task_9.3

Machine Learning (WiSe 2025/2026)

Author: Suvansh Shukla
Matriculation No. 256245

Assignment 9 Task 3

Part A

Regression and classification both involve prediction of target values. The only difference is that regression involves predicting numerical values while classification involves predicting categorical values. Another important distinction is that regression deals with continuous values while classification deals with discrete values. This also means that the attributes involved in prediction would be numerical for regression and categorical for classification. Both are also somewhat interchangeable, meaning a classification model may be adjusted to perform regression.

The following are a few algorithms discussed in the lectured sorted according to classification or regression:

- Decision trees: primarily classification, as it uses class labels or ranges of numerical values
- Naive Bayes: primarily classification, as it uses attributes with categorical values
- K-Nearest Neighbour: classification as it uses categorical attributes
- linear regression: like the name implies it tries to predict numerical values

Part B

The performance of regression can easily be evaluated by using error functions like Sum of Squared Errors (SSE) or Mean Squared Error (MSE) where the lower the output of the error functions, the better the performance of the models.

Performance of classifiers can be evaluated by using statistical summaries like F1-score or Kappa statistic, or simply by comparing the classifier with a benchmark, like a random classifier or a different pre-existing classifier.

Part C

Overfitting in instance based learning (especially in the case of K-means) looks as if each instance has its own cluster. Compared to Neural networks, it would not showcase high complexity but rather it would exhibit no generalization at all, even on the test set. Compared to decision trees, we cannot prune or fix the clusters after the model has been developed as the outcome of the algorithm itself results in all instances being their own predictions or clusters.

Part D

The optimal number of 'k' can be determined using the elbow method. It involves the following steps:

1. perform KNN multiple times on the dataset with a different value of K
2. Calculate its accuracy each time
3. Graph the values of the error against each value of K
4. Looking at this graph we choose the value of K which resembles an elbow the most, i.e. the curve towards the bottom that is not quite the lowest possible value for error.

This relates to the previous task because choosing the correct value of K is detrimental to the predictions that will be made. Even K values may lead to ties, K values that are too small may lead to misclassification, same with too high values of K.