SEP 786 Course Project
Comparing Feature Extraction Approaches

Introduction

In this course, we have looked at a couple of feature extraction approaches:

- PCA
- Feature Selection

Project Objective

In this project, your primary task is to **apply both of these approaches** to a larger scale data classification or regression problem. In doing so, you will choose a **single data set** which you will classify with each approach. Further, you must choose two different classification or regression approaches. In other words, you will have 4 experiments – PCA features with classification / regression approaches one and two, and raw features selected via search with classification approaches one and two. The data set must have greater than a **thousand** labeled examples. These examples must be classified into **one of two classes** (for classification) or have a **single scalar output** (for regression). Note that it is possible to find a large data set which has more than two classes from which you could always select examples from only two of the classes. If a dataset contains mixed data (numeric and other types) you can always simply ignore the non-numeric data.

Once the data set has been classified, your job is to **compare the results** from the classifiers or regressors that you chose. The comparison (in general – details will follow below) must include the following analyses:

- A confusion matrix (classifier)
- Mean Square Error (regressor)

Project Methodology

Due to the size of the data set, you must use a **computer based implementation** to solve the problem. For this, you must develop the **Python** code for the purposes of PCA and feature extraction however you can use classifiers or regressors implemented in third party libraries like sklearn..

Choosing a Dataset

You are free to choose any dataset that you like, subject to the constraint of **data types** mentioned below. The only other constraint is that the dataset must include at least **1000 labeled examples for two classes**. One excellent source of data is the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/index.php). These datasets are very popular amongst **researchers** in machine learning. You can also use data from a research project that you are currently working on.

Choosing a Classifier or Regressor

You must select two approaches – either 2 (different) classifiers or two (different) regressors. The focus of this project is on the behavior of the feature extraction, so it isn't critical to understand the details of the workings of the classification or regression approach that you choose.

Training the Classifier/ Regressor

For each classifer / regressor you must use two different approaches to select features – PCA and feature selection via search. Using these approaches you must vary the number of features that you retain for the purposes of classification. Your results discussion must include the variation of the classification results as you modify the number of features.

You must use at least 750 data points to train your classifier (the remaining 250 or more will be used for testing). In performing the training there are four analyses that you must perform:

Computational Time

To get a sense of how long it takes to train a classifier using different approaches with a moderate amount of data, measure the amount of time it takes to perform the training and testing. This measurement can be built into your code using a Python time function or you can simply record the time with a watch (to the nearest second is accurate enough) if the training or testing time is sufficiently long.

If you are investigating classification: Recording the Results of Classification

A minimum of **25 percent** of your dataset must be reserved for testing. The testing results of the algorithms must be shown using a confusion matrix (see below).

Class B **misclassified** as Class A (**False positive** or Type 1 error)	Class A **correctly** classified (**True positive**)
Class B **correctly** classified (**True negative**)	Class A **misclassified** as Class B (**False negative** or Type 2 error)

In this matrix, you can record either the number in each category or a percentage. Note that the sum of elements in the first column must equal the number of elements in class B and the sum of elements in the second column must equal the number of elements in class A.

If you are investigating regression: Recording the Results of Regression

All that is required in the way of reporting the output of a regression model is the Mean Square Error (MSE).

Project Submission

Individual submission is required. You cannot work in groups for this project – simply because the scope and time requirements are relatively small. Your submission must contain the following (and will be assessed based on the marks allocated for each part):

- Computational Times for both training and testing (5 marks)
- Reporting of results (10 marks)
- Submission of Python code for all algorithms (5 marks)
- A two page (approximately) summary (10 marks) of:
    - the data that you used
    - the experiments you conducted
    - discussion of the results (including why you think things ended up the way they did)

Your submission is due **November 12, 2019**. No extensions can be given, since this is close to the last day that I am permitted to submit the marks. Please upload your submission as a .zip file to the dropbox on the course website.