Computational Molecular Medicine: Project

Due April 14, 2023

Apply any machine learning technique to learn a predictor for distinguishing among the particular classes (phenotypes, sub-types, etc.) described in your dataset. The choice of method is entirely up to you. It can be one studied in the course, or mentioned on the slides but not presented in detail or one you learned about elsewhere or even invented yourself. Your effort will be evaluated by various criteria, including creativity, (mathematical) coherence, parsimony and proper validation. Obtaining high accuracy is also valuable, but is very difficult and not necessary to do a good job.

You might want to see what happens if you filter the genes based on differential expression. If you are going to evaluate the effect of the level of filtering I suggest you just try orders of magnitude, like retaining 10 vs 100 vs 1000 genes. The number of genes used by your predictor is one aspect of "parsimony."

Think carefully about how you will evaluate your classifier. There are many possibilities corresponding to different partitioning schemes in n-fold cross-validation. Also, you might want to divide your original training set into three pieces: learning (your classifier), validation and testing. You can do all your training on the first part and then use the validation part to get a sneak preview of what you might see when you finally use your test, allowing you to go back and forth between learning and validation to make changes. Of course eventually this may not be very different than just taking all but the test data for training in one shot.

Finally, in order to allow us frame your results in a more clinically realistic setting, determine the specificity on your test that can be achieved while maintaining 80% (or 90% sensitivity). As usual, sensitivity (respectively, specificity) is the fraction of the more pathological class (resp., other class(es)) which are correctly classified. Usually, we want to maintain a high sensitivity at the expense of reduced specificity; for example, in a "progression" vs. "no progression" case, not treating a patient who

will progress (a false negative) is more serious than treating a patient who will not progress (a false positive). Summarizing the steps for this last type of analysis:

- Define a (real-valued) discriminant function ("score") $g(x)$ for your approach, where large values of $g(x)$ are associated with $Y = 1$. Whatever approach you use this has likely already been done, perhaps implicitly. For example, if you chose a k for the kNN classifier (or the $kTSP$ classifier), the natural discriminant function is simply the number of the k nearest neighbors with class 1 (number of pairs voting for class 1). If you built a random forest, the discriminant function could be the number of trees which vote for class 1. Virtually any method has a natural discriminant.

- Construct an ROC curve for your g using the test set. That is, apply the classifier $F_t(x) = \delta\{g(x) \geq t\}$ to the samples x of the test set, compute the sensitivity $sens(t)$ and specificity $spec(t)$ for each t, and plot $(1 - spec(t), sens(t))$.

- Find the point on the ROC curve where $sens(t) \geq .80$, i.e., the largest value of t for which

$$\hat{P}(g(X) \geq t | Y = 1) \geq .8,$$

  where $\hat{P}$ refers to the estimate based on the test set. Call this point t_{80}. Your specificity at 80% sensitivity is then $spec(t_{80}) = \hat{P}(g(X) < t_{80} | Y = 0)$.

This is actually an optimistic estimate of $spec(t_{80})$ because in practice you would need to estimate the threshold t_{80} *based on the training data.* But this still gives us a single number for each approach. (Again, you are not being "graded" based on the specificity you report.) Just show the ROC curve and report $spec(t_{80})$.

Final note: present your results in a civilized form. In particular, do **not** insert any code in the presentation; the report should be readable for somebody who knows nothing about programming. Also, divide the report into sections, something like Intro (where you state the problem being addressed, briefly describe the data and perhaps summarize the proposed method of prediction), Preliminary analyses (any feature selection or filtering to reduce dimensionality, etc.), Results (with appropriate lists, tables and figures, and any biological or medical connections of interest which contextulize the work), Methodology (for details about your particular prediction method, especially if not standard), and Conclusions (what did you learn?).