Home Assignment 1: Evaluation Tools

TRAN Duy Nghia

1. KMeans and KNearest Neighbors

a. Which one does classification, which one is a clustering method?

- Classification: KNearest Neighbors
- Clustering: KMeans ### b. One of the methods is a supervised method, one is unsupervised. Which one is which, and what does that mean?
- The KNN classifier is a supervised method, which means the training dataset are labeled and the test data are associated to the existing labels. k is the number of nearest neighbors.
- KMeans is an unsupervised method, which means the data is not labeled and the method creates clusters based on the input data. k is the total number of clusters.

2. Evaluation tools

a. Confusion Matrix

In a confution matrix produced with

```
metrics.confusion_matrix( <test_labels>, <predicted>)
```

The column is the prediction. The row is the ground truth label.

b. Classification report

In a classification report generated by

```
metrics.classification_report( <test_labels>, <predicted>)
```

The rows corresponds to classes (or labels). The columns are:

- Precision: True positive / (True positive + False positive) ~ The number of people that are really sick out of the predicted sick people
- Recall: True positive/ (True positive + False negative) ~ The number of predicted sick people out of the real sick people
- F1-score: is the harmonic mean of precision and recall
- Support: counts the number of elements classified into that class

There are also global metrics:

- The accuracy is the total of correctly classified elements over the whole dataset.
- Macro-avg: calculate the average of all the metrics.
- Micro-avg: calculate the average using the number of TP, TN, FP and FN of each class.
- Weighted avg: calculate the score with weight based on the support of the class

c. Evaluating Clustering Algorithm

Several measures are used to evaluate a clustering algorithm.

```
metrics.completeness_score( <train_labels>, <clustered>)
metrics.homogeneity_score( <train_labels>, <clustered>)
metrics.adjusted_mutual_info_score( <train_labels>, <clustered>)
```

- Completeness score: measures how well a member of a given class is classified in the same cluster
- Homogeneity score: measures how well a cluster only contains a single class
- Adjusted mutual info score: This is a measure of the similarity of two labels of the same data. By applying it to clusters, we can evaluate how they are similar and how much we can understand one cluster when looking at the other

d. Using clustering result for classification

By using the clustering on the dataset, we can seperate the data into clusters. We can then:

- Use the cluster as a feature for the dataset
- Divide the dataset into smaller groups and train a classifier on each group