Machine Learning Exercises 21 (Friday 17 November)

Principal Component Analysis

The **crabs** data describes five morphological measurements on 50 crabs each of two colour forms and both sexes, of the species Leptograpsus variegatus collected at Fremantle, W. Australia[1]. The measurements are all in millimetres and are: Frontal lobe size, rear width, carapace length, carapace width, and body depth. The classes of interest are the four combinations of sex and species.

Before you start, split the data into training and test data.

Exercise 1. Perform a Principal Components Analysis on the training data in two ways:

1. Directly via an eigendecomposition (which you can use Python to find)
2. Via the `PCA` method in sklearn[2].

Familiarise yourself with the built-in PCA method so that you know where and how to find the relevant output: transformation matrix, variances, principal component scores. Python, as many other software packages, use something called the Singular Value Decomposition rather than the Eigendecomposition as it is computationally superior. You can always use the things you compute by hand to verify your understanding of the Python output.

Exercise 2. Make two scatterplot matrices (plot of pairs of features against each other) – one for the raw training data and one for the principal components – where observations are colored by class.

Exercise 3. Find the loadings of the first principal component and explain what they are and how to use them for interpreting the principal component.

Exercise 4. Find the variances of the principal components and make three plots

1. Variance against component number
2. Proportion variance explained against component number
3. Cumulative sum of variances for the first k components against component number k.

Discuss how the plots can be used to inform the choice of how many principal components to use in further analyses.

Exercise 5. Train a 5-nearest neighbours classifier based on the first two principal components.

Exercise 6. Compute the test error for your 5-nn classifier. Remember that you need to apply the PCA transformation obtained from the training data to the test data – do not perform a new PCA on the test data!

Note that the number of components k can be seen as a hyperparameter and could, as such, be selected by cross validation.

Exercise 7. Perform another PCA, this time where you first standardise each feature to have mean 0 and variance 1. Explain why standardising variables can be a good idea.

[1] Campbell, N.A. and Mahon, R.J. (1974) A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22, 417–425.

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html`

Exercise 8. Sphere the training data by first applying PCA and then scaling each principal component by its standard deviation. The result is data where not only features are uncorrelated, but they also have unit variance. Plot the sphered data in a scatterplot matrix or similar to see the effect of the transformation.

Sphering – also referred to as whitening – can be done directly in sklearn by specifying `whiten = True` in the call to `PCA`. Sphering can be used as a preprocessing step before applying other machine learning methods, such as knn.