Machine Learning *COMP 6630, Spring 2020*

HW04: Linear models

Remember that only PDF submissions are accepted.

1. What values of λ in the regularized loss objective (Slide #4) will lead to overfitting? What values will lead to underfitting?

$$\min_{\mathbf{w}} \sum_{n} \underbrace{\mathbf{1}[y_n \mathbf{w}^T \mathbf{x}_n < 0]}_{\text{fewest mistakes}} + \overset{\text{hyperparameter}}{\lambda} \underbrace{R(\mathbf{w})}_{\text{simpler model}}$$

Lambda is what we call a hyper parameter. This parameter cannot be learned from the data itself. It is one that we choose ourselves (chosen by humans).

Let's show what will happen with three different values for lambda [-1000, 0, 1000]

-1000: The model will prefer complicated solutions and will therefor lead to overfitting.

0: Nothing will happen, when lambda is zero it neither complicates or simplifies the problem.

1000: The model will prefer simpler solutions and will therefore tend to become underfitting.

2. The solution to the least-squares regression problem involves inverting the matrix ($\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_D$) which might not be invertible. Is this actually a problem?

Yes, this is a problem. If the matrix is not invertible, then our solutions will not be correct. We need an invertible matrix in this scenario.

3. Explain why the squared loss is not suitable for binary classification problems.

The main reason not to use squared loss as the cost function for logistic regression or binary classification is because you don't want the cost function to be non-convex in nature. If the cost function is non convex, then it is hard for the function to optimally converge. In case of binary classification, the cost function curve comes out as non-convex. Also, squared loss tends to be heavily dominated by outliers which is typically not a good thing for what we do.

4. One disadvantage of the squared loss is that it has a tendency to be dominated by outliers – the overall loss $\sum_n (y_n - \hat{y}_n)^2$, is influenced too much by points that have high $|y_n - \hat{y}_n|$. Suggest a modification to the squared loss that remedies this.
   Instead of squaring $(y_n - \hat{y}_n)$, we can take the absolute value. This will make it so that outliers do not dominate anymore.