BIOL-GA 2031. Statistics and Machine Learning in Genomics Homework 02

TA. Jaime Cascante Vega
Email. jc12343@nyu.edu

In the jupyter notebook LOGISTIC_REGRESSION_PY.IPYNB I wrote some Python code for loading the data and performing a logistic regression + Bootstrapping on some Dengue disease data. You could use that code I wrote or write one yourself to answer the following numerals. The equivalent R code I wrote can be found in LOGISTIC_REGRESSION_R.IPYNB, I'm not super versed in R so that code might have a lot of bits to improve.

1. A bootstrap sample of a dataset of size N $[x_1, x_2, x_3, \cdots, x_N]$ correspond to sample with replacement with equal probabilites each x_i and forming a new dataset $[b_1, b_2, \cdots, b_N]$. In the $N = 2$ case we have $[x_1, x_2]$ and the possible bootstrap samples are $[x_1, x_1]$, $[x_1, x_2]$, $[x_2, x_2]$, $[x_2, x_1]$.

   How many possible bootstrap samples are in a dataset of size N?

2. Let X_1, X_2, $\cdots$, X_n be independent and identically distributed (iid) random variables with mean μ and variance σ^2. Consider a bootstrap sample of those variables of size n denoted by Y_1, Y_2, $\cdots$, Y_n.

   2.1 Calculate $\mathbb{E}[Y_i]$ and $\text{Var}[Y_i]$.

   2.2 Denote $\bar{Y}$ the mean of the bootstrap sample as show below. Calculate the conditional expected value of the mean and variance of the bootstrap sample given the original dataset, i.e. calculate $\mathbb{E}[\bar{Y}|X_1, X_2, \cdots, X_n]$ and $\text{Var}[\bar{Y}|X_1, X_2, \cdots, X_n]$.

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

3. How does the first two moments, mean and variance, of the prediction accuracy change with the numbers of bootstraps?

4. Use just the 5 genes significant in the ANOVA (volcano plot) and do the bootstrapping+logistic regression again. How does the estimate of the expected prediction accuracy change? Use Figures to present and discuss your results.

5. Repeat 1. again but just with the genes significant in the ANOVA.

6. Compare 1. and 3., and discuss. Does the previous knowledge about the significantly expressed genes during disease contribute to the estimate of the expected prediction accuracy? Did you require a lower number of bootstrap samples?

7. Write the leave-one-out cross validation code with partitions of 10 each (some will have more because there are 53 data-points).

   7.1 How does the expected estimated prediction accuracy compare to that one with bootstrapping? Use figures to present your results and discuss.

   7.2 How does the variance of the prediction accuracy compare to the one estimated with bootstrapping?