CS6923 Machine Learning, Fall 2023
Prof. Linda Sellie, NYU School of Engineering

Homework 4

Submit on NYU Classes by October 6th at 11:00 p.m.

Part I: Written Exercises

1. Suppose you are given the following dataset, where the target variable is MED:

RM	RAD	DIS	MED
6.6	1	4.0	24.0
6.4	2	5.0	21.6
7.2	2	5.0	34.7
6.4	2	5.0	21.6
7.2	2	5.0	34.7

   Using the data above, write the *equation* derived in the lecture notes to compute the closed form solution for ridge regression where $\lambda = 0.1$. You do not need to actually calculate the coefficient vector - just set up the formula using the numbers given above.

2. Consider a binary classification problem, where $y \in \{0, 1\}$. The independently and identically distributed (i.i.d) examples

$$D = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \dots, (\mathbf{x}^{(N)}, y^{(N)})\}$$

   are divided into two disjoint sets, D_{train} and D_{val}.

   (a) Suppose you fit a model h using the training set, D_{train}, and then estimate its error using the validation set, D_{val}. If the size of D_{val} was 100 (i.e., $|D_{\text{val}}| = 100$), how confident are you that the true error of h is within 0.1 of its average error on D_{val}?

   (b) Repeat the previous question, but now with $|D_{\text{val}}| = 200$, i.e., you have 200 examples in your validation set.

   (c) Suppose you fit two models, h_1 and h_2 (both fitted using D_{train}), and then you selected the model with the smallest error on your validation set, D_{val}. If $|D_{\text{val}}| = 100$, how confident are you that the selected model's true error is within 0.1 of its average error on D_{val}?

   (d) *(Do not turn in this question)* When dividing the set of examples D into D_{train} and D_{val}, how large should D_{val} be if you want to be 90% confident that the true error of h is within 0.05 of the average error your hypothesis makes on D_{val}?

   In solving this problem, use the Hoeffding bound we discussed in class. An additional resource is `https://www.cs.cmu.edu/~avrim/ML14/inequalities.pdf`

3. *(Do not turn in this question)* This question explores the Bayesian connection to ridge regression, building on the probabilistic view discussed in Lecture 3, where we assumed:

$$y^{(i)} = \mathbf{w}^T \mathbf{x}^{(i)} + \epsilon^{(i)},$$

   with noise being independent and identically distributed as $\epsilon^{(i)} \sim \mathcal{N}(0, \sigma^2)$.

Now, we will specify a *prior distribution* $p(\mathbf{w})$ on the parameters $\mathbf{w}$. Incorporating a prior allows us to denote that some values of $\mathbf{w}$ are more probable than others, serving as a regularizer for the parameters.

Recalling Bayes' rule, discussed in the second lecture, we can combine the prior $p(\mathbf{w})$ with the likelihood $\prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w})$ to obtain:

$$\mathbf{w}_{\mathrm{MAP}} = \arg\max_{\mathbf{w}} \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)};\mathbf{w})p(\mathbf{w}).$$

For this question, assume that each w_i for $i = 1, \ldots, d$ is independent, identically distributed according to a Gaussian distribution, denoted as $w_i \sim \mathcal{N}(0, \rho^2)$ for $i = 1, \ldots, d$. We use ρ instead of σ to avoid confusion with the noise term.

What is $\mathbf{w}_{\mathrm{MAP}}$? You do not need to re-derive the results given in the lecture notes.

Argue how the ridge regression estimate aligns with the MAP (maximum a posteriori probability) estimate when $\mathbf{w}$ has the prior distribution described above.

Part II: Programming Exercise

1. **Linear Regression on Boston Housing Dataset**

   In this problem, you will investigate a linear regression problem using the real-world Boston Housing dataset. The objective is to estimate the median price of owner-occupied houses in towns near Boston by utilizing 13 attributes. Perform the following tasks:

   (a) **Linear Regression Model**

   - Apply 10-fold cross-validation to fit a linear regression model using the closed-form solution discussed in class.
   - Compute and print the average in-sample error E_{in} (i.e. sum the squared errors received for each fold and divide by N.)
   - Compute and print the cross-validation error E_{cv}, calculated as:

     $$E_{\mathrm{cv}} = \frac{1}{N}\sum_{i=1}^{k}\sum_{j \in \mathrm{fold}\ i} (y^{(j)} - \hat{y}^{(j)})^2$$

     where k is the number of folds, $y^{(j)}$ is the true target value for the j^{th} example, and $\hat{y}^{(j)}$ is the predicted value. (See the note below regarding the modification from Lecture 2).
   - Print both E_{cv} and E_{in}.
   - Some code is provided in the notebook to assist you.

