldsymbol{w}$$

Here, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is a positive definite matrix and $\|\cdot\|_2$ stands for the ℓ_2 norm. Find the closed form solution for $\boldsymbol{w}_*$. Proceed in a similar way as how we derived the general least-squares solution in class. (This optimization problem is a generalization of ℓ_2 *regularization*, which we will see in class.)

Answer: Setting the gradient $2\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + (\boldsymbol{M} + \boldsymbol{M}^T)\boldsymbol{w}$ to be $\boldsymbol{0}$ and using the fact that $\boldsymbol{M}$ is invertible gives

$$\boldsymbol{w}_*' = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \frac{\boldsymbol{M} + \boldsymbol{M}^T}{2}\right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$

Q4 Assume we have a training set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, where each outcome y_i is generated by a probabilistic model $\boldsymbol{w}_*^{\mathrm{T}}\boldsymbol{x}_i + \epsilon_i$ with ϵ_i being an independent Gaussian noise with zero-mean and variance σ^2 for some $\sigma > 0$. In other words, the probability of seeing any outcome $y \in \mathbb{R}$ given $\boldsymbol{x}_i \in \mathbb{R}^d$ is

$\varepsilon_i = y_i - w_*^T x_i$

$$\Pr(y \mid \boldsymbol{x}_i; \boldsymbol{w}_*, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y - \boldsymbol{w}_*^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\sigma^2}\right).$$

Assume σ is fixed and given, find the maximum likelihood estimation for $\boldsymbol{w}_*$. In other words, first write down the probability of seeing the outcomes $y_1, \ldots, y_n$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as a function of the value of $\boldsymbol{w}_*$; then find the value of $\boldsymbol{w}_*$ that maximizes this probability. You can assume $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ is invertible, where $\boldsymbol{X}$ is the data matrix with each row corresponding to the features of an example. You may find it helpful to review the steps we took in Lecture 2 to find the maximum likelihood solution for the logistic model.

Answer: The probability of seeing the outcomes $y_1, \ldots, y_n$ given $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ for a linear model $\boldsymbol{w}$ is

$$\mathcal{P}(\boldsymbol{w}) = \prod_{i=1}^n \Pr(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, \sigma) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)^2}{2\sigma^2}\right).$$

Taking the negative log, this becomes

$$F(\boldsymbol{w}) = n\ln\sqrt{2\pi} + n\ln\sigma + \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)^2 = n\ln\sqrt{2\pi} + n\ln\sigma + \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2.$$

$(X\omega - y)_i$
$= x_i^T \omega - y_i$

Maximizing $\mathcal{P}$ is the same as minimizing F, which is clearly the same as just minimizing $\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$, the same objective as for least square regression. Therefore the MLE for $\boldsymbol{w}_*$ is exactly the same as the least square solution:

$$\boldsymbol{w}_* = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}.$$