Data Mining

Classification I

1 K-nearest neighbor classification

The objective is to study the use of the method k-nn and

1. Select two datasets with labels.
2. Run k-nn on these two datasets. Calculate the classification error (by comparing the class labels obtained with the prediction and the original labels of test data).
3. Vary the value of k, comment on the results
4. Try to normalize the input dataset, is the performance better?
5. Apply PCA and SVD on the dataset, then what is the performance of k-nn on the new projected data? Justify the answer.
6. Propose an approach to improve the performance of k-nn with the use of k-cross validation.
7. Apply leave-one-out and calculate the error of classification.

2 SVM classifier

Suppose that we use the SVM classifier from sklearn in Python.

- Select two datasets with labels.
- Analyze the dataset. How is the data distribution? linearly separable or non-linearly seperable?

- Set up the SVM parameters suitable to the selected datasets.
- What is the performance of SVMs?
- How can SVM handle multi-class datasets?