Written Assignment 0

Due: Thursday 09/18/2025 by 11:59pm EST

Disclaimer

Written assignments must be typeset. The `.tex` files that were used to generate this pdf are included with this assignment: I encourage you to use them when writing your solutions. These questions are **proof** questions, meaning that you will need to write a two-column proof with mathematical steps on the left column and english justification on the right column. Show all of your steps.

Note that you are **NOT** allowed to use any help from LLMs, online solutions, old solutions, etc. when solving these problems. Your solutions are your own. You **are** allowed to chat with your classmates, but not with detail granular enough to copy each others work.

Question 1: Probablity and Statistics Review (20 points)

Let $X_1, X_2, \cdots, X_n$ be n continuous random variables with expectations $\mathbb{E}[X_1], \mathbb{E}[X_2], \cdots, \mathbb{E}[X_n]$ and variances $\mathbb{V}[X_1], \mathbb{V}[X_2], \cdots, \mathbb{V}[X_n]$. For constants $a_1, a_2, \cdots, a_n$ prove the following:

(a) $\mathbb{E}\Big[\sum_{i=1}^{n} a_i X_i\Big] = \sum_{i=1}^{n} a_i \mathbb{E}[X_n]$.

(b) $\mathbb{V}\Big[\sum_{i=1}^{n} a_i X_i\Big] = \sum_{i=1}^{n} a_i^2 \mathbb{V}[X_i] + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j Cov(X_i, X_j)$ where $Cov(X_i, X_j)$ is the covariance of X_i and X_j.

Question 2: MLE Practice (20 points)

Consider N i.i.d scalar samples $\{x_1, x_2, \cdots, x_N\}$ drawn from distribution P. For each distribution, calculate the MLEs of the parameters that control that distribution:

(a) $P = Pr[x; \theta] = \theta e^{-\theta x^2}$ for $x \geq 0$

(b) $P = Pr[x; \theta] = \frac{1}{1-\theta}$ for $\theta \leq x \leq 1$

(c) $P = Pr[x; \alpha, \beta] = \frac{1}{\pi\alpha\left(1+(\frac{x-\beta}{\alpha})^2\right)}$. For this problem, don't actually solve for α, β. Instead just simplify as much as you can.

You may assume that any value x outside the support of P has probability zero.

Question 3: Naive Bayes MLE (20 points)

Consider a binary dataset $D = \{(\vec{x}^{(i)}, y^{(i)})\}_{i=1}^{M}$ where each $\vec{x}^{(i)} \in \mathbb{B}^n$ and each $y^{(i)} \in \mathbb{B}$ drawn from random variables $X_1, X_2, \cdots, X_n, Y$. Let us define function $c(y)$ that counts the number of occurances that label y appears in D:

$$c(y) = \sum_{(\vec{x}^{(i)}, y^{(i)}) \in D} \mathbb{I}[y^{(i)} = y]$$

where $\mathbb{I}[\cdot]$ is the indicator function. Let us define function $c(j, y)$ that counts the number of occurances that label y appears in D *and* the corresponding observation $\vec{x}_i^{(j)} = 1$ (e.g. the j^{th} entry of observation $\vec{x}^{(i)}$ is 1:

$$c(j, y) = \sum_{(\vec{x}^{(i)}, y^{(i)}) \in D} \mathbb{I}[y^{(i)} = y, \vec{x}_j^{(i)} = 1]$$

Let us define paramter b as $Pr[Y = 1]$ and one parameter b^{jy} for each $1 \leq j \leq n$ as $Pr[X_j = 1 | Y = y]$. Prove that the following estimators are MLE for these parameters:

(a) $\hat{b}_{MLE} = \frac{c(1)}{|D|}$

(b) $\hat{b^{jy}}_{MLE} = \frac{c(j,y)}{c(y)}$

Question 4: Ridge and Lasso Regression (20 points)

Consider a fixed data matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ where each row $\vec{x}^{(i)} \in \mathbb{R}^d$. We also have access to a ground truth vector $\vec{y} \in \mathbb{R}^m$, where each scalar $y^{(i)}$ is the ground truth for observation $\vec{x}^{(i)}$. Let us assume a linear model $y^{(i)} = \vec{w}^T \vec{x}^{(i)} + \epsilon^{(i)}$ where $\vec{w} \in \mathbb{R}^d$ and $\epsilon^{(i)} \in \mathbb{R}$ are unknown.

We want to figure out the value of $\vec{w}$ that minimizes the squared error to our ground truth:

$$\vec{w}^* = \arg\min_{\vec{w} \in \mathbb{R}^d} \mathcal{L}(\vec{w}) = \arg\min_{\vec{w} \in \mathbb{R}^d} \sum_{i=1}^{m} \left(y^{(i)} - \vec{w}^T \vec{x}^{(i)} \right)^2$$

However, we don't want just **any** solution. Instead, we have some preferences regarding what solutions we want our optimization algorithm to find. For instance, if we define solutions with small values to be better than solutions with large values, we would change our objective $\mathcal{L}$ accordingly:

$$\vec{w}^* = \arg\min_{\vec{w} \in \mathbb{R}^d} \mathcal{L}(\vec{w}, \lambda_r) = \arg\min_{\vec{w} \in \mathbb{R}^d} \left(\sum_{i=1}^{m} \left(y^{(i)} - \vec{w}^T \vec{x}^{(i)} \right)^2 \right) + \lambda_r ||\vec{w}||_2^2$$

where $\lambda_r \in \mathbb{R}^{>0}$ is a hyperparameter to control the weight of the l2 term relative to the squared error. This is called **Ridge Regression**.

