DS-UA 301
Advanced Topics in Data Science
Instructor: Parijat Dube
TAs: Astha Gupta, Rahul Sankar
Due: September 24, 2022

Homework 1

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 2.

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 $y(x) = f(x) + \mathcal{N}(0, 0.3)$, where $\mathcal{N}(0, 0.3)$ is 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 $f(x)$. Use scatter plot for y and smooth line plot for $f(x)$. **(5)**

2. Use weighted sum of polynomials as an estimator function for $f(x)$, in particular, let the form of 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 $y(x), f(x), g_1(x), g_3(x), g_{10}(x)$. 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 testing set for each of the estimators showing the tradeoff between bias and variance with model complexity. **(8)**
   (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 part 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 it contains the noise (Normal distribution) component.
- Sample a test set (of size 10) before sampling any training dataset. Then sample training set (of size 40) for each 100 dataset 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 train set is different.

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

Homework 1

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 a 80-20 train-test split.

You will use 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 same as Euclidean distance and for $p = 1$ it is called Manhattan distance.

1. With $K = 4$ and $p = 2$ train a KNN classifier and evaluate its misclassification error, Accuracy, Precision, Recall, 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. **(5)**
4. What is the best value of K with Eucledian distance? Is this value the same with Manhattan distance? What combination of p and K gives the best classifier (one with the minimum misclassification error). **(5)**

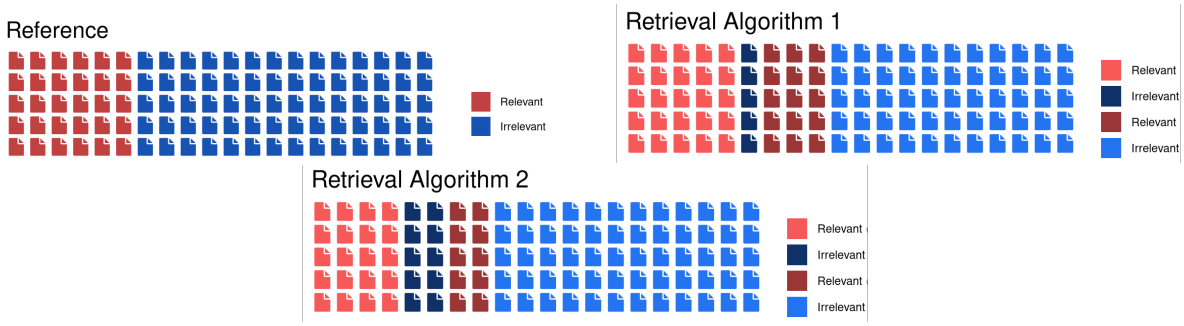
For cross validation you will use `cross_val_score` from sklearn. For reference you can look at this post where parameter tuning is done in KNN using `cross_val_score`.

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 is shown below for 100 samples (with relevant being positive and irrelevant being negative class):

1. Create confusion matrix for 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 which has better performance on the negative classes. Your friend suggests to use Balanced accuracy instead of accuracy to identify the best algorithm. Your instructor suggests to use F-1 score instead. Who is right here and why? Support your answer with numbers. **(4)**
3. Did the advise of your friend or your instructor helped 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) do you think will give 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 2. Here we are doing regularization with logistic regression using the IRIS dataset. The dataset was introduced to you in Lab 1.

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 $\texttt{C} = 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000, 10000, 100000$ 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 L2 norm) for each value of `C` and penalty `'l1'` and `'l2'`. Plot this ratio on x-axis and value of the four coefficients on y-axis for different values of `C`. You will get similar graphs as we discussed in the 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)**

Problem 5 - *Algorithmic Performance Scaling* 25 points

OpenML (https://www.openml.org) has thousands of datasets for classification tasks. Select any sufficiently large (having greater than 50K instances) dataset from OpenML with multiple (greater than 2) output classes.

1. Summarize the attributes of the selected dataset: number of features, number of instances, number of classes, number of numerical features, number of categorical features. Is the dataset balanced? Plot the distribution of number of samples per class. **(5)**

2. For each dataset, select 80% of data as training set and remaining 20% as test set. Generate 10 different subsets of the training set by randomly subsampling $10\%, 20\%, \ldots, 100\%$ of the training set. Use each of these subsets to train two different classifiers: *Decision Tree* and *Gradient boosting* in sklearn. You will work with default hyperparameters for these classifiers in sklearn. When training a classifier also measure the wall clock time to train. After each training, evaluate the accuracy of trained models on

Homework 1

the test set. Report model accuracy and training time for each of the 10 subsets of the training set for the two models in a table. **(8)**

3. Using the data collected in part 2 you will create *learning curve* for the two classifiers. A learning curve shows how the accuracy changes with increasing size of training data. You will create one chart with horizontal axis being the percentage of training set and vertical axis being the accuracy on test set. On this chart you will plot learning curve for Decision Tree and Gradient Boosting. **(5)**

4. Next using the data collected in part 3 you will create a chart showing the training time of classifiers with increasing size of training data. So, for each classifier you will have one plot showing the training time as a function of training data size. **(3)**

5. Study the scaling of training time and accuracy of classifiers with training data size using the two figures generated in part 3 and 4 of this problem. Compare the performance of classifiers in terms of training time and accuracy and write 3 main observations. Which gives better accuracy ? Which has shorter training time ? **(4)**