DS-UA 301
Advanced Topics in Data Science
Instructor: Parijat Dube

Homework 1

Due: October 1, 2023

General Instructions

This homework must be turned in on both Gradescope and Brightspace by 11:59 pm on the due date. It must be your own work and your own work only—you must not copy anyone's work, or allow anyone to copy yours. This extends to writing code. You may consult with others, but when you write up, you must do so alone. Your homework submission must be written and submitted using Jupyter Notebook (.ipynb). **No handwritten solutions will be accepted**. You should submit:

1. On **Gradescope**: one Jupyter Notebook(.ipynb) containing your solutions for each Problem in this assignment. Each Jupyter Notebook should be named as (*netid*)_hw(*#Homework*)_p(*#Problem*).ipynb.

   For example, *bz2058_hw1_p1.ipynb*.

2. On **Brightspace**: one .zip file containing all the notebooks, datasets, and other supporting materials involved in this homework. The .zip file should be named as (*netid*)_hw(*#Homework*).zip.

   For example, *bz2058_hw1.zip*.

Please make sure your answers are clearly structured in the Jupyter Notebooks:

1. Copy the Template Notebooks provided on Brightspace and write your solutions there.
2. Label each question part clearly. Do not include written answers as code comments. The code used to obtain the answer for each question part should accompany the written answer.
3. All plots should include informative axis labels and legends. All codes should be accompanied by informative comments. All output of the code should be retained.
4. Math formulas can be typesetted in Markdown in the same way as LaTeX. A Markdown Guide is provided on Brightspace for reference.

For more homework-related policies, please refer to the syllabus.

Problem 1 - *Bias Variance Tradeoff* 25 points

Read carefully the article at https://dustinstansbury.github.io/theclevermachine/bias-variance-tradeoff. We will review this in Lab 1.

Let $y(x) = f(x) + \epsilon$ be the measured relationship and $\hat{y} = g(x)$ be the model predicted value of y. Then MSE over test instance x_i, $i = 1, \ldots, t$, is given by:

$$MSE = \frac{1}{t}\sum_{i=1}^{t}(f(x_i) + \epsilon - g(x_i))^2$$

Recall that the expected mean squared error of a regression problem can be written as

$$E[MSE] = Bias^2 + Variance + Noise$$

1. Consider the case when $f(x) = x + \sin(1.5x)$ and $y(x) = f(x) + \mathcal{N}(0, 0.3^2)$, where $\mathcal{N}(0, 0.3^2)$ is a normal distribution with mean 0 and standard deviation 0.3. Create a dataset of size 20 points by randomly generating samples from y. Display the dataset and $f(x)$ in one plot. Use scatter plot for y and smooth line plot for $f(x)$. **(5)**

Homework 1

2. Use the weighted sum of polynomials as an estimator function for $f(x)$, in particular, let the form of the estimator function be:

$$g_n(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_n x^n$$

Consider four candidate estimators, g_1, g_3, g_5, and g_{10}. Estimate the coefficients of each of the four estimators using the sampled dataset and plot $y(x), f(x), g_1(x), g_3(x), g_{10}(x)$ in the same plot. Which estimator is underfitting? Which one is overfitting? **(8)**

3. Generate 100 datasets (each of size 50) by randomly sampling from y.

   (a) Next fit the estimators of varying complexity, i.e., $g_1, g_2, ...g_{15}$ using the training set for each dataset. Then calculate and plot the squared bias, variance, and error on the testing set for each of the estimators showing the tradeoff between bias and variance with model complexity. **(8)** (**Note**: You will have one plot, where the x-axis will be model complexity in polynomial degree, i.e. from 1 to 15. You will plot three lines: Squared bias, variance, and error(MSE), all calculated on the testing set only)

   (b) Identify the best model, i.e., the model with the smallest Mean Squared Error. What is the value of bias and variance for this model? **(4)**

Note: For parts 1 and 2 of this problem, limit the range of x range for the 20 points generated to lie between some range, say 0 and 5, to observe overfitting and underfitting. Remember to use the same range for training and testing. Additionally, please note to sort the points (increasing x) before plotting. The graph must contain a scatter plot of the points and line plot of the functions.

For part 3 of this problem, there are two different ways to sample x and y when creating 100 datasets.

- Follow the post https://dustinstansbury.github.io/theclevermachine/bias-variance-tradeoff. The idea is to keep the value of x same across all the 100 datasets. The y values will vary since they contain the noise (Normal distribution) component.
- Sample a test set (of size 10) before sampling any training dataset. Then sample the training set (of size 40) for each 100 datasets but make sure that none of the 10 test set samples should show in any of the 100 datasets. So all the datasets share this common test set but their training set is different.

  The key is to have a fixed test set even though you have 100 independently sampled training sets

Problem 2 - *KNN hyperparameter tuning using cross validation* 20 points

For this problem, you should read the article at: https://www.analyticsvidhya.com/blog/2021/01/a-quick-introduction-to-k-nearest-neighbor-knn-classification-using-python/ to review how to work with K-Nearest Neighbor (KNN) in sklearn. We will use the same Social Network ads dataset that is used in this post. You will work with an 80-20 train-test split.

You will use the KNN algorithm to predict whether an individual will buy a product or not. As discussed in the class, there are two hyperparameters: the number of neighbors (K) and the distance metric. For distance between two n-dimensional points $\bar{x}_1 = \{x_{1,1}, x_{1,2}, \ldots, x_{1,n}\}$ and $\bar{x}_2 = \{x_{2,1}, x_{2,2}, \ldots, x_{2,n}\}$ we consider Minkowski distance given by:

$$\left(\sum_{i=1}^{n} |x_{1,i} - x_{2,i}|^p \right)^{1/p}.$$

where p is a parameter. For $p = 2$, this distance is the same as Euclidean distance and for $p = 1$ it is called Manhattan distance.

Homework 1

1. With $K = 4$ and $p = 2$ train a KNN classifier and evaluate its misclassification error, Accuracy, Precision, Recall, and F-1 score on the test set. **(5)**

2. You will use 5-fold cross-validation to identify the best value of K. First, fix $p = 1$ and for $K \in [1, 2, \ldots, 15]$ calculate the misclassification error and plot it as a function of K for different values of K. **(5)**

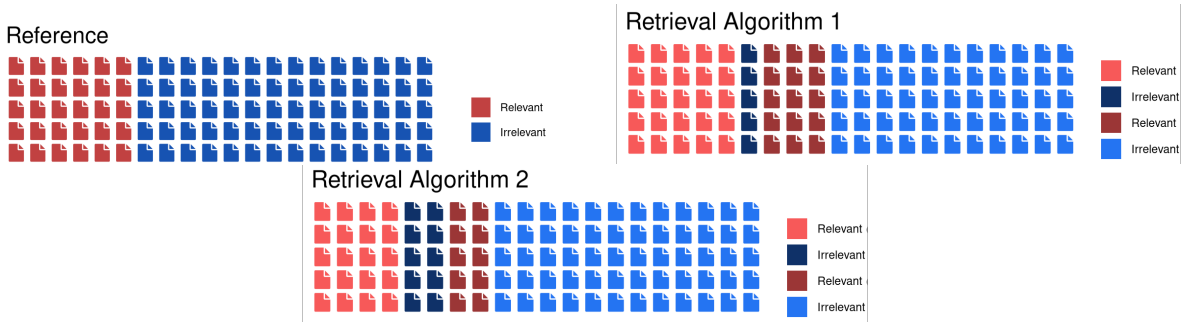
3. Next, fix $p = 2$ and again using 5-fold cross-validation for $K \in [1, 2, \ldots, 15]$ calculate the misclassification error and plot it as a function of K. This should be plotted in the same graph as for $p = 1$ in part 2 of this problem. (**Note**: Keep the line plotted in part 2 for $p = 1$, and plot a new line for $p = 2$) **(5)**

4. What is the best value of K with Euclidean distance? Is this value the same as the Manhattan distance? What combination of p and K gives the best classifier (one with the minimum misclassification error). **(5)**

For cross-validation, you will use `cross_val_score` from sklearn. For reference, you can look at this post where parameter tuning is done in KNN using `cross_val_score`.

Problem 3 - *Which Algorithm is Better?* 10 points

You want to compare two algorithms for document retrieval. You need to answer this problem manually, showing explicit calculations. The ground truth and performance of the two algorithms are shown below for 100 samples (with relevant being positive and irrelevant being negative class):



1. Create a confusion matrix for each of the two algorithms showing TP, FP, FN, TN. Note you need to compare ground truth labels from reference with corresponding labels from different algorithms to count these quantities. Follow the example discussed in class. **(2)**

2. You are interested in finding the algorithm that has better performance on the negative classes. Your friend suggests using Balanced accuracy instead of accuracy to identify the best algorithm. Your instructor suggests using the F-1 score instead. Who is right here and why? Support your answer with numbers. **(4)**

3. Did the advice of your friend or your instructor help you in identifying the right algorithm? If yes, you are good. If not, explain why the metrics suggested by them did not work. **(2)**

4. List all the metric(s) you think will help you make the right selection. **(2)**

Homework 1

Problem 4 - *Logistic Regression with Regularization* 20 points

Regularization with linear regression will be covered in Lab 1. Here we are doing regularization with logistic regression using the IRIS dataset. The dataset was introduced to you in Lab 0.

1. Read documentation of sci-kit learn on `LogisticRegression` class and understand its parameters. In sci-kit learn `LogisticRegression` class takes different parameters: `C, solver, penalty,` and `multi_class`. Explain the significance of each of these parameters and their possible values. **(2)**

2. The parameter penalty of `LogisticRegression class` in sklearn specifies the type of regularization. What is the meaning of `'l1'` and `'l2'` penalty? **(2)**

3. Using `penalty='l1'` and `penalty='l2'` fit 10 logistic regression models one for each of 10 different values of `C` (total 20 models, 10 for 'l1' and 10 for 'l2'), with `C` $= 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000, 10000, 100000$ and `multi_class='ovr'`. Collect the weight coefficients for the two features (petal width and petal length) of class 0 and 2 and plot them for different values of C. What is your observation from these graphs for `'l1'` and `'l2'` penalty? **(8)**

4. Let $\boldsymbol{\beta}_C$ denote the weight coefficients learned for a model for a given `C`. Calculate the ratio $\frac{\|\boldsymbol{\beta}_C\|_2}{\|\boldsymbol{\beta}_{100000}\|_2}$ (the double brackets indicate the L_2 norm) for each value of `C` and penalty `'l1'` and `'l2'`. Plot this ratio on the x-axis and the value of the four coefficients on the y-axis for different values of `C`. You will get similar graphs as we discussed in class for regularization with linear regression. This will show you how the ratio between the total magnitude of coefficients with varying degrees of regularization and with `C`=100000. What is your observation from these graphs for `'l1'` and `'l2'` penalty? **(8)**

Note: For parts 3 and 4 of this problem, you will have two plots, one for `'l1'` penalty and another one for `'l2'` penalty. In plots for part 3, the x-axis would be 10 different values of `C` (or $\log(C)$ preferably). In plots for part 4, the x-axis would be 10 different ratios of L_2 norm. In both parts, the y-axis would be the value of four coefficients: *petal width of class 0*, *petal length of class 0*, *petal width of class 2*, and *petal length of class 2*.