Assignment 1

Pooria Azarakhsh (Audit Student)

October 17, 2025

Note: Although I have attempted to verify the answers, I do not have a CE or CS background. Therefore, please cross-check the answers with other resources.

Exercise 1

(a) Softmax is defined as follows:

$$\sigma(\mathbf{z})_i = \frac{\exp(z_i)}{\sum_{j=1}^{K} \exp(z_j)} \quad \text{for } i = 1, \ldots, K$$

Softmax is used because of the following reasons:

1. It normalizes the output between $(0, 1)$, so it produces a valid probability distribution.
2. It is differentiable and it has a unique form of derivative.
$$\frac{\partial \sigma\left(z_i\right)}{\partial z_i} = \sigma\left(z_i\right)\left(1 - \sigma\left(z_i\right)\right)$$
3. It highlights larger values and exponential function amplifies difference between logits.
4. It works well with cross-entropy loss and their combination has numerically stable implementations.

(b) High variance in a model means that the model is trained in such a way to effectively capture noise. In other words, the model is over-fitted on the training dataset. Also, the model is highly sensitive to any changes in training dataset. In order to reduce the variance of a model one can reduce model complexity (for example lower polynomial degree) or use regularization, more data for training and ensemble methods.

(c) It is because of the nature of Lasso regularization forcing some features to be zero leading to some information loss. On the other hand, Ridge shrinks weights toward zero but keeps all features. Therefore, when there are correlations between features and the output using Lasso is not a good idea, and using Ridge keeps the solution more stable.

(d) Let's look at the L2 (Ridge) regularization equation:

$$L(\mathbf{w}) = -\sum_{i=1}^{n} \left[y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right] + \lambda \|\mathbf{w}\|_2^2$$

L2 regularization increases the bias and decreases the variance since it adds a penalty to the main loss function. Without any regularization weights can be achieved large values. Consequently, adding L2 regularization handle the complexity of the model.

Exercise 2

(a) In a general form, weight can be achieved as follows:

$$\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$$

regression on feature j-th means regression with using just one feature. The above equation is reduced to the following:

$$w_j = \frac{X_j\mathbf{y}}{X_jX_j^T}$$

(b) If features are independent, the covariance matrix $\mathbf{X}^T\mathbf{X}$ will be diagonal. So, its inverse equals to inverse of each diagonal element. This assumption results in calculating weights j-th from the previous part since there is no other element in the covariance matrix affecting weights.

(c) It is like to find the In this case, w_0 and w_j can be obtained as follows:

$$y_j = w_0 + w_jx_{ij} + \varepsilon_i \ , \quad i = 1, \cdots, n \rightarrow \begin{cases} \sum y_i = nw_0 + w_j\sum x_{ij} \\ \sum x_{ij}y_i = w_0\sum x_{ij} + w_j\sum x_{ij}^2 \end{cases}$$

Solving for w_0 and w_j,

$$w_0 = \bar{y} - w_j\bar{x}_j$$

$$w_j = \frac{\text{Cov}\left(x_j, y\right)}{\text{Var}\left(x_j\right)}$$

Exercise 3

Exercise 4

(a)

- If $i = k$,

$$\frac{\partial\hat{y}_i}{\partial z_i} = \frac{\exp(z_i)\left[\sum_{j=1}^n \exp\left(z_j\right)\right] - [\exp(z_i)\exp(z_i)]}{\left[\sum_{j=1}^n \exp\left(z_j\right)\right]^2}$$

$$\frac{\partial\hat{y}_i}{\partial z_i} = \frac{\exp(z_i)}{\sum_{j=1}^n \exp\left(z_j\right)}\left(\frac{\sum_{j=1}^n \exp\left(z_j\right) - \exp(z_i)}{\sum_{j=1}^n \exp\left(z_j\right)}\right)$$

$$\frac{\partial\hat{y}_i}{\partial z_i} = \hat{y}_j\left(1 - \hat{y}_j\right)$$

- If $i \neq k$,

$$\frac{\partial\hat{y}_i}{\partial z_i} = \frac{-[\exp(z_i)\exp(z_{ik})]}{\left[\sum_{j=1}^n \exp\left(z_j\right)\right]^2}$$

$$\frac{\partial\hat{y}_i}{\partial z_i} = -\hat{y}_j\hat{y}_k$$

(b) One can write,

$$\frac{\partial L}{\partial z_k} = \sum_{i=1}^{n} \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial z_k}$$

$$\frac{\partial L}{\partial \hat{y}_i} = -\frac{y_i}{\hat{y}_i}$$

$$\frac{\partial \hat{y}_i}{\partial z_k} \rightarrow \begin{cases} \hat{y}\,(1-\hat{y}) i = k \\ -\hat{y}_i \hat{y}_k i \neq k \end{cases}$$

$$\frac{\partial L}{\partial z_k} = -y_k\,(1-\hat{y}_k) + \hat{y}_k \underbrace{\sum_{i \neq k} y_i}_{1-\hat{y}_k}$$

$$\frac{\partial L}{\partial z_k} = \hat{y}_k - y_k$$