Last updated on Monday, May 1, 2017

2017 CAB320 - Assignment Two (Machine Learning)

Assessment information

- Code and report submission due on **Monday 5th June, 08.30am**
- Use **Blackboard** to submit your work
- Group size: up to three people per submission

Overview

- You are provided with a dataset of tumour measurements
- The aim of this assignment is to build some classifiers and evaluate their performance. The classification task is to predict whether a tumour is malignant (M) or benign (B).
- Using the *sklearn* library, you will build the following classifiers
    - a *naive Bayes classifier*
    - a *nearest neighbours* classifier
    - a *decision tree* classifier
    - a *support vector machine* classifier

Dataset

The records are stored in a text file named "medical_records.data". Each row corresponds to a patient record. The diagnosis is the attribute predicted. In this dataset, the diagnosis is the second field and is either *B* (benign) or *M* (malignant). There are 32 attributes in total (ID, diagnosis, 30 real-valued input features).

For the purpose of this assignment, you can ignore the details of the attribute information given below. This information is only provided for the sake of completeness.

Attribute Information

1) ID number

2) Diagnosis (M = malignant, B = benign)

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

All feature values are recorded with four significant digits.

Your tasks

Preprocessing

Complete the function *prepare_dataset* that loads the data records from the text file, and converts the information to numpy arrays.

Build Classifiers

Complete the *build_αβ_classifier* functions. There are four functions of this type to implement in the provided python file (naive Bayes, nearest neighbours, decision tree, and support vector machine).

Whenever a classifier has parameters that affects the capacity/complexity of the classifier, you should use cross-validation to estimate the best value of one of these parameters. It is sufficient to restrict the search to one parameter per classifier.

You have to write code to split the whole dataset into a training, validation and testing sets. You should report the prediction errors on *train_data* as well as on the *validation_data* and *test_data.* These errors are best reported in tables.