COMP 364 - Tools for the Life Sciences

Assignment #3: Class Prediction
Due: Tues, March 29th, 2016
10% of total grade

The goal of the assignment is to explore using Naive Bayes' Classifiers (NBCs) to predict patient outcome using gene expression.

Question 1: [10% of assignment] Building the "default" model based on just clinical variables

Using the class notes (contained in `naiveBayes.R`), construct an Naive Bayes' Classifier (NBC) to predict `event.5` in `vanvliet` using ER, Her2, lymph node status and grade of the patient. You just need to re-do what was done in class except that you should build the model *only* on the first 450 patients (those that correspond to rows 1-450 in the clinical data). This is your *training set*.

The remaining (947 - 450 = 497) patients are used as the *validation dataset*.

Using the validation dataset, calculate the *accuracy* of your predictor. Report the number of true/false positive/negatives using the `table()` function as in the notes.

$$\text{accuracy} = \frac{\text{number of true positives} + \text{number of true negatives}}{\text{number of true positives} + \text{false positives} + \text{false negatives} + \text{true negatives}}$$

So a true positive is when the classifier predicts good outcome, and the patient is actually observed to have good outcome (event.5==FALSE). A true negative is when the classifier predicts poor outcome and that matches event.5==TRUE. False positive and negatives are the mistakes between predicted and observed.

https://en.wikipedia.org/wiki/Accuracy_and_precision

Question 2: [40% of assignment] Comparing predictors

Build a NBC as in Question 1 to predict `event.5` but instead of using clinical variables (e.g. ER, Her2, lymph and grade), use the 20 probes that are differentially expressed between good and bad outcome (`event.5`) as per Assignment #2 (smallest p-values from a t-test). Don't forget to report the gene names of these probes.

Using the same training and validation dataset as Question 1 here, report the accuracy of your predictor and use the `table()`function to display your results. Report the names of the genes associated with the 20 probes from above. Show your R code, of course.

Question 5: [20% of assignment] Across datasets

Using the genes from the NBC in Question 3, predict the outcome of patients in `curtis.discovery`. This will require that you train a new NBC using these genes in each of these datasets - use the first half for training and the rest for validation.

Comment on the performance and what you observe in a short paragraph. What difficulties did you encounter when trying to port your predictor to this new dataset?

Question 6: [20% of assignment] Leave one out vs half/half split

In the above questions, you used half of the dataset for training and half for validation. Only using half for training and leaving out half for validation might be overkill. Another approach is to use "leave one out cross validation (LOOCV)".

LOOCV proceeds as follows. We go through each patient in the dataset. So for `vanvliet` we would consider each one of the 947 patients. At the *i*th iteration, leave patient i out of the training set (patient i is the entire validation dataset), and build an NBC using the 10 most differentially expressed probes as per Question #3 using all patients except patient i. Then we predict the outcome of the single patient in the validation set: the *i*th patient.

The accuracy of LOOCV is computed from the number of true positives, true negatives, false positives and false negatives over the 947 different iterations.

Implement this procedure and try it out on `vanvliet`. Compare it to the performance of the predictor in Question #3.

Question 3: [50% of assignment] Predictors stratified by subtype

In this question, you will repeat Question 2 but instead of building an NBC across all patients, you will first stratify by clinical subtype: ER+/HER-, ER+/HER2+, ER-/HER2+, ER-/HER2-. Then you will build a predictor for each subtype using only the good and bad outcome patients of that subtype. For example, you will perform a t-test between good and bad outcome patients that are in the ER-/HER2- subtype, and then combine the top 20 into a NBC. Use all of the patients in each subtype (do not split into training and validation datasets).

Part A (25%): First report the gene names for the top 20 probes in each of the four NBCs, and the NBC from Question 2. Highlight any common genes (if any). If they are different genes, do they still have the same molecular function and role? Do the genes belong to different pathways or biological processes? Do the genes, pathways and processes change between the five predictors.

Part B (25%): Apply each of the five classifiers to each subtype. That is, apply the "unstratified" NBC from Question 2, to the ER+/HER-, ER+/HER2+, ER-/HER2+, and ER-/HER2- cohorts individually, and calculate the accuracy in each subtype. Then apply the ER+/HER- NBC to the ER+/HER2+, ER-/HER2+, and ER-/HER2- cohorts, and record the accuracy in each cohort. Repeat this for the remaining three classifiers. Report in a 5 by 4 table the accuracy of each of the 5 classifiers in each of the 4 subtypes. Are there any marked differences in accuracy? Can you suggest why? Are the classifiers better in the subtypes they were trained for? Why might that be?