Exercise 12: Molecular Classification of Cancer

May 17, 2016

One of the challenges in cancer treatment is correct identification of the type of cancer in order to apply the most suitable treatment. We would like to test the feasibility of using gene expression data obtained by DNA microarrays to classify cancer. Our focus is on two types of leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL).

Our initial leukemia data set consisted of 38 bone marrow samples (27 ALL, 11 AML) obtained from acute leukemia patients at the time of diagnosis. The MATLAB file `ex12.mat` contains three variables:

- `expression` – a 7129x38 matrix containing the log2-scale data from the DNA microarrays.
- `genes` – a 7129x1 cell array with the gene names
- `true_labels` – a 38x1 matrix indicating the types of leukemia. 1 means ALL, and 2 means AML.

1 Visualizing high dimensional data

IMPORTANT: At the beginning of your script, initialize the random number generator using this seed `rng(2016);`

1.1 Univariate analysis

Start by performing the standard analysis using `mattest, mafdr` and `mavolcanoplot`. Use the volcano plot to find out which up regulated gene has the most significant p-value and the same for the down regulated genes. Use FOLDCHANGE threshold of 4.

Hint: Use `DataALL = expression(:, true_labels==1)` and `DataAML = expression(:, true_labels==2)` to define the two groups of samples.

1.2 Scatter plot using 2 genes

Find the index of the two genes from the previous questions. Plot the expression of these two genes using a scatter plot and use `true_labels` for the colors.

Hint: To find the indices use `find(ismember(genes, 'xxx'))`. Then select the corresponding rows from the `expression` matrix. Finally use the `scatter` function to make the plot.

1.3 Use k-means to see if it manages to distinguish the two clusters automatically

Using only the expression of the two genes you selected in 1.1, run the `kmeans` function and make another scatter plot with the resulting cluster labels. Do you spot any differences between the true labels and the clusters? How many samples were misclassified?

Hint: Remember that `kmeans` assumes that the rows are samples and the columns are feature dimensions. Therefore, the data from the `expression` matrix needs to be transposed.

1.4 Dimension reduction using PCA

Run PCA on the entire `expression` matrix and select the first two principal components. Project the original data on these two components using matrix multiplication. Then make a scatter plot using the resulting 2D matrix and use the true labels to color the points.

Hint: In order to get the two principal components run:

```
PC = pca(expression', 'NumComponents', 2);
```

Then to project the data on these principal components use:

```
projected_data = PC' * expression;
```

1.5 Use k-means to see if it manages to distinguish the two clusters automatically

Repeat what you did in Question 1.3, but instead of using two specific genes – use the projected data on the two principal components. Does k-means perform better now? How many misclassified samples do you observe this time?

1.6 Select only highly significantly changing genes

Now, combine the two approaches of selecting genes by p-value and PCA for dimension reduction. First, find the indices of those genes whose (FDR-

corrected) p-value is lower than 0.001. Then, select only these rows from the `expression` matrix and perform a PCA projection into 2D (same as in 1.4). Again, plot the result in a scatter plot and use the true labels for colors.

1.7 Use k-means to see if it manages to distinguish the two clusters automatically

Once again, use k-means on the newly projected values and plot the result in a scatter plot. Does k-means perform better now? How many misclassified samples do you observe this time?