CO902
Written Report based on Computer Project
Due in class 9:00am Monday 11 Feb

The MATLAB file `project_data.mat` (available from the course website) contains data obtained from a cancer study. The abundance of each of 70 proteins was measured across 500 tumour samples obtained from different patients. Each of these 500 tumours was then treated with an anti-cancer drug called "U0126". Some samples responded, in the sense that the drug was successfully able to kill off the tumour cells, others did not. There is currently no good way to tell, ahead of time, whether or not a drug of this kind will work in a specific case. The only way to find out is to actually undertake the course of treatment, which can, unfortunately, lead to serious, even devastating, side-effects, often with no benefits. A question of great interest then is whether molecular measurements can be used to predict drug response. Hence, the goal is to build a predictor of U0126 response from the protein measurements.

The MATLAB file contains three items: (i) a 70 × 500 binary matrix $\mathbf{X}$, (ii) a 1 × 500 binary row vector Y and (iii) a list of names `proteins`. Matrix $\mathbf{X}$ contains protein measurements binarised into present/absent calls: that is, $X_{ij} = 0/1$ indicates the absence/presence of protein i in tumour j. Row vector Y contains labels $Y_j = 0/1$ which record whether the drug U0126 was unsuccessful/successful in treating tumour j. The variable `proteins` contains[1] the names of the 70 proteins under study (these names are provided for completeness, and should not play any role in your analysis).

1. Build a "Naive Bayes" classifier which, given binary protein measurements as inputs, predicts whether or not the drug U0126 is likely to work. (The input to the classifier is a vector $X_i \in \{0,1\}^{70}$ and its output is a probability $P(Y_i = 1|X_i)$. For the purpose of making a prediction, this probability can easily be mapped to a binary decision $\hat{Y}_i \in \{0,1\}$. You may assume that the class prior is flat, i.e. $P(Y_i = 1) = P(Y_i = 0) = 0.5$ and that the protein data X_i are independent.)

2. The data $\{\mathbf{X}, Y\}$ in the file `project_data.mat` is the only relevant data available. It is therefore important to make best use of the data to evaluate the predictive ability of your classifier. Once you have written the required code, test your classifier on the dataset using *leave-one-out cross-validation*. To be clear, since there are 500 data vectors, your cross-validation will require 500 steps: at each step the training set will consist of 499 datapoints with the remaining datapoint held out for testing. Accuracy is obtained by computing (over the 500 tumors) the proportion of how many times your classifier is correct.

3. It is highly unlikely that all 70 proteins play a central role in determining the response to U0126. Furthermore, there is much scientific interest in understanding which proteins

[1] The variable proteins is a MATLAB "cell-array". Typing `proteins{12}` will return the name of the 12^{th} protein.

are most predictive of response. Can you think of a way to determine *which* of the proteins are especially important in your classifier? (If possible, provide some results, but a good discussion of *how* you would address this question will be sufficient).

4. Are you happy with the accuracy rate you obtained with cross validation? In words (i.e. without writing any more code!) can you suggest any ways to improve your predictor? Why do you think the extension(s) you suggest would actually buy you anything in terms of predictive ability over the vanilla Naive Bayes classifier?

Your submission should consist of:

- A written report. This should be in 12pt and no more than 4 sides A4 (including figures), but should contain a concise, mathematical description of your classifier, a presentation of your empirical results, and a discussion of these results and possible extensions to your classifier.

- Computer code. Please append a printout of your computer code to the report. Without going over-board, please ensure it is documented... the beginning of each file should describes what the code does, and each block of code should have at least minimal description.

Please consider ...

- The mathematical description should be concise, but sufficient for an informed reader to go away and replicate your results. While a derivation of everything involved is certainly not required nor desired, it's not enough to say you use the "usual" results, without stating, however briefly, what these results are. Thus "...we then make use of the usual estimators for the parameters..." is not enough, but "...we then make use of the usual estimators for the parameters, namely: (one set of equations)" is fine.

- The report should not describe your work in a chronological fashion: rather, it should concisely summarise and present your method, findings and insights. In particular, please use the standard format of scientific articles, using 4 main sections: Introduction, Methods, Results, Discussion.

  (a) The **Introduction** sets the scene and spells out what you aim to do and (when there is no abstract) briefly summarises the results. It brings the reader on board.

  (b) The **Methods** dryly describe what tools you use and the steps you take. The methods *should not* say anything/much about the outcome or "Results". This of course is artificial, because in practice you iteratively get results and change/build your method. But, again, the format is *not* chronological, and this artifice is essential to help your reader rapidly digest your work.

  (c) The **Results** dryly describe what happened when you did what you said you were going to do in the Methods section. There *should not* be speculative or wide-ranging interpretation of the findings. That's for the next section.

(d) The **Discussion** pulls everything together, giving more broad-ranging interpretation of the results, insights gained, and offers possible directions for future work. By contrast to the Methods and Results, this can be more 'lively'.

I found the article "The Science of Scientific Writing" by Gopen & Swan (on the webpage) immensely helpful for my own writing; I encourage you to read it and put its principals to work in your own writing.