Mathematics and Big Data – STATISTICS part B . Yo can do the excercises in groups of at most 3 people.

Exercise 1

Solve exercises 3, 4, 5 and 10 from CHAPTER 6 of the 1st Edition of the book **An Introduction to Statistical Learning and Applications, James, Witten, Hastie, Tibshirani**

Exercise 2

The following table, represents the set of counts, for a particular threshold, of a statistical test applied to some number of features. Let significant and non-significant represent the results of a particular test with the given threshold, and positive represent the number of truly alternative features, and negative represent the number of truly null features.

	non-significant	significant
negative	1945	54
positive	188	192

(a) How many false positives, false negatives, true positives, and true negatives are there?
(b) What is the false positive rate for this threshold? What is the false discovery rate?
(c) If you are going to give the set of significant features to your collaborator, what are you going to tell that collaborator about the make-up of that set in terms of the false positive rate and the false discovery rate?
(d) What are the sensitivity and specificity in terms of these results?

Exercise 3

We are going to study the `diabetes` dataset (Efron, Hastie, Johnstone and Tibshirani (2003) "Least Angle Regression" (with discussion) *Annals of Statistics*) which is part of the `lars` package.

The dataset consists of patient level data on the progression of diabetes.

The dataset has three matrices `x`, `x2` and `y`. While `x` has a smaller set of independent variables, `x2` contains the full set with quadratic and interaction terms. `y` is the dependent variable which is a quantitative measure of the progression of diabetes.

(a) Study the relationship of each of the predictors with the dependent variable.

(b) Regress `y` on the predictors in `x` using OLS. We will use this result as benchmark for comparison.

(c) Use the `glmnet` function to plot the path of each of x's variable coefficients against the L1 norm of the beta vector. This graph indicates at which stage each coefficient shrinks to zero. Identify the most relevant coefficients.

(d) Use the `cv.glmnet` function to get the cross validation curve and the value of lambda that minimizes the mean cross validation error.

(e) Using the minimum value of lambda from the previous exercise, get the estimated beta matrix. Indicates which predictors are important in explaining the variation in y.

(f) To get a more parsimonious model we can use a higher value of lambda that is within one standard error of the minimum. Use this value of lambda to get the beta coefficients. How many non-zero coefficients are left?

(g) As mentioned earlier, `x2` contains a wider variety of predictors. Using OLS, regress `y` on `x2` and evaluate results.

(h) Repeat the previous analyses for the new model.

Exercise 4

The dataset `Golub_Merge` that can be obtained from library library(`golubEsets`) (that has to be downloaded from Bioconductor) contains a famous dataset which has the expression levels of 7,129 genes from 72 subjects, with two different types of diabetes (ALL and AML).

golubMerge {golubEsets} R Documentation

Combined Test and Training Sets from the Golub Paper

Description

golubMerge is deprecated. use Golub_Merge instead.

The data are from Golub et al. These are the combined training samples and test samples. There are 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). The samples were assayed using Affymetrix Hgu6800 chips and data on the expression of 7129 genes (Affymetrix probes) are available. The data were obtained from the Web site listed below and transformed slightly. They were installed in an `exprSet`.

Usage

```
data(golubMerge)
data(Golub_Merge)
```

Format

There are 11 covariates listed.

- `Samples`: The original sample numbers.
- `ALL.AML`: Whether the patient had AML or ALL.
- `BM.PB`: Whether the sample was taken from bone marