1 Hypothesis Class

The answer for this problem is $false$. The reason is because even if I know that a class contains the optimal hypothesis for a given problem, I cannot be certain that I will find the optimal hypothesis. In addition, if we know the optimal hypothesis, we can do too well that the hypothesis might have poor prediction and generalization power. (i) One method that we can use here is $cross - validation$ to reduce cost of the process. And also if we are not sure to find optimal hypothesis and if model is doing over-fitting, then we can use regularization and reduct over-fitting problem by using $ramda$(put more bias and make it under-fitting). Because there are too many data floating, we can divide our training data. Hence, by using regularization, we can get benefit from selecting correct hypothesis.

2 Loss Function

2.1

Not valid. The reason is that if $\hat{y}$, which is our prediction for y, is really large, then the loss function will turn to negative. But since we want to minimize the loss function, then we will increase the $\hat{y}$, which will direct to negative infinity. In addition, loss function cannot go to negative. Therefore, this loss function is not valid.

2.2

$Valid$. This function is valid since the loss function is not going to zero and does not penalize correct prediction. Even if the prediction,$\hat{y}$, is large, the loss function will not go to zero. Therefore, this function $l(y, \hat{y}) = \frac{1}{3}(y - \hat{y})^2$ is valid. For classification, the loss function is not suitable. For binary classification, squared loss function will have equal loss for better prediction and poor prediction, which means it is penalizing the correct answer. As the prediction is more correct, the loss function will penalize more and more. Thus, the function is not suitable.

2.3

Not valid. The function is not valid since if we predict negative number, then the loss function $l(y, \hat{y})$ goes to negative. Suppose we have negative number for $\hat{y}$ and have positive for y, then we will have negative loss with the given function. Hence, the loss function is not valid.

2.4

$Valid$. This function is valid(0-1 loss function). For binary classification, this function is also suitable. The reason is that if we incorrectly predict the value,

then the sign will be different, which will make $1 - y * \hat{y}$ positive and large. Therefore, loss function will go up. If we correctly predict, then $1 - y * \hat{y}$ will be negative, which means that loss function goes to zero. This will add 1 if we predict incorrectly and do nothing if we predict correctly. Therefore, it is suitable and valid function.

3 Ranking

3.1

(i) If we are using classification algorithm and using valid loss function, then we can use suitable loss function to rank a set of instances. We can bring the idea of pairwise classification. Since the model has already been trained for this purpose, we can assume that the instances are divided by the generalized classifiers(not sure how many in the instance). With the valid function, we can calculate the loss function from classification algorithm . If we have all the loss function magnitude of instance, then now we need to sort them based on their magnitude to rank instances. Hence, if we use sorting algorithm from the data structure, we can sort the loss function based on it's vector. Or we can also make function to take input of list of classification loss function at each instance and label each instance. If we return the function with ranked instances. Let L has $l(i), l(j), ..., l(n)$ with n instances in the list. If $l(i) > l(j)$, then $instance(i)$ should ranked before $instance(j)$.
(ii) For the linear regression model, if the linear regression is trained, then it will give linear regression line to calculate the loss function. Based on the linear regression line, we can also calculate the loss function. Therefore, we can rank them based on the loss magnitude with sorting algorithm implemented by some optimal function.

3.2

Since we are using loss function $l(y, \hat{y})$ from the classification algorithm, we can express a loss function as

$$l(y, \hat{y}) = max(0, 1 - y * \hat{y})$$

Magnitude will come from based on the dot product of w and x_i
For linear regression, we will have

$$l(y, \hat{y}) = (y - \hat{y})^2$$

4 Regularization

$$\hat{\beta}_0 = \underset{\beta_0}{\operatorname{argmin}} \|y - X_1\beta_0\|_2^2, \tag{1}$$

$$(\hat{\beta}_1, \hat{\beta}_2) = \underset{\beta_1,\beta_2}{\operatorname{argmin}} \|y - X_1\beta_1 - X_2\beta_2\|_2^2, \tag{2}$$

$$\hat{\beta}_3 = \underset{\beta_3}{\operatorname{argmin}} \|y - X_1\beta_3\|_2^2 + \lambda\|\beta_3\|_2^2, \tag{3}$$

where $\lambda > 0$, $y \in \mathbb{R}^n$, $X_1 \in \mathbb{R}^{n\times d_1}$, and $X_2 \in \mathbb{R}^{n\times d_1}$. (**??**) is well known as the ridge regression. The square norm acts as a penalty function to reduce overfitting. Prove

$$\|y - X_1\hat{\beta}_3\|_2^2 \geq \|y - X_1\hat{\beta}_0\|_2^2 \geq \|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2. \tag{4}$$

1. The proof of the left side is straight forward. Assume that $\|y - X_1\beta_0\|_2^2$, will be minimized with the $\hat{\beta}_0$. Also $\|y - X_1\beta_3\|_2^2 + \lambda\|\beta_3\|_2^2$ will be minimized with $\hat{\beta}_3$. Then we will have two regression function looks like $\|y - X_1\hat{\beta}_0\|_2^2$ and $\|y - X_1\hat{\beta}_3\|_2^2 + \lambda\|\hat{\beta}_3\|_2^2$. But since the term $\|y - X_1\hat{\beta}_0\|_2^2$ is minimized by $\hat{\beta}_0$. This means that $\|y - X_1\hat{\beta}_3\|_2^2$ has to be at least $\|y - X_1\hat{\beta}_0\|_2^2$ for positive *ramda*
   Therefore, $\|y - X_1\hat{\beta}_3\|_2^2 \geq \|y - X_1\hat{\beta}_0\|_2^2$ holds.

2. Now let's look at the right side of the inequalities. Now consider the equation $\|y - X_1\beta_4 - X_2\beta_5\|_2^2$ and put $\beta_4 = \hat{\beta}_0$ and $\beta_5 = 0$. According to the equation to (2), it seems like $\hat{\beta}_1$ can be at most close to $\hat{\beta}_0$ to minimize the function. If we know that $\hat{\beta}_0 \approx \hat{\beta}_1$ and $\|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2$ is minimized with $\hat{\beta}_1$ and $\hat{\beta}_2$, we can say that $\|y - X_1\hat{\beta}_0\|_2^2 \geq \|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2$. If $\hat{\beta}_2$ equals to zero, then $\|y - X_1\hat{\beta}_0\|_2^2 \geq \|y - X_1\hat{\beta}_1\|_2^2$ as we have mentioned. And if $\hat{\beta}_2$ is not equal to zero, then $\|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2$ will be strictly less as it is defined in equaltion (2).

Hence, $\|y - X_1\hat{\beta}_3\|_2^2 \geq \|y - X_1\hat{\beta}_0\|_2^2 \geq \|y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2\|_2^2$ proved. Most left side and most right side will be satisfied because of transitivity rule.