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

Prof. Manpreet S Katari
TA. Jaime Cascante Vega
Email. jc12343@nyu.edu

1. A bootstrap sample of a dataset $\mathscr{D}$ of size N, $\mathscr{D} = [x_1, x_2, x_3, \cdots, x_N]$ correspond to sample with replacement with equal probabilities each x_i and forming a new dataset $\mathscr{B} = [b_1, b_2, \cdots, b_N]$.
   In the $N = 2$, $\mathscr{D} = [x_1, x_2]$ and the possible bootstrap samples are $\mathscr{B}_1 = [x_1, x_1]$, $\mathscr{B}_2 = [x_1, x_2]$, $\mathscr{B}_3 = [x_2, x_2]$, $\mathscr{B}_4 = [x_2, x_1]$.

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

2. Let X_1, X_2, …, 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, …, Y_n.

   If you feel more comfortable you could start with $n = 5$, then increase it to $n = 10$ and lastly to $n = 100$ and provide answer for both.

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

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

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

   For the next problems use the gene expression data, q-PCR, of 35 candidate genes on patients suspected of Dengue disease. The classifications for disease are Dengue Shock Syndrome (*DSS*), Dengue Hemorrhagic Fever (*DHF*), and Dengue Fever (*DF*). The goal is to potentially identify a subset of genes to use as biomarkers. For this train a logistic regression model, using only two categories: *healthy* and *disease*. Use the 3 classifications of disease, *DSS*, *DHF* and *DF* as the disease category. Use a bootstrap to estimate the expected prediction accuracy, i.e. accuracy of prediction averaged over multiple bootstrapped test sets.

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

4. In the Figure 1 below, I present an ANOVA comparing the expression of the genes during disease. Use just the 5 genes significant in the ANOVA (volcano plot shown in the Figure below) 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.

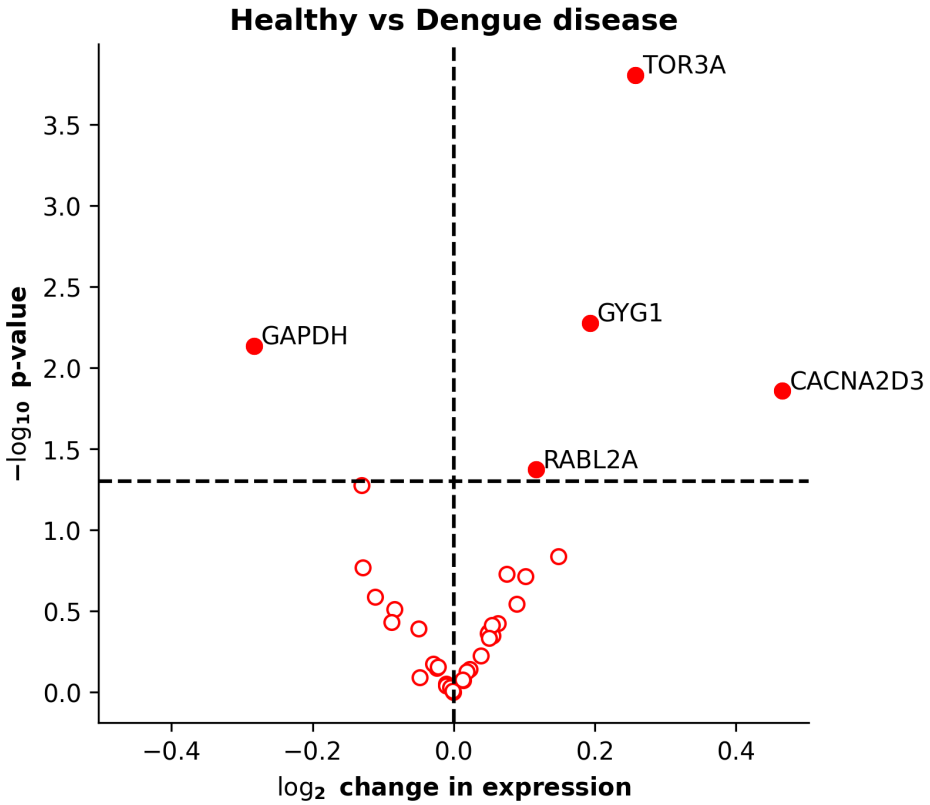

Figure 1: **Volcano plot.** Genes with increased expression are presented to the right of vertical dashed line, and those in healthy patients to the left. The significance line p-value $= 0.05$ is highlighted with a horizontal blacked dashed line. The genes significant are highlighted by filled circles, and with the name of the genes.

5. Repeat 3. again but just with the genes significant in the ANOVA.

6. Compare 3. and 5., 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. Divide the dataset in folds, each with 10 data-points (some will have more because there are 53 data-points). Estimate the expected prediction accuracy using a leave-one out 5 fold cross-validation procedure.

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?