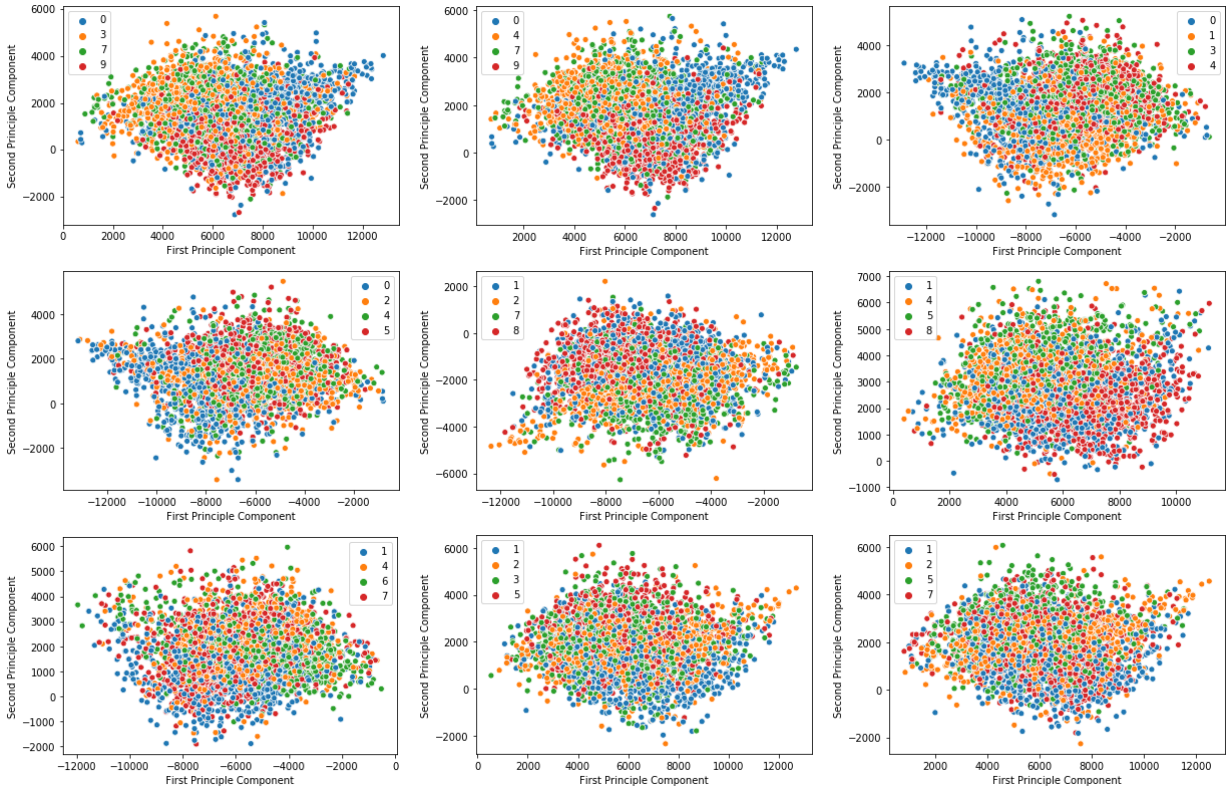
COMP4702/COMP7703/DATA7703 - Machine Learning
Homework 5 - Dimensionality Reduction
Solutions

Marcus Gallagher

Core Questions

1. The CIFAR-10 dataset is a widely-used benchmark dataset in machine learning (see https://www.cs.toronto.edu/~kriz/cifar.html for details). A subset of CIFAR-10 is available on the course blackboard site (`cifar10_data_batch1.mat`). Perform PCA on **four**[1] of the classes in this CIFAR-10 dataset. Submit a plot of the data projected onto the first two principal components. Use a different colour for each class.

   Answer: here are some samples:



[1] Choose classes using the four least significant digits of your student number. If you have duplicate digits, choose your other class(es) at random.

2. Fisher's Linear Discriminant Analysis is described in this week's lecture for a dataset with two classes. Alpaydin discusses (p.143-144) how to generalise this for $K > 2$ classes. Using the first two features of the Iris dataset, calculate the between-class scatter matrix (correct to four decimal places).

   The intention was to calculate S_B before the projection, but people might have done the between-class scatter after the projection. Before:

$$S_B = \begin{pmatrix} 63.2121 & -19.9527 \\ -19.9527 & 11.3449 \end{pmatrix}$$

   After: 50.0735 Matlab code:

```
%Script for HW5 Q2, 2020
load fisheriris.mat
x = meas(:,1:2);
%m is the overall mean
m = mean(x)
m1 = mean(x(1:50,:))
m2 = mean(x(51:100,:))
m3 = mean(x(101:150,:))
disp('Answer before projection:');
Sb = 50*((m1-m)'*(m1-m)) + 50*((m2-m)'*(m2-m)) + 50*((m3-m)'*(m3-m))
S1 = cov(x(1:50,:))
S2 = cov(x(51:100,:))
S3 = cov(x(101:150,:))
SW = S1+S2+S3
[v,d]=eig(inv(SW)*Sb)
%Solution is the largest eigenvectors, is this case just one because data
%was 2D
W = v(:,1);
%Finally, the between class scatter AFTER projection:
disp('Answer after projection:');
Sbafter = W'*Sb*W
```

3. What is the general type of optimisation algorithm used for training in t-SNE?

   (Stochastic) gradient descent.

Extension Questions

Guyon and Elisseeff[1] is a very well-known paper about feature selection. You will find it on the course blackboard site (under Books and Primary References). Please refer to it to answer the following questions.

4. Techniques for choosing features in machine learning are sometimes categorized as *wrappers* and *filters*. Explain in 3 sentences or less the difference between wrapper and filter methods.

   From the paper (p10): "Wrappers utilize the learning machine of interest as a black box to score subsets of variable according to their predictive power. Filters select subsets of variables as a pre-processing step, independently of the chosen predictor." So the essential point is that wrappers use

a specified ML technique (e.g. classifier, regressor) to perform feature selection and so performance depends on that technique. Filters don't.

5. One very common way of performing feature selection is to calculate the correlation coefficients between all pairs of features in the dataset and then remove features that have a very high (absolute) correlation value. Guyon and Elisseeff show a simple example where this is a bad idea (Fig.2b in [1]). After reading this example, explain in your own words (approx. three sentences) why this example demonstrates that selecting features based on their correlation might be a bad idea.

   The histograms in Fig.2b show that the classes overlap with respect to either feature. The scatter plots show that when both features are present, the classes are well-separated, despite very high correlation between the features. The means of the class-conditional distributions are not captured by the covariance/correlation.

1 References

[1] Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.