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

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.

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 **before the projection** (correct to four decimal places).

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

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.

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.

1 References

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

[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.