CHEM 6164-Project

The objective will be to apply machine learning (ML) methods to two types of spectra – mass spectra and infrared spectra – taken from four types of fruit and evaluate the models you build for understanding and classifying the spectra.

Data Sets provided

You are provided with two sets of data. Mass spectra, taken in positive ion mode, and FTIR spectra. You are provided with 60 spectra per fruit: orange, tomato, lime and lemon.

The data files are all in the .csv format. The files are described as follows:

Mass spectra are provided with 40,000 m/z data points per spectrum over the range 10-2000 m/z. There are 12 files per fruit, each containing 5 spectra (60 spectra per fruit in total). The first column is each file is m/z. The 2^{nd} – 6^{th} columns are the intensity.
For mass spectra, the fruit were cut in half and the probe was interested into the flesh of the fruit, and the probe was inserted into the mass spectrometer for analysis. Spectra were collected in positive ion mode, with a "high temperature, low-fragmentation" setting for the ion source.

FTIR spectra were recorded by placing a sample of skin/peel into the ATR attachment of the FTIR spectrometer. A different part of the skin/peel was used for each of the 60 spectra, which are each in a different file. Spectra were recorded over 400 – 4000 cm^{-1} and each is averaged over 10 scans. In these files, the first column is wavenumber, the 2^{nd} column is transmittance.

You will also be provided with example code in a Google colab notebook to help you. These will be made available on Blackboard.

Project Outline

1. Provide a *brief* background on why ML could be useful to classify the spectra of samples. Consider, for example, the classification of spectra from healthy and cancerous tissues.
2. Examine the data provided. Plot example spectra (mass spectra and infrared spectra) from each fruit class to get an idea of the differences. Describe some of the differences that you see.

Dimension Reduction

The mass spectra as raw data have a dimension of approximately 40,000 (0.05 m/z resolution over a range from 10-2000 m/z). You will want to reduce the dimensionality to pick out the components of the spectra that show the maximum variance across the data. This can be done using Principal Components Analysis (PCA) and clustering methods (there are other possibilities too but the project will focus on PCA).

PCA

3. Perform PCA on the mass spectra data using all data points.
    a. Examine the variance described by the principal components to decide on a suitable number of principal components that you will use as input to machine learning methods for classification of the data. Describe how you have made this decision. You will have the opportunity to examine the dependence of the machine learning methods on the number of principal components later in the project.
    b. Visualise the results of PCA. Plot scatterplots of the variation of the mass spectra with respect to the first three principal components. You can try a 3-D scatterplot or a series of 2-D scatterplots.
    c. Describe what clustering you see of the data in the plots of the first three principal components, or is the data overlapping? If you see clustering, does clustering of the data correspond to the classes of the fruit data? Remember that you know which fruit each spectrum was taken from.
4. Repeat (a)-(c) above on the infrared spectra from the fruit samples, again using all data.
5. Discuss differences in the effectiveness of PCA on the mass spectra and IR spectra.

CHEM 6164 project

***Unsupervised* learning of fruit type from spectra**

6. Perform k-means clustering on the mass spectra, and separately on the IR spectra, using the chosen principal components from 3a and 4a above.
    a. Do this initially with k=4 because you know that there are four fruit types in the data. Discuss how well the clustering performs in classifying the data. Some ideas are provided below for measuring the success of the clustering.
    b. Test the sensitivity of clustering to the choice of principal components. Choose higher and lower numbers of principal components and re-do clustering so that you can comment on the success of clustering and its dependence on the number of principal components.
    c. Perform clustering again (on both mass spectra and IR spectra) with a range of k and see what optimal value of k you would choose if you did not know the number of fruit types. Describe how you find the best value of k. You can pick what you think is the most suitable set of principal components from the results in (a) and (b).

***Supervised* learning of fruit type**

Choose two pairs of fruits based on your visualisation of the data during the principal component analysis. Choose a pair that you think will be hard to correctly distinguish and a pair that you think will be easier to distinguish. In this section, you will build support vector machines for classification of the data, using the mass spectra and infrared spectra separately. (Note that multi-class support vector machines can be created, but we will focus on two-class classification, as we did in the lectures. Feel free to look into multi-class SVMs, but it is not required here.)

Start by splitting the data into a training set and test set. You will train the SVM on the training set and test it on the test set.

7. Build a *linear* support vector machine classifier using the mass spectra principal components chosen in 3a and 70% of data in the training set. You should split the data into a set of training and testing spectra.
    a. By default, the SVM in scikit-learn will train a soft margin model with $C = 1$, where C is the weight of the hinge loss term (they use different notation from in our lecture notes). A large value of C puts more weight on misclassification of data points (leading to a larger margin hyperplane), while a small value of C puts less weight on misclassified points (leading to a smaller margin). Examine the effect of adjusting C up and down.
    b. Examine how the performance changes as the training set is decreased to 50% and 30%. Use a consistent testing set of spectra for evaluating 70%, 50% and 30% training sets.
    c. Repeat (a) and (b) with a non-linear model, using a radial basis function kernel. You have an extra parameter to play with: γ. Investigate the sensitivity of the SVM to the value of γ. You don't need to try a large number of values, but enough that you can comment on the effect of changing γ, which affects whether the model considers only nearby points or a large range of points.
8. Repeat with the IR data and compare the performance of classification between mass spectra and IR data.
9. Study the effect of changing the number of principal components used. You don't need to look at all models, but pick one of your models from parts 7 and 8 and vary the number of principal components, as was done when performing k-means clustering.

10. Discuss the performance of Supervised and unsupervised classification. Describe some other methods that could be used.

In assessing classification and clustering, consider using some or all of the following metrics. These are based on the following four parameters, which can be calculated for each class:

- True positives (TP_i): the number of times a data point is predicted to belong to class i and does belong to that class;
- True negatives (TN_i): the number of times a data point is predicted not to belong to class i and does not belong to that class;

- False positives (FP_i): the number of times a data point is predicted to belong to class *i* but does not belong to that class;
- False negatives (FN_i): the number of times that a data point is predicted not to belong to class *i*, but actually does belong to that class.

Summary statistics

- Accuracy, $A_i = (TP_i + TN_i)/(TP_i + TN_i + FP_i + FN_i)$
- Recall for class $i = TP_i/N_i$
- Where N_i is the number of data points belonging to class *i*.
- Precision: $P_i = TP_i/(TP_i + FP_i)$. This is the fraction of times that the model is correct when it predicts that a data point belongs to class *i*.

You can also summarise results using a confusion matrix. https://en.wikipedia.org/wiki/Confusion_matrix

The assessment material you hand in should consist of the following:

1. A formal written report, length equivalent to about up to 8 A4 pages including figures and references (min font size = 12).
   You may also use an Appendix for additional figures if you want to include figures that do not fit in the 8-page limit.
2. Final notebooks. Submit copies of your final notebook(s) used to perform PCA, k-means clustering and SVM.