If we instead want to try and get the sparsest solution, we would change our objective $\mathcal{L}$ accordingly:

$$\vec{w}^* = \arg\min_{\vec{w} \in \mathbb{R}^d} \mathcal{L}(\vec{w}, \lambda_l) = \arg\min_{\vec{w} \in \mathbb{R}^d} \left(\sum_{i=1}^{m} \left(y^{(i)} - \vec{w}^T \vec{x}^{(i)} \right)^2 \right) + \lambda_l ||\vec{w}||_1$$

where $\lambda_l \in \mathbb{R}^{>0}$ is a hyperparameter to control the weight of the l1 term relative to the squared error. This is called **Lasso Regression**.

With these two regression variants, please answer the following questions:

(a) Derive a closed form expression for $\vec{w}^*$ for Ridge Regression. Your expression should be a function of $\mathbf{X}, \vec{y}$, and λ_r.

(b) Prove that Lasso Regression is not differentiable at some points.

(c) Prove that the maximum value for λ_l is $2||\mathbf{X}^T \vec{y}||_\infty$ (i.e. for any $\lambda_l \geq 2||\mathbf{X}^T \vec{y}||_\infty$, the solution $\vec{w}^*$ is guaranteed to be the zero vector). Prove this using the following three steps:

(i) The one-sided directional derivative of function $f(x)$ in the direction $\vec{u} \in \mathbb{R}^d$ is defined as follows:

$$f'(x; \vec{u}) = \lim_{h \to 0^+} \frac{f(x + h\vec{u}) - f(x)}{h}$$

where $\vec{u}$ is a unit vector. Compute the one-sided directional derivative of our objective function $\mathcal{L}(\vec{0}, \lambda_l)$ in direction $\vec{u}$. Your expression $\mathcal{L}'(\vec{0}, \lambda_l; \vec{u})$ should be a function of $\mathbf{X}, \vec{y}, \lambda_l$, and $\vec{u}$.

(ii) Show that for any $\vec{u} \neq \vec{0}$, $\mathcal{L}'(\vec{0}, \lambda_l; \vec{u}) \geq 0$ iff $\lambda_l \geq C$ where C is some constant that depends on $\mathbf{X}, \vec{y}$, and $\vec{u}$. Find the expression for C.

(iii) Due to the convexity of Lasso regression, $\vec{w}^*$ is a minimizer of $\mathcal{L}(\vec{w}, \lambda_l)$ iff $\forall \vec{u} \neq \vec{0}$, $\mathcal{L}'(\vec{w}^*, \lambda_l; \vec{u}) \geq 0$. Show that $\vec{w} = \vec{0}$ is the minimizer of $\mathcal{L}(\vec{w}, \lambda_l)$ iff $\lambda_l \geq 2||\mathbf{X}^T \vec{y}||_\infty$.

Question 5: Matrix Derivatives and Multi-Target Linear Regression (20 points)

Consider a fixed data matrix $\mathbf{X} \in \mathbb{R}^{m \times d}$ where each row $\vec{x}^{(i)} \in \mathbb{R}^d$. We also have access to a ground truth matrix $\mathbf{Y} \in \mathbb{R}^{m \times k}$, where each row $\vec{y}^{(i)} \in \mathbb{R}^k$ is the ground truth for observation $\vec{x}^{(i)}$. Let us assume a linear model $\mathbf{Y} = \mathbf{XW} + \epsilon$ where $\mathbf{W} \in \mathbb{R}^{d \times k}$ is unknown and each $\epsilon_j^{(i)} \in \mathbb{R}$ is drawn from a normal distribution with mean 0 and variance σ. Since each $\epsilon_j^{(i)}$ is drawn from the same distribution, the matrix ϵ is drawn from a multivariate gaussian $\mathcal{N}(\vec{0}, \sigma\mathbf{I})$ with mean vector $\vec{0}$ and covariance $\sigma\mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

Rearranging the linear model, we can see

$$\begin{aligned} \mathbf{Y} &= \mathbf{XW} + \epsilon && \text{the model} \\ \mathbf{XW} - \mathbf{Y} &= \epsilon && \text{algebra} \\ \mathbf{XW} - \mathbf{Y} &\sim \mathcal{N}(\vec{0}, \sigma\mathbf{I}) && \epsilon \sim \mathcal{N}(\vec{0}, \sigma\mathbf{I}) \rightarrow \mathbf{XW} - \mathbf{Y} \sim \mathcal{N}(\vec{0}, \sigma\mathbf{I}) \end{aligned}$$

then if we wish to find the value of $\mathbf{W}$, we will want to maximize this probability:

$$\mathbf{W}^* = \underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\arg\max}\, \mathcal{L}(\mathbf{W}) = \underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\arg\max} \log Pr[\mathbf{Y}|\mathbf{X}; \mathbf{W}] = \underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\arg\max} \log \mathcal{N}(\mathbf{Y} - \mathbf{XW}; \vec{0}, \sigma\mathbf{I})$$

Using this setup, please show the following:

(a) Show that $\underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\arg\max}\, \mathcal{L}(\mathbf{W}) \rightarrow \underset{\mathbf{W} \in \mathbb{R}^{d \times k}}{\arg\min}\, ||\mathbf{Y} - \mathbf{XW}||_F^2$

(b) Show that $\frac{\partial ||\mathbf{Y} - \mathbf{XW}||_F^2}{\partial} = -2\mathbf{X}^T(\mathbf{Y} - \mathbf{XW})$

(c) Show that $\mathbf{W}^*_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$