Theoretical (50)

1 *(6 points)*

Part1: (c) A high magnitude suggests that the feature is important. However, it may be the case that another feature is highly correlated with this feature and it's coefficient also has a high magnitude with the opposite sign, in effect cancelling out the effect of the former. Thus, we cannot really remark on the importance of a feature just because it's coefficient has a relatively large magnitude.

Part2:

- True , we can not reduce the bias and cover our model's limits by increasing the training data.
- False , the obvious counterexample that can be mentioned is overfiting on training data.
- False , increasing model complexity could be useful for this case , but not always.
- False

2 *(19 points)*

2.a:

$$
\begin{gathered}
J(w) = \sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})^2 \\
\nabla_w J(w) = -2\sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})(x^{(i)})^T = 0 \\
\sum_{i=1}^{n} y^{(i)}(x^{(i)})^T = w^T \sum_{i=1}^{n} x^{(i)}(x^{(i)})^T \\
x^T y = (x^T x)w \\
w^* = (x^T x)^{-1} x^T y
\end{gathered}
$$

2.b:

$$
\begin{gathered}
J(w) = \sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})^2 + \lambda w^T w \\
\nabla_w J(w) = -2\sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})(x^{(i)})^T + 2\lambda \sum_{j=1}^{M} w_j = 0 \\
\sum_{i=1}^{n} y^{(i)}(x^{(i)})^T = w^T \sum_{i=1}^{n} x^{(i)}(x^{(i)})^T + \lambda \sum_{j=1}^{M} w_j \\
x^T y = (x^T x)w + \lambda I w \\
x^T y = (x^T x + \lambda I)w \\
w^* = (x^T x + \lambda I)^{-1} x^T y
\end{gathered}
$$

2.c:
First, we should prove that if $\Sigma X = XF$ then $w^*_{new} = w^*$

$$
\begin{aligned}
\Sigma X = XF \Rightarrow \Sigma = XFX^{-1} \Rightarrow \Sigma^{-1} = XF^{-1}X^{-1} \\
\Rightarrow \Sigma^{-T} = X^{-T}F^{-T}X^T \Rightarrow \Sigma^{-1} = X^{-T}F^{-T}X^T \\
w^*_{new} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y = (X^T X^{-T}F^{-T}X^T X)^{-1}X^T X^{-T}F^{-T}X^T y \\
= (X^T X)^{-1}F^T F^{-T}X^T y = (X^T X)^{-1}X^T y = w^*
\end{aligned}
$$

The first side of the term was proved.
Then, we should prove that if $w^*_{new} = w^*$ then $\Sigma X = XF$

$$
\begin{aligned}
w^* = (X^T X)^{-1}X^T y = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}y \Rightarrow (X^T X)^{-1}X^T = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1} \\
X^T\Sigma = (X^T X)(X^T\Sigma^{-1}X)^{-1}X^T \Rightarrow \Sigma X = X[(X^T X)(X^T\Sigma^{-1}X)^{-1}]^T
\end{aligned}
$$

now name $[(X^T X)(X^T\Sigma^{-1}X)^{-1}]^T$ as F then we conclud the other side so this estimator is equal to ordinary least-square estimator w^*_{opt}, if and only if there exist a nonsingular matrix F, such that $\Sigma X = XF$.

2.d:
L1 regularized loss function for Y' and X' is:

$$L = |y' - X'w|^2 + \lambda_2 \|w\|_1$$

now we need to define some Y' and X' which gives that:

$$|y' - X'w|^2 + \lambda_2 \|w\|_1 = |y - Xw|^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \|w\|_1$$

now we define X' and Y' as follows:

$$\text{X'} = \begin{bmatrix} \sqrt{\lambda_1} & & \cdot & \cdot & \cdot & \\ & \sqrt{\lambda_1} & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ \cdot & \cdot & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \sqrt{\lambda_1} \\ & & X & & & \end{bmatrix}_{m \times m} \Rightarrow \text{X' w} = \begin{bmatrix} \sqrt{\lambda_1} w_1 \\ \sqrt{\lambda_1} w_2 \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{\lambda_1} w_n \end{bmatrix}$$

$$\text{Y'} = \begin{bmatrix} 0 \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ Y \end{bmatrix}$$

By replacing the equalities for X' and Y' we can conclude that $|y - Xw|^2 + \lambda_1 \|w\|_2^2 + \lambda_2 \|w\|_1$ is equivalent to L1 regularized loss function by adding some data points.

3 *(10 points)*

The Kullback-Leiblerdivergence takes the form

$$KL(P||Q) = -\int p(x) \ln q(x) dx + \int p(x) \ln p(x) dx$$

Substituting the Gaussian for $q(x)$ we obtain

$$\begin{aligned} KL(P||Q) &= -\int p(x) \{-\frac{1}{2} \ln |\Sigma| - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\} dx + \text{const} \\ &= \frac{1}{2}\{\ln |\Sigma| + Tr(\Sigma^{-1} E[(x-\mu)(x-\mu)^T])\} + \text{const} \\ &= \frac{1}{2}\{\ln |\Sigma| + \mu^T \Sigma^{-1} \mu - 2\mu^T \Sigma^{-1} E[x] + Tr(\Sigma^{-1} E[xx^T])\} + \text{const} \end{aligned} \tag{1}$$

