Project 1
k-NN

Data

Data: wine.data (can be opened with Text Editors, e.g. Notepad, Word, …)

These data came from the chemical analysis results of different categories of wine

For each sample, 13 attributes are provided

Please read the readme.txt carefully!

Source: http://archive.ics.uci.edu/ml/datasets/Wine

Problem: classify the wine into one of three categories

Requirement

Basic (7/10)

- 1) Develop a k-NN classifier with Euclidean distance and simple voting
- 2) Perform 5-fold cross validation, find out which k performs the best (in terms of accuracy)
- 3) Use PCA to reduce the dimensionality to 6, then perform 2) again. Does PCA improve the accuracy?

Plus (at most 3/10)

- Explore the data before classification using summary statistics or visualization (+0.5)
- Pre-process the data (such as denoising, normalization, feature selection, …) (+0.5~1)
- Try other distance metrics or distance-based voting (+0.5)
- Try other dimensionality reduction methods (+0.5)
- How to set the k value, if not using cross validation? Verify your idea (+1.5)

Report

.pdf is best, .doc and .ppt are acceptable

English is encouraged

List all references

Source code is a plus