DS5220: Supervised Machine Learning and Learning Theory Fall 2022

Problem Set 5

Instructor: Hongyang Ryan Zhang *Due:* ***December 9, 2022, 11:59pm***

Instructions:

- You are expected to write up the solution on your own. Discussions and collaborations are encouraged; remember to mention any fellow students you discussed with when you turn in the solution.
- There are up to three late days for all the problem sets and project submissions. Use them wisely. After that, the grade depreciates by 20% for every extra day. Late submissions are considered case by case. Please reach out to the instructor if you cannot meet the deadline.
- Submit your written solutions to Gradescope and upload your code to Canvas. You are recommended to write up the solution in LaTeX.

Problem 1 (20 points). Write down the answer to the following questions.

- [**5 points**] Let D be a probability distribution over d-dimensional vectors in $\mathbb{R}^d$. Write down the definition/expression of the expectation of D and the covariance matrix of D.
- [**5 points**] Let $M \in \mathbb{R}^{m\times n}$ be an m by n matrix. Explain what is the singular value decomposition (SVD) of M; that is, write down the expression of SVD and describe each part of the expression. Then, describe at least two use cases of SVD.
- [**5 points**] Write down the expression of ridge regression and lasso expression. Then, explain how would you choose the regularization parameter in ridge regression and lasso regression. Finally, describe the commonality as well as the difference between ridge regression and lasso regression.
- [**5 points**] Explain what is the Dropout regularization for training neural networks.

Problem 2 (40 points). This problem is a continuation of Problem 3 (from Problem Set 3) to predict the acceptance rate using all variables other than `Accept` and `Apps` in the `College` data set. We will be using regression with Gradient Boosting. Feel free to re-use the code from Problem Set 3 to load the dataset.

(a) **[0 points]** Split the data into a training set and a test set with 80% observations in the training set and 20% observations in the test set.

(b) **[10 points]** Fit a Boosted Regression Tree (BRT) model to the training set. Plot the tree, and interpret the results. What test MSE do you obtain? [Hint: You may find the function `GradientBoostingRegressor()` in the `sklearn.ensemble` library useful.]

(c) **[15 points]** Determine the optimal set of parameters for `n_estimators`, `max_depth` and `min_samples_split` for the boosted tree model using cross-validation. [Hint: The range for the number of tree estimators is $[1, 400]$. The range for the maximum tree depth is $[1, 10]$. The range for the minimum samples per split is $[1, 10]$.]

(d) **[15 points]** Based on the selected parameters, make a horizontal bar plot for the feature importance scores and discuss your results. [Hint: Keep the strongest predictive features at the top and the weakest predictive features at the bottom of the plot for better visualization.]

Problem 3 (40 points). With a softmax activation function, we apply the softmax function to a vector $z \in \mathbb{R}^K$ for some positive integer K such that the probability value of the j-th output neuron, denoted by p_j, is equal to

$$p_j = \frac{\exp(z_j)}{\sum_{i=1}^{K} \exp(z_i)}, \text{ for any } j = 1, 2, \ldots, K.$$

(a) **[10 points]** Let j be any fixed integer from 1 to K. Show that $\frac{\partial p_j}{\partial z_k}$—the partial derivative of p_j over z_k—is positive if $k = j$, and otherwise negative if $k \neq j$.

(b) **[15 points]** Based on (a), argue that increasing the value of z_j must increase the corresponding probability value p_j. Additionally, increasing z_j will decrease all other probability values p_k for any $k \neq j$.

(c) **[15 points]** Let $c > 0$ be any positive value. Consider

$$q_j(c) = \frac{\exp(c \cdot z_j)}{\sum_{i=1}^{K} \exp(c \cdot z_i)}, \text{ for any } j = 1, 2, \ldots, K.$$

Let $M = \max_{i=1}^{K} z_i$ be the maximum over all z_i, for $i = 1, \ldots, K$. Suppose that there are n entries among $z_1, z_2, \ldots, z_K$ whose value is equal to M. Show that the following statement regarding the limit of $q_j(c)$ holds when c increases to positive infinity $+\infty$:

$$\lim_{c \to +\infty} q_j(c) = \begin{cases} 0 & \text{if } z_j < M \\ \frac{1}{n} & \text{if } z_j = M \end{cases}$$