   (b) **Ridge Regression Model**

   - Apply 10-fold cross-validation to fit a ridge regression model using the closed-form solution discussed in class.
   - Run 10-fold cross-validation multiple times to determine the optimal λ using values from `np.logspace(-5, 1, num=15)`. Note: In Python, use `alpha` as a substitute for λ since `lambda` is a reserved keyword.
   - For each value of λ, calculate and print both the average E_{in} and E_{cv} as described earlier.

(c) **Polynomial Transformation**

Repeat the exercises mentioned above, applying a polynomial transformation of degree 2 to the features of the dataset.

(d) **Experiment**

- Experiment with different values of λ and various data transformations.
- Document your results, specifying the chosen λ and any transformations applied.

(e) **Optional: Feature Selection**

- Examine the effect of various feature selection methods on model performance.
- Identify which features are most influential in making predictions.
- Assess whether omitting any features leads to an improvement in the model.

2. **Questions and Responses**

Provide the answers to the following questions in a PDF file named `proganswers.pdf`:

(a) List the E_{cv} and average E_{in} values obtained for different λ values, including the case where $\lambda = 0$ (calculated in programming question 1a).

(b) Given a choice, which model would you select to predict the average house price in a town and why? Refit the best-performing model using all the available data and specify the parameters (i.e. $\mathbf{w}$). Predict the average price of a house with features: $[1, 0.1, 11, 7, 0, 0.4, 6, 70, 4, 6, 300, 16, 360, 10]$, ensuring the features are scaled appropriately before making predictions.

(c) Discuss your findings and insights gained from the experiments conducted in question 1d.

Note on Cross-Validation Error Calculation

In most lecture notes and literature on k-fold cross-validation, the procedure for calculating the cross-validation error typically involves computing the mean of the errors obtained from each fold. However, in the context of our analysis, given the relatively small size of the dataset and the possibility of unequal numbers of samples in each fold, this traditional approach might not be mathematically rigorous.

To address this, our approach for calculating the cross-validation error will deviate slightly from the traditional method. Instead of merely averaging the errors from each fold, we will sum up the errors across all folds and then divide by N, the total number of training examples. This ensures that our error estimate is unbiased and takes into account the potential discrepancy in the number of samples across different folds.

Mathematically, the cross-validation error, E_{cv}, is computed as:

$$E_{\text{cv}} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j \in \text{fold } i} (y^{(j)} - \hat{y}^{(j)})^2$$

where k is the number of folds, $y^{(j)}$ is the true target value for the j^{th} example, and $\hat{y}^{(j)}$ is the predicted value.

Part III: Ethical Considerations in the Boston Housing Dataset

The Boston Housing dataset, widely used for regression analysis in machine learning, includes several features to predict housing prices in Boston neighborhoods in the 1970s. However, it is scrutinized for incorporating a feature, denoted as 'B', that indirectly encodes racial information. This feature is calculated as $1000(Bk - 0.63)^2$ where Bk represents the proportion of Black residents by town.

Including such a feature introduces significant ethical concerns, as models trained on this dataset might learn and perpetuate racial biases, thereby contributing to inequality and unfair treatment. It is imperative for practitioners in machine learning and data science to recognize and address these concerns, aiming to create models that are fair, unbiased, and representative of diverse populations.

Due to time and curriculum constraints, we will not delve deeply into this critical topic in class; however, you are strongly encouraged to explore this issue further independently.

Please reflect on the following two questions about how the ‘B‘ feature might influence model predictions and consider potential strategies for addressing and mitigating the biases present in this dataset.

Do not submit your answers.

1. How might the inclusion of the B feature influence the model’s predictions, and what are the potential real-world implications of this?

2. Can you think of ways to address or mitigate the biases present in this dataset?