Advanced Statistical Inference
Feature selection

1 Aims

- To pick some features from the leukaemia data.

2 Feature selection

Like last week, this lab is quite open-ended. On the course webpage, you can download a (slightly reduced) version of the leukaemia data. It consists of some training data `X` and associated labels `t` as well as independent test data `X_test` and `t_test`. Your task is to select some features and classify the data – in previous labs we have provided code for classification (SVM, KNN, Bayes) and you may use whichever you like.

2.1 Feature selection

You are free to use any of the feature selection schemes we covered on Tuesday – scoring the features, PCA, or clustering them (you should have K-means code from last week to do this). In each case, you should use the test set to investigate how performance varies as you change the number.

2.1.1 Scoring

Use the score described in the lectures (only on the training data) to get a set of features (corresponding to a subset of the original ones). You should make sure you extract this subset from the test data too!

2.1.2 Clustering

You can use the K-means code you wrote last week. Remember to pass the data the correct way around and only cluster the training data. Transforming the test data is a little more complex – your kmeans should give you assignments of features to clusters. Use these alongside the test data to compute the cluster means for the test data (you do not need to re-assign the clusterings). Ask me or Dom if this doesn't make sense!

2.1.3 Principal components

Matlab has a built in PCA function `princomp`. If you do the following:

```
[W,Z] = princomp(X)
```

`W` and `Z` will correspond to the symbols in the lecture notes. Note, it returns all of the possible principle components. To use just the first K:

```
Zk = Z(:,1:K);
X_testk = X_test*W(:,1:K); %These are the test points in the new space.
```