Department of Computer Science
Ashoka University

Introduction to Machine Learning

Assignment 1

Collaborators: None **Name: Rushil Gupta**

Question 1

Question 1.1

Since we want λ to be the coefficient of x, we have:

$$\eta = \log \lambda \implies \lambda = e^{\eta} \implies e^{\eta \cdot x} = e^{\log \lambda^{x}} = \lambda^{x}$$

So, we want the remaining $e^{-\lambda}$ term to be $e^{-e^{\eta}}$. So, we get:

$$A(\eta) = e^{\eta}$$

So, this gives us:

$$p(x \mid \lambda) = b(x) \cdot e^{\eta \cdot x - A(\eta)} = b(x)\lambda^{x} e^{-\lambda}$$

Clearly, setting $b(x) = \frac{1}{x!}$, we get the Poisson distribution in exponential family form. So, we have:

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

Question 1.2

It is known that:

$$E[X] = A'(\eta) \quad \text{and} \quad \mathrm{Var}(X) = A''(\eta).$$

Part (a)

We have $A(\eta) = e^{\eta}$. So, we get:

$$A'(\eta) = e^{\eta} = \lambda$$

Also, since we know $p(x)$ is a Poisson distribution, we know that $E[X] = \lambda$. Hence, correct.

Part (b)

We have $A'(\eta) = e^{\eta} = \lambda$. So, we get:

$$A''(\eta) = e^{\eta} = \lambda$$

Again, since $p(x)$ is Poisson, we know that $\mathrm{Var}(X) = \lambda$. Hence, correct.

Question 2

Question 2.a

We know that $Z = w^T X$. So, we have:

$$E[Z] = E[w^T X] = E\left[\sum_{i=1}^{N} w_i X_i\right]$$

By linearity of expectation, we get:

$$E[Z] = \sum_{i=1}^{N} w_i E[X_i] = \sum_{i=1}^{N} w_i \mu_i = w^T \mu$$

So, $E[Z] = w^T \mu$.

Question 2.b

Since we know $Z = w^T X$, we have:

$$\text{Var}(Z) = \text{Var}(w^T X) = \text{Var}\left(\sum_{i=1}^{N} w_i X_i\right)$$

Since X_i are not independent, but we know the covariance matrix of $X = \Sigma$, we get:

$$\text{Var}(Z) = \sum_{i=1}^{N} w_i^2 \,\text{Var}(X_i) + 2\sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j \,\text{Cov}(X_i, X_j)$$

Now, we know $Var(X_i) = \Sigma_{i,i}$ and $\text{Cov}(X_i, X_j) = \Sigma_{i,j}$. So, we get:

$$\text{Var}(Z) = \sum_{i=1}^{N} w_i^2 \Sigma_{i,i} + 2\sum_{i=1}^{N} \sum_{j=i+1}^{N} w_i w_j \Sigma_{i,j} = \sum_{i=1}^{N} \sum_{j=1}^{N} w_i \Sigma_{i,j} w_j = w^T \Sigma w$$

So, $\text{Var}(Z) = w^T \Sigma w$.

Question 2.c

We know that the correlation coefficient between X_1 and X_2 is given by:

$$\rho_{X_1,X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_{X_1} \sigma_{X_2}}$$

By the Cauchy-Schwarz inequality, we have:

$$|\,\text{Cov}(X_1, X_2)| \leq \sqrt{\text{Var}(X_1)\,\text{Var}(X_2)} = \sigma_{X_1} \sigma_{X_2}$$

This gives:

$$\frac{|\,\text{Cov}(X_1, X_2)|}{\sigma_{X_1} \sigma_{X_2}} = |\rho_{X_1,X_2}| \leq 1 \implies -1 \leq \rho_{X_1,X_2} \leq 1$$

Question 2.d

We want to evaluate:

$$\int_{-\infty}^{\infty} xf(x)dx$$

where:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

We will use u-substitution to solve this integral. Let:

$$u = \frac{x-\mu}{\sigma} \implies du = \frac{dx}{\sigma} \implies dx = \sigma du$$

Now, after substitution, we get:

$$\begin{aligned}\int_{-\infty}^{\infty} (\sigma u + \mu)\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{u^2}{2}\right)\sigma du &= \int_{-\infty}^{\infty} (\sigma u + \mu)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right) du \\ &= \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \exp\left(-\frac{u^2}{2}\right) du + \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du\end{aligned}$$

