CS 475 Machine Learning: Homework 1
Supervised Classifiers 1
Solution

1) Hypothesis Class (3 points) True/False (and why): You are using an algorithm that selects a hypothesis from a class that you know contains the optimal hypothesis for a given problem. In this case, there is no benefit to using regularization.

(1 point) False.

(2 points) Example: Suppose the optimal hypothesis is a sparse linear model, which belongs to the hypothesis class of all linear models. Then we can still improve the estimation by imposing sparsity-inducing regularization.

There are multiple right answers. The key idea is that regularization can help you find the optimal hypothesis, or that when it is too difficult to find the optimal hypothesis, regularization can lead to better hypotheses.

2) Loss Function (4 points) For each of the following, state if the function is a valid loss function. If it is, state whether it would make a suitable loss function for binary classification. If it is not a valid loss function, why not? Here y is the correct label, and $\hat{y}$ is a decision confidence value, which can be converted to a label by $\text{sign}(\hat{y})$. Moreover, a larger $\hat{y}$ implies a larger confidence on the classification.

1. $\ell(y, \hat{y}) = y - \hat{y}$.

   This is not necessarily a valid loss function, because it is not lower bounded, and can return $-\infty$.

2. $\ell(y, \hat{y}) = \frac{1}{3}(y - \hat{y})^2$

   This can be a good loss function for regression. However, it is not appropriate for binary classification. Suppose $y = \text{sign}(\hat{y})$ and $|\hat{y}| > 1$. Then a larger $|\hat{y}|$ gives a larger loss. In other words, even though predictions are correct you can get increasingly large loss.

3. $\ell(y, \hat{y}) = |(y - \hat{y})|/\hat{y}$

   This is not necessarily a valid loss function; When $|\hat{y}|$ is very large, it always takes a value close to 1, regardless $\hat{y}$ is positive or negative.

4. $\ell(y, \hat{y}) = \max(0, 1 - y \cdot \hat{y})$

   This is a valid loss function. This is the hinge loss, which is used in SVMs for binary classification.

1 point for each problem. No half point.

3) Ranking (7 points) Ranking is a common supervised machine learning problem, such as in ranking search engine results. In a ranking task, the predictor is given a set of instances and returns an ordering over the instances. We haven't discussed how to design or train ranking algorithm, but some of the algorithms we have learned about can be used within a ranking task.

1. How would you use a trained regression or classification algorithm to rank a set of instances? Assume that the model has already been trained for this purpose. You only need to describe how it will be used at test time.

   The classifier takes in a pair of instances, which can be featurized to create a single instance. The classifier then predicts if instance A should be ranked higher than instance B ($\hat{y} = 1$), or B ranked higher than A ($\hat{y} = -1$). Using this comparator, sort all of the instances in the given set.

2. Provide a loss function suitable for ranking. This loss function need not be related to your answer in the first part of the question.

   Swapped pairs loss. Consider all pairs of instances in the ranked list. Increase the loss by 1 for every pair that is out of order, i.e. where A is incorrectly ranked higher than B.

   There are many possible answers to this question, and we'll accept any correct answer.

4) Regularization and Overfitting. (11 points) Statisticians love linear models because these models are very simple and interpretable. Many variants of linear models have been proposed, and most of them are formulated as a (penalized) least squares objective. Consider these three least squares objectives:

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

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

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

where $\lambda > 0$, $y \in \mathbb{R}^n$, $X_1 \in \mathbb{R}^{n\times d_1}$, and $X_2 \in \mathbb{R}^{n\times d_1}$. (3) is well known as the ridge regression. The square norm acts as a penalty function to reduce overfitting. Prove

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

(2 points) Since $(\hat{\beta}_0, 0)$ is a feasible solution to (2), we have

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

(2 points) Since $\hat{\beta}_3$ is a feasible solution to (1), we have

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

There are multiple ways to prove these relations. We will accept any correct proof.