Differentiating this w.r.t. μ, using results from Appendix C (from Bishop book), and setting the result to zero, we see that

$$\mu = E[x]. \tag{2}$$

And with the use of equalitiy below

$$\begin{aligned} cov[x, y] &= E_{x,y}[\{x - E[x]\}\{y^T - E[y^T]\}] \\ &= E_{x,y}[xy^T] - E[x]E[y^T] \end{aligned} \tag{3}$$

We can say that, similarly, differentiating the equality number (1) w.r.t. Σ^{-1}, again using results from Appendix C (from Bishop book) and also making use of equalities number (2) and number (3),we see that

$$\Sigma = E[xx^T] - \mu\mu^T = cov[x]$$

4 *(15 points)*

4.a

In this particular case, it looks like our data matrix is equal to x_j. So according to the linear regression relation we have equality below:

$$w_j = (x_j^T x_j)^{-1} x_j y = \frac{x_j^T y}{x_j^T x_j}$$

4.b

Note that the columns are orthogonal therefore their internal multiplication is zero. As a result, the value of $X^T X$ will be diagonal. More precisely:

$$X^T X = diag(x_1^T x_1, ..., x_m^T x_m) \Rightarrow (X^T X)^{-1} = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1})$$

now we have:

$$\begin{gathered} w = (X^T X)^{-1} X^T y = diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y \Rightarrow \\ w_j = (diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}) X^T y)_j = (diag((x_1^T x_1)^{-1}, ..., (x_m^T x_m)^{-1}))_j (X^T y)_j \\ = (x_j^T x_j)^{-1} (X^T y)_j = \frac{(X^T y)_j}{x_j^T x_j} = \frac{(X^T)_j y}{x_j^T x_j} = \frac{x_j^T y}{x_j^T x_j} \end{gathered}$$

Combining the recent equality with the previous part, results to concluding that öptimal parameters from training the regressor on all features is the same as the optimal parameters resulting from training on each feature independently".

4.c

Note that in the matrix mode our data is as follows:

$$\mathrm{X} = \begin{bmatrix} & 1 \\ & 1 \\ x_j & . \\ & . \\ & . \\ & 1 \end{bmatrix}$$

That x_j is a column. For the internal multiplication of $(x_j)(1, ..., 1)$, which is equal to the sum of the x_j elements. We use the $sum(x_j)$ symbol. Now with this in mind we have:

$$X^T X = \begin{bmatrix} x_j^T x_j & sum(x_j) \\ sum(x_j) & n \end{bmatrix} \Rightarrow (\mathrm{X}^T X)^{-1} = \frac{1}{n\|x_j\|^2 - sum(x_j)^2} \begin{bmatrix} n & -sum(x_j) \\ -sum(x_j) & x_j^T x_j \end{bmatrix}$$

Note that:

$$X^T y = (\begin{bmatrix} & 1 \\ & 1 \\ x_j & . \\ & . \\ & . \\ & 1 \end{bmatrix})^T y = \begin{bmatrix} x_j^T y \\ sum(y) \end{bmatrix}$$

Then:

$$[w_j, w_0] = w = (X^T X)^{-1} X^T y = \frac{1}{n\|x_j\|^2 - sum(x_j)^2} \begin{bmatrix} n & -sum(x_j) \\ -sum(x_j) & x_j^T x_j \end{bmatrix} \begin{bmatrix} x_j^T y \\ sum(y) \end{bmatrix}$$

$$w_j = \frac{n x_j^T y - sum(x_j) sum(y)}{n \|x_j\|^2 - sum(x_j)^2} = \frac{\frac{x_j^T y}{n} - \frac{sum(x_j)}{n}\frac{sum(y)}{n}}{\frac{\|x_j\|^2}{n} - (\frac{sum(x_j)}{n})^2} = \frac{\mathbb{E}[x_j y] - \mathbb{E}[x_j]\mathbb{E}[y]}{\mathbb{E}[x_j^2] - \mathbb{E}[x_j]^2} = \frac{cov(x_j, y)}{var(x_j)}$$

$$\begin{aligned} w_0 &= \frac{sum(y)\|x_j\|^2 - sum(x_j)(x_j^T y)}{n\|x_j\|^2 - sum(x_j)^2} = \frac{sum(y)\|x_j\|^2 - sum(x_j)(x_j^T y)}{n\|x_j\|^2 - sum(x_j)^2} = \frac{n^2 \mathbb{E}[y]\mathbb{E}[x_j^2] - n^2 \mathbb{E}[x_j]\mathbb{E}[x_j y]}{n^2 var(x_j)} \\ &= \frac{\mathbb{E}[y]\mathbb{E}[x_j^2] - \mathbb{E}[x_j]\mathbb{E}[x_j y]}{var(x_j)} = \frac{\mathbb{E}[y](\mathbb{E}[x_j^2] - \mathbb{E}[x_j]^2) + \mathbb{E}[y]\mathbb{E}[x_j]^2 - \mathbb{E}[x_j]\mathbb{E}[x_j y]}{var(x_j)} = \\ &= \mathbb{E}[y] + \frac{\mathbb{E}[y]\mathbb{E}[x_j]^2 - \mathbb{E}[x_j]\mathbb{E}[x_j y]}{var(x_j)} = \mathbb{E}[y] + \mathbb{E}[x_j]\frac{\mathbb{E}[y]\mathbb{E}[x_j] - \mathbb{E}[x_j y]}{var(x_j)} = \mathbb{E}[y] - w_j \mathbb{E}[x_j] \end{aligned}$$