We know that:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{u^2}{2}\right) du = \sqrt{2\pi}$$

So, out final integral looks like:

$$\frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} u \exp\left(-\frac{u^2}{2}\right) du + \mu$$

Now, look at the other integral:

$$\int u \exp\left(-\frac{u^2}{2}\right) du$$

Since $\exp\left(-\frac{u^2}{2}\right)$ is an even function, u is an odd function, and we have symmetric bounds, we get:

$$\int_{-\infty}^{\infty} u \exp\left(-\frac{u^2}{2}\right) du = 0$$

So, our final answer is:

$$\int_{-\infty}^{\infty} xf(x)dx = \mu$$

Question 3

Question 3.a

First, let's look at what a prediction $\hat{y}$ is. We have:

$$\hat{y}^{(i)} = \Theta^T x^{(i)}$$

So, we can then construction a matrix X with columns as $x^{(i)}$ and a matrix Y with columns as $y^{(i)}$. Observe that $X \in \mathbb{R}^{n \times m}$ and $Y \in \mathbb{R}^{p \times m}$. So, we can predictions $\hat{Y}$ as:

$$\hat{Y} = \Theta^T X$$

Now, we know our earlier definition of the cost function was the sum of all squared errors component-wise in each vectors, for all data points. That is equivalent to taking the square of the differences of $\hat{Y} = \Theta^T X$ and Y and summing them up. So, we have:

$$J(\Theta) = \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{p} \left((\Theta^T X)_{ij} - Y_{ij} \right)^2$$

Now, if we say $A = \Theta^T X - Y$, we know that the cost function is the Frobenius norm of A squared. So, we can write:

$$J(\Theta) = \frac{1}{2} \|A\|_F^2 = \frac{1}{2} \|\Theta^T X - Y\|_F^2$$

Lastly, since we know $\|A\|_F^2 = \operatorname{tr}(A^T A)$, we can write:

$$J(\Theta) = \frac{1}{2} \operatorname{tr} \left((\Theta^T X - Y)^T (\Theta^T X - Y) \right)$$

Question 3.b

Expand the term inside the trace:

$$(X\Theta - Y)^T (X\Theta - Y) = \Theta^T X^T X \Theta - \Theta^T X^T Y - Y^T X \Theta + Y^T Y$$

Thus, the cost function becomes:

$$J(\Theta) = \frac{1}{2} \operatorname{tr} \left(\Theta^T X^T X \Theta \right) - \operatorname{tr} \left(Y^T X \Theta \right) + \frac{1}{2} \operatorname{tr} \left(Y^T Y \right)$$

Since the last term, $\operatorname{tr}(Y^T Y)$, does not depend on Θ, it can be ignored when finding the minimizer. To find the minimum, we differentiate $J(\Theta)$ with respect to Θ:

$$\begin{aligned}
\nabla_\Theta \left(\frac{1}{2} \operatorname{tr} \left(\Theta^T X^T X \Theta \right) \right) &= \frac{1}{2} \operatorname{tr} \left(\nabla_\Theta \left(\Theta^T X^T X \Theta \right) \right) \\
&= X^T X \Theta \\
\nabla_\Theta \left(\operatorname{tr} \left(Y^T X \Theta \right) \right) &= Y^T X
\end{aligned}$$

This gives us:

$$\nabla_\Theta J(\Theta) = X^T X \Theta - X^T Y$$

We know at the minimum, the gradient is zero. So, we get:

$$\begin{aligned}
X^T X \Theta &= X^T Y \\
\Theta &= \left(X^T X \right)^{-1} \left(X^T Y \right)
\end{aligned}$$

Question 3.c

Consider the case where we solve for each output dimension separately. We would consider p independent linear models:

$$y_j^{(i)} = \theta_j^T x^{(i)}, \quad j = 1, \ldots, p,$$

Here, with little inspection, we see that $\Theta_j \in \mathbb{R}^n$ is the parameter vector for the j-th output dimension. The closed-form solution for each θ_j is given by:

$$\theta_j = (X^T X)^{-1} X^T y_j$$

where y_j is the vector of the j-th output values over all samples. So now, our final predictor will be p such predictors, and we can collect them into a matrix $\Theta \in \mathbb{R}^{n \times p}$:

$$\Theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{bmatrix} = \theta_{:,j}$$

This is identical to the multivariate solution we saw earlier.