DS 5220: Supervised Machine Learning and Learning Theory (Spring 2023) Dr. Roi Yehoshua
Student name: (Due) February 16, 2023

PS1: Regression

1 Quadratic Regression (20%)

You are given the following set of five points: (-2, 0), (-1, 0), (0, 1), (1, 0) and (2, 0), where the first element of each pair is the predictor variable x, and the second element is the response variable y.

1. Find the parabola (a function of the form $y = ax^2 + bx + c$) that best fits these data points using the normal equations.
2. Draw a plot that shows the data points and the best fitting parabola.

2 Regularized Linear Regression (40%)

Recall the problem of linear regression and the derivation of its various learning algorithms. We are given a data set $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$ (i.e., each data point has d features) and $y_i \in \mathbb{R}$. The hypothesis class we consider in linear regression is, for $\mathbf{w} \in \mathbb{R}^{d+1}$:

$$h(\mathbf{x}) = w_0 + w_1 x_1 + \ldots + w_d x_d = w_0 + \sum_{j=1}^{d} w_j x_j$$

In *ridge regression*, we add a regularization term to the ordinary least squares error function, such that the cost function to minimize is:

$$J(\mathbf{w}) = \sum_{i=1}^{n} (h(\mathbf{x}_i) - y_i)^2 + \lambda \sum_{j=0}^{d} w_j^2$$

1. Prove that the vector $\mathbf{w}^*$ that minimizes $J(\mathbf{w})$ is:

$$\mathbf{w}^* = \left(X^T X + \lambda I\right)^{-1} X^T \mathbf{y}$$

   where X is the $n \times (d+1)$ design matrix, whose ith row is $\mathbf{x}_i$, and $\mathbf{y} = (y_1, ..., y_n)^T$.

2. λ is known as a *regularization constant*; it is a *hyperparameter* that is chosen by the algorithm designer and fixed during learning. What do you expect to happen to the optimal $\mathbf{w}^*$ when $\lambda = 0$? $\lambda \to +\infty$? $\lambda \to -\infty$? Explain your answer.

3. Find the partial derivatives $\frac{\partial J}{\partial w_j}, j \in \{0, \ldots, d\}$. (You can only derive once for an arbitrary index j.)

4. Write out the update rule for gradient descent applied to the error function $J(\mathbf{w})$ above. Compare with the gradient descent update rule for standard linear regression; what is the difference? How does this difference affect gradient descent, assuming $\lambda > 0$? Consider what happens both when w_j is positive and when it is negative.

5. In order for the problem to be well-defined, an appropriate regularization constant λ must be chosen for the given problem (data set). How should the designer choose λ?

3 California House Prices Prediction (40%)

The file housing.csv contains data on median house prices in California districts, derived from the 1990 census data. The data set contains 20,640 rows, one row per district (house block). Each row contains the following features:

- `longitude`: how far west a house is; a higher value is farther west.
- `latitude`: how far north a house is; a higher value is farther north.
- `housing_median_age`: median age of a house within the block; a lower number is a newer building.
- `total_rooms`: total number of rooms within the block.
- `total_bedrooms`: total number of bedrooms within the block.
- `population`: total number of people residing within the block.
- `households`: total number of households (a group of people residing within a home unit) within the block.
- `median_income`: median income for households within the block (measured in tens of thousands of US dollars).
- `median_house_value`: median house value for households within the block (measured in US dollars).
- `ocean_proximity`: location of the house with respect to the ocean. Can have one of the following values: NEAR BAY, NEAR OCEAN, <1H OCEAN, INLAND, ISLAND.

The objective in this data set is to predict the median house value in a given district based on the values of the other features.

1. Explore the data set and display summary statistics of the data. What can you learn from it on the data?

2. Compute the correlation between each feature and the target `median_house_value`. Which features have strong correlation with the target?

3. Which actions do you need to take in order to prepare the data set for the learning algorithm? Identify at least four such actions.

4. Clean the data set using the data preprocessing techniques discussed in class. Show a sample of the data set before and after the cleaning.

5. Extract at least two new features from the data set that have strong correlation with the target feature.

6. Run linear regression on the transformed data set. Use 80% of the data set as your training set and 20% as your test set. Compute RMSE and R^2 score both on the training and the test sets.

7. Does adding regularization improve the results? If yes, which value of λ provides the best RMSE on the test set? If no, what is the reason for it?

8. Now replace linear regression with `DecisionTreeRegressor` (in `sklearn.tree`). What is the RMSE on the training and test this time? Which phenomenon is observed here?