Technical University of Dortmund Summer Semester 2023
Data Science Exercise Sheet 01
Group member name(s): Ahmed Arian Sajid, Sabrina Sultana, Sharmin Ahmad
Group member UID(s): 235061, 235062, 230239

Advanced Statistical Learning

Answer to Exercise 1.a

The explicit form of the constant model $f(x)$, which does not take into account other features in the data, is given by:

$$f(x) = \beta_0$$

For this model, the empirical risk is given by:

$$R_{emp}(f(x)) = \sum_{i=1}^{n}(L(y^{(i)}, f(x)^{(i)}) = \sum_{i=1}^{n}(y^{(i)}, \beta_0)$$

$$= \sum_{i=1}^{n}(y^{(i)} - \beta_0)^2$$

$$= \sum_{i=1}^{n}((y^{(i)})^2 - 2y^{(i)}\beta_0 + (\beta_0)^2)$$

To optimize, we need to differentiate the empirical risk with respect to β_0

$$\frac{\partial R_{emp}(f(x))}{\partial \beta_0} = \sum_{i=1}^{n}(-2y^{(i)} + 2(\beta_0)) = 0$$

$$\implies 2\sum_{i=1}^{n}(\beta_0) = 2\sum_{i=1}^{n}(y^{(i)})$$

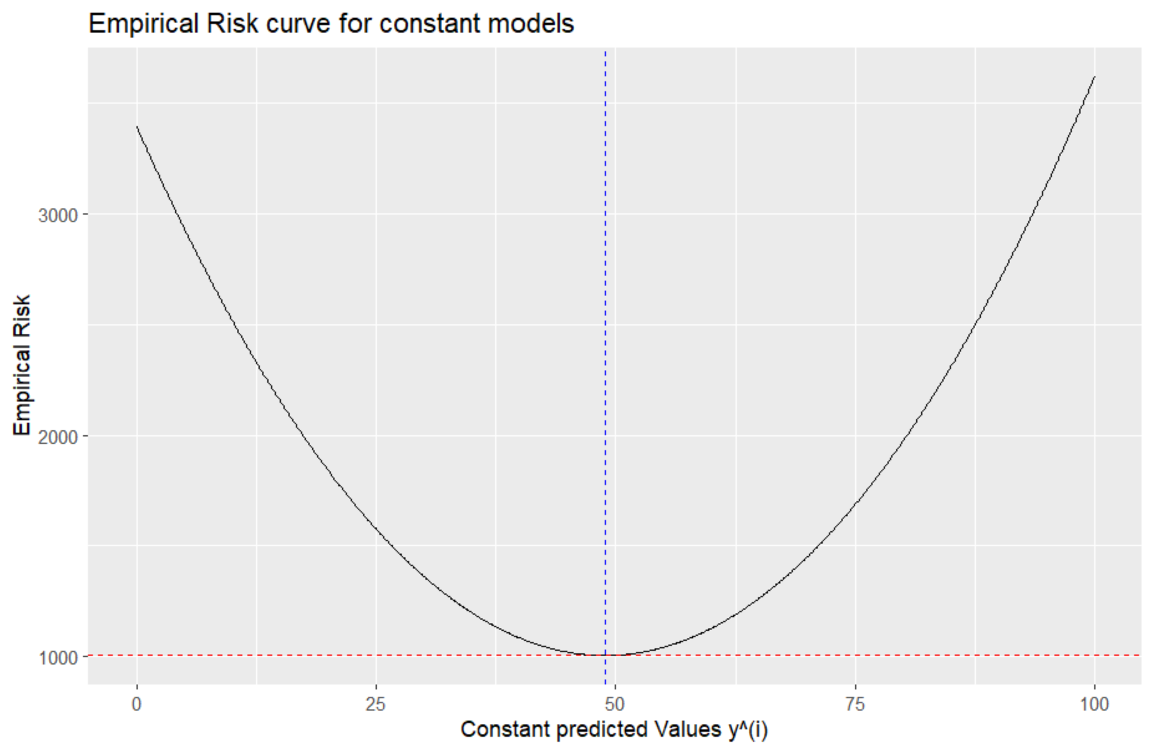
$$\implies n(\beta_0) = \sum_{i=1}^{n}(y^{(i)})$$

$$\hat{\beta_0} = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)})$$

So from the above equation, we can show that the arithmetic mean of the responses is the optimal constant that optimizes the empirical risk induced by the quadratic loss function.

Answer to Exercise 1.b

A data set with response y = 10, 28, 94, 83, 47, 86, 24, 19. To visualize the empirical risk using the constant model, we take different values of β_0 from 0 to 100. According to the answer of 1.a, we should get the optimal value $of R_{emp}$ if we choose β_0 as the average of the response y, which is: $\frac{10+28+94+83+47+86+24+19}{8} = 48.875$



The visualization shows that the optimal constant value is 48.875, which is the average of response y data set value.

Answer to Exercise 1.c

To modify the model from b) by adding a coefficient describing the effect of a single feature x_1, we can rewrite the function as: $f(x) = x^T\theta$ where $x = \{1, x_1\}$and $\theta = \{\theta_0, \theta_1\}$ Decomposing the problem:

1. Representation:

The hypothesis space is given by: $\mathcal{H} = \{f(x) = x^T\theta | \theta \in \mathbb{R}^{p+1}\}$

2. Cost Function:

The cost function is induced by L2 Loss(Quadratic Loss): $R_{emp}(f(x)) = \sum_{i=1}^{n}(L(y^{(i)}, f(x)^{(i)})$ where $L(y^{(i)}, f(x)^{(i)}) = (y^{(i)} - f(x)^{(i)})^2 = ||y - x\theta||_2^2$

3. Optimization:

As the above equation has a closed-form solution, it can be minimized analytically by derivation w.r.t.θ.

Answer to Exercise 1.d

To derive the estimator of the linear regression model specified in c): The quadratic loss of the training data:

$$SSE = \sum_{i=1}^{n}(y^{(i)} - \theta^T x^{(i)})^2$$

which can be written in terms of Matrix Multiplication:

$$SSE = ||y - x\theta||_2^2$$

$$= (y - x\theta)^T(y - x\theta)$$

$$= y^T y - \theta^T x^T y - y^T x\theta + \theta^T x^T x\theta$$

$$= y^T y - 2y^T x\theta + \theta^T x^T x\theta$$

We minimize the SSE by setting the first derivatives to 0:

$$\frac{\partial SSE}{\partial \theta} = -2X^T y + 2X^T X\theta = 0$$

$$\implies 2X^T X\theta = 2X^T y$$

$$\implies X^T X\theta = X^T y$$

$$\hat{\theta} = (X^T X)^{-1} X^T y$$

Answer to Exercise 1.e

No, evaluating the model's error rate on the training dataset (i.e., the training error rate) is insufficient because it performs well on the training data but poorly on unseen data. To get a more estimated model's performance/accuracy/generalization error, we must evaluate the model on a separate test dataset that was not used during training.
If we would want to evaluate the error in our model, we have split the dataset into two separate partitions - training and testing. We should train the model using the training partition, and then the generalization error of that model is estimated using the testing partition. This approach is known as **Holdout**.

Answer to Exercise 2

Given that,

$$L(y, \pi(x)) = -y\ln(\pi(x)) - (1-y)\ln(1-\pi(x))$$

Emperical risk,

$$\begin{aligned} R(\pi) &= \sum_{i=1}^{n} L(y_i, \pi(x_i)) \\ &= \sum_{i=1}^{n} \left(-y_i \ln(\pi(x_i)) - (1-y_i)\ln(1-\pi(x_i))\right) \end{aligned}$$

Let, $\pi(x) = c$

$$R = \sum_{i=1}^{n} \left(-y_i \ln c - (1-y_i)\ln(1-c)\right)$$

$$\begin{aligned} \frac{\partial R}{\partial c} &= \frac{\partial}{\partial c} \sum_{i=1}^{n} \left(-y_i \ln c - (1-y_i)\ln(1-c)\right) \\ &= -\sum_{i=1}^{n} \frac{y_i}{c} + \sum_{i=1}^{n} \frac{1-y_i}{1-c} \end{aligned}$$

To find an optimal constant model, we need to put $\frac{\partial R}{\partial c} = 0$

$$
\begin{aligned}
&-\sum_{i=1}^{n} \frac{y_i}{c} + \sum_{i=1}^{n} \frac{1-y_i}{1-c} = 0 \\
&\implies \sum_{i=1}^{n} \frac{y_i}{c} = \sum_{i=1}^{n} \frac{1-y_i}{1-c} \\
&\implies \frac{1}{c}\sum_{i=1}^{n} y_i = \frac{1}{1-c}\sum_{i=1}^{n} 1 - y_i \\
&\implies \sum_{i=1}^{n} y_i - c\sum_{i=1}^{n} y_i = nc - c\sum_{i=1}^{n} y_i \\
&\implies \sum_{i=1}^{n} y_i = nc \\
&\implies \hat{c} = \frac{1}{n}\sum_{i=1}^{n} y_i
\end{aligned}
$$

So, the relative frequency of class 1 is the optimal constant for the empirical risk $R(\pi)$