In The Name of Allah

Machine Learning (Fall 2022)

Instructor: Mahdi Yazdian

TA: Darezereshki & Ramazankhani

Homework#4: Instance-based Learning & Ensemble Learning

Due Date: 1401.10.19

A. K-NN algorithm for classification

Consider a data set $\{x(n), y(n)\}_{n=1}^{300}$ consisting of 300 points $\mathbf{x}(n)=(x_1(n),x_2(n))$ and their corresponding labels $y(n)$, such that the first 100 points have label $y(n) = 0$ and are generated according to a Gaussian distribution, $x(n) \sim \mathcal{N}([-1,0], \delta^2 I)$, other 100 points have label $y(n) = 1$ and are generated according to a Gaussian distribution $x(n) \sim \mathcal{N}([1,0], \delta^2 I)$, and the remaining 100 points have label $y(n) = 2$ and are generated according to a Gaussian distribution $x(n) \sim \mathcal{N}([0,1], \delta^2 I)$. Split the data set into a training and test set, containing respectively 75% and 25% of the dataset.

1. Implement k-Nearest Neighbor (k-NN) algorithm.
2. Generate one data set with $\sigma^2 = 0.10$. Classify the test data set using the k-NN algorithm with respect to the Euclidean distance for $K \in \{1,3,5,\dots,21\}$. In order to visualize all dataset and the corresponding label, plot all generated points in a coordinate plane and the estimated decision boundary for the best k, plot the training and test error for each k and discuss about the results.
   Hint: Use matplotlib.pyplot.countourf to plot the decision boundary.
3. Fix the variance $\sigma^2 = 0.10$ and $k = 1$. Run the algorithm over 50 randomly generated datasets (training and test dataset), compute the average error rate of the test dataset and its standard deviation. Repeat it with $\delta^2 \in \{0.15, 0.20, 0.25\}$. Plot the average error rate of the test dataset versus variance σ^2, use error bars to represent the standard deviation. Discuss about the result.

B. K-NN algorithm for regression

Dataset:

In this part we will apply the k-NN algorithm for prediction.

Let's use the Szeged-weather dataset that can be downloaded in https://www.kaggle.com/budincsevity/szeged-weather/data. We want to predict the apparent temperature.

Short description: The Szeged-weather data-set is a daily/hourly summary for Szeged, Hungary area, between 2006 and 2016, in terms of temperature, humidity, apparent temperature, pressure, wind speed, among other measurements.

Data Preparation: For simplicity we will consider only three attributes: apparent temperature, humidity, and temperature, and only the first 2000 samples of the dataset. Permute the order of the 2000 samples uniformly at random, and split the dataset into 5 partitions (folds). Each fold is then used once as a test while the 4 remaining folds form the training set.

1. Implement k-Nearest Neighbor (*k*-NN) regression.
2. Visualize all dataset in terms of temperature (x-axis), humidity (y-axis), and apparent temperature (color).
3. Consider the first 2000 samples of the data set. The prediction, given by the *k*-NN algorithm, is computed by taking into account the average of the values of *k* nearest neighbors. Predict the apparent temperature given humidity and temperature using *k*-NN algorithm with respect to the Euclidean distance for $k = 1$. Repeat five times, using each fold as test at a time. Compute the mean squared error of the test dataset and its standard deviation.
4. Repeat the previous step for $K \in \{3, 5, 7, 10, 15\}$. Plot the mean squared error of the test data versus the parameter *k*, use error bars to represent the standard deviation. Discuss about the results.
5. Implement *k*-NN Regression algorithm with Gaussian kernel.
6. Repeat the above steps using $\sigma \in \{0.01, 0.04, 0.1, 0.4, 1, 4, 10, 40\}$ for Gaussian *k*-NN Regression algorithm and report the results.
7. Which one performs better, *k*-NN regression or kernel *k*-NN regression? Why?

C. Ensemble Learning

Dataset:

This exercise will test your knowledge of ensemble learning.

Download the Heart dataset in https://www.kaggle.com/datasets/shubamsumbria/statlog-heart-data-set. This dataset provides information on the risk factors for heart disease.
Split the data into a test (20 %) and training (80 %) set.

1. Learn a random forest on the training set.

    a. Use 10-fold cross-validation on training data to set the number of trees in the forest.
    b. Plot the train and validation accuracies and their corresponding variances of ten-fold cross-validation for different values of n_estimators and interpret the results when varying the number of trees.
    c. Select the best value for the number of trees in the forest and train the model using the whole training data again, then report confusion-matrix, recall, precision and f-measure for test set.

2. Learn a decision tree on the training set and report confusion-matrix, recall, precision and f-measure for test set.

3. Which of the previous models has the highest accuracy on the test set? Compare decision tree with random forest and itemize your conclusions regarding the results of this exercise.

Report

Prepare a report in PDF format including the figures, answer to the questions and discussions mentioned in the homework.

- Make a folder including your report and you codes (Note that your code is needed to be self-comment)

- Submit all things in a zipped **folder** named as "YourName_YourFamily.rar"

Good Luck.