Statistical Learning

Prediction error and information criteria

Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N\left(0, \sigma = \frac{1}{\sqrt{1000}}\right)$. Then generate the vector of the response variable according to the model

$$Y = X\beta + \epsilon \ ,$$

where $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$ and $\epsilon \sim N(0, I)$.

1. Perform the following analyses using the model with

   i) 2 first variables

   ii) 5 first variables

   iii 10 first variables

   iv) 100 first variables

   v) 500 first variables

   vi) all 950 variables.

   - For each of the considered models
     a) Estimate β with the Least Squares method and calculate residual sum of squares and the true expected value of the prediction error

        $$PE = E||X(\beta - \hat{\beta}) + \epsilon^\star||^2 \ ,$$

        where $\epsilon^\star \sim N(0, I)$ is a new noise vector, independent on the training sample.
     b) Use the residual sum of squares to estimate PE assuming that σ is known and replacing σ with its regular unbiased estimator.
     c) Estimate PE using leave-one-out crossvalidation (do not perform analysis 1000 times but apply the formula for leave-one-out cross-validation error provided in class).
   - Select the optimal model using two versions of AIC: for known and unknown σ.

   - Repeat the above calculations 100 times and
     * for each of the considered models compare the boxplots of $\hat{PE} - PE$ for three estimates of PE, mentioned above.
     * Provide histograms of the number of false negatives and false positives produced by both versions of AIC (with known and unknown σ).

2. Use BIC, AIC, RIC, mBIC i mBIC2 (you can use *bigstep* library in R) to identify important covariates when the search is performed over the data base date consisting of

   i) 20 first variables

   ii) 100 first variables

   iii) 500 first variables

   iv) all 950 variables.

   a) Report the number of false and true discoveries and the square error of the estimation of the vector of expected values of Y : $||X\beta - \hat{Y}||^2$.

b) Repeat point a) 100 times and report the estimated power, FDR and mean squared error of the estimation of expected values of Y for all criteria considered above. Critically summarize the results.

3. Data set realdata.Rdata contains the expression levels of 3221 genes for 210 individuals.

   a) Randomly select 30 individuals for the test.

   b) Use the training set (180 individuals) to construct the regression model explaining the expression level of gene 1 (first column in the data set) as the function of expression levels of other genes. Select explanatory variables using AIC, BIC, mBIC and mBIC2. Test the accuracy of your model predictions on the test set. Which criterion yielded the best prediction results.

Malgorzata Bogdan