ASSIGNMENT 1A

CS 3410: INTRODUCTION TO MACHINE LEARNING

KRISHNA PRANEETH SIDDE
AUID: 1020231796

1. Problem 1

(i) For a random variable $X \sim P(\lambda)$ with the pmf

$$\mathbb{P}(x|\lambda) = \frac{\lambda^x e^{-\lambda}}{x!},$$

where $\lambda > 0$, comparing this with the exponential family form

$$\mathbb{P}(x|\eta) = b(x)\exp(\eta \cdot x - A(\eta)),$$

we get that

$$\begin{aligned} b(x) &= \frac{1}{x!} \\ \eta &= \ln \lambda \\ A(\eta) &= e^{\eta}. \end{aligned}$$

(ii) (a) For $X \sim P(\lambda)$, let us first compute $\mathbb{E}[X]$.

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda. \end{aligned}$$

Hence, we verify that

$$A'(\eta) = \frac{d}{d\eta}\left(e^{\eta}\right) = e^{\eta} = e^{\ln \lambda} = \lambda = \mathbb{E}[X]$$

indeed stands true.

Date: February 27, 2025.

(b) Similarly, for $Var[X]$,

$$\begin{aligned}
Var[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\
&= \left(\sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!}\right) - \lambda^2 \\
&= \left(\sum_{y=0}^{\infty} (y+1)^2 e^{-\lambda} \frac{\lambda^{(y+1)}}{(y+1)!}\right) - \lambda^2 \\
&= \lambda \left(\sum_{y=0}^{\infty} (y+1) \frac{e^{-\lambda}\lambda^y}{y!}\right) - \lambda^2 \\
&= \lambda \left(\sum_{y=0}^{\infty} y \cdot \frac{\lambda^y e^{-\lambda}}{y!} + \sum_{y=0}^{\infty} \frac{\lambda^y e^{-\lambda}}{y!}\right) - \lambda^2 \\
&= \lambda(\lambda + 1) - \lambda^2 \\
&= \lambda.
\end{aligned}$$

Hence, we can verify that

$$A''(\eta) = \frac{d}{d\eta}\left(\frac{d}{d\eta} e^{\eta}\right) = e^{\eta} = e^{\ln \lambda} = \lambda = Var[X]$$

is also true.

2. Problem 2

(a) Claim: $\mathbb{E}[Z] = \omega^T \mu$

Proof.

$$\begin{aligned}
\mathbb{E}[Z] &= \mathbb{E}\left[\sum_{i=1}^{N} \omega_i X_i\right] \\
&= \sum_{i=1}^{N} \omega_i \mathbb{E}[X_i] \\
&= \sum_{i=1}^{N} \omega_i \mu_i \\
&= \omega^T \mu,
\end{aligned}$$

where

$$\omega = \begin{bmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_N \end{bmatrix}, \text{ and } \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_N \end{bmatrix}.$$

■

(b) Claim: $Var[Z] = \omega^T \Sigma \omega$

Proof.

$$\begin{aligned}
Var[Z] &= \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2 \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N} \omega_i X_i\right)^2\right] - (\omega^T\mu)^T(\omega^T\mu) \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i\omega_j X_i X_j\right] - (\omega^T\mu)^T(\omega^T\mu) \\
&= \sum_{i=1}^{N}\sum_{j=1}^{N} \omega_i\omega_j \mathbb{E}[X_i X_j] - (\omega^T\mu)^T(\omega^T\mu) \\
&= \omega^T(\Sigma + \mu\mu^T)\omega - (\omega^T\mu)^T(\omega^T\mu) \\
&= \omega^T\Sigma\omega + \omega^T\mu\mu^T\omega - \omega^T\mu\mu^T\omega \\
&= \omega^T\Sigma\omega.
\end{aligned}$$

■

(c) Using the fact that $Cov(X_1, X_2) = \mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]$, and the Cauchy-Schwarz inequality, we can observe that

$$\begin{aligned}
|\rho_{X_1,X_2}| &= \left|\frac{Cov(X_1, X_2)}{\sigma_{X_1}\sigma_{X_2}}\right| \\
&= \frac{|\mathbb{E}[(X_1 - \mathbb{E}[X_1])(X_2 - \mathbb{E}[X_2])]|}{\sigma_{X_1}\sigma_{X_2}} \\
&\leq \frac{\sqrt{\mathbb{E}[(X_1 - \mathbb{E}[X_1])^2] \cdot \mathbb{E}[(X_2 - \mathbb{E}[X_2])^2]}}{\sigma_{X_1}\sigma_{X_2}} \\
&\leq \frac{\sqrt{\sigma_{X_1}^2 \cdot \sigma_{X_2}^2}}{\sigma_{X_1}\sigma_{X_2}} \\
&\leq 1.
\end{aligned}$$

Therefore, we get that

$$-1 \leq \rho_{X_1,X_2} \leq 1.$$

(d)

$$\begin{aligned}
I &= \int_{-\infty}^{\infty} x f(x)\, dx \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} x \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\, dx.
\end{aligned}$$

Suppose $z = (x - \mu)/\sigma$. Then, $dz = dx/\sigma$ and $x = \mu + \sigma z$. Hence,

$$
\begin{aligned}
I &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} (\mu + \sigma z)\, exp\left(-\frac{z^2}{2}\right)\, \sigma dz \\
&= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \left[\mu \int_{-\infty}^{\infty} exp\left(-\frac{z^2}{2}\right)\, dz + \sigma \int_{-\infty}^{\infty} z\, exp\left(-\frac{z^2}{2}\right)\, dz\right] \\
&= \frac{\sigma}{\sqrt{2\pi\sigma^2}} \cdot \mu \cdot \sqrt{2\pi} \\
&= \mu.
\end{aligned}
$$

3. Problem 3

(a) Let us first start with expressing the training set $\{(x^{(i)}, y^{(i)}), 1 \leq i \leq m\}$ where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathbb{R}^p$ using matrices. We can observe that

$$
X = \begin{bmatrix} (x^{(1)})^T \\ (x^{(2)})^T \\ \vdots \\ (x^{(m)})^T \end{bmatrix} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix},
$$

and

$$
Y = \begin{bmatrix} (y^{(1)})^T \\ (y^{(2)})^T \\ \vdots \\ (y^{(m)})^T \end{bmatrix} = \begin{bmatrix} y_1^{(1)} & y_2^{(1)} & \cdots & y_p^{(1)} \\ y_1^{(2)} & y_2^{(2)} & \cdots & y_p^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(m)} & y_2^{(m)} & \cdots & y_p^{(m)} \end{bmatrix},
$$

can be used to denote the training data in terms of matrices. Here, $X \in \mathbb{R}^{m \times n}$ and $Y \in \mathbb{R}^{m \times p}$. Given a parameter matrix $\Theta \in \mathbb{R}^{n \times p}$, we can now express $J(\Theta)$ as

$$
\begin{aligned}
J(\Theta) &= \frac{1}{2} (X\Theta - Y)^2 \\
&= \frac{1}{2} (X\Theta - Y)^T (X\Theta - Y).
\end{aligned}
$$

(b)

$$
\begin{aligned}
\nabla_\Theta J(\Theta) &= \nabla_\Theta \frac{1}{2} (X\Theta - Y)^T (X\Theta - Y) \\
&= \frac{1}{2} \nabla_\Theta ((X\Theta)^T X\Theta - (X\Theta)^T Y - Y^T (X\Theta) + Y^T Y) \\
&= \frac{1}{2} \nabla_\Theta ((\Theta^T X^T)(X\Theta) - Y^T (X\Theta) - Y^T (X\Theta)) \\
&= \frac{1}{2} \left[\nabla_\Theta \left(\Theta^T (X^T X)\Theta\right) - \nabla_\Theta \left(2(X^T Y)^T \Theta\right)\right] \\
&= \frac{1}{2} \left[2X^T X\Theta - 2X^T Y\right] \\
&= X^T X\Theta - X^T Y.
\end{aligned}
$$

Hence, $\nabla_\Theta J(\Theta) = 0$ implies that

$$X^T X\Theta = X^T Y$$
$$\implies \Theta = (X^T X)^{-1} X^T Y.$$

(c) Now, if we were to compute $y_j^{(i)}$ separately for each $1 \leq j \leq p$ using p individual linear models of the form $\theta_j \in \mathbb{R}^n$, each of the independent linear models would be computing $X\theta_j$ and the cost function would be comparing them with respective y_j. Hence, we get that $\theta_j = (X^T X)^{-1} X^T y_j$ from our solution above. However, our multivariate solution gives a parameter matrix Θ, which is simply a concatenation of all the θ_j's in each row of Θ. In the approach of considering j linearized models, we are performing an evaluation over every single θ_j and concatenating them in order to get Θ, which can just be computed from our closed-form equation given above. Hence, both of these values essentially compare to the same matrix Θ, but our multivariate solution is faster in evaluating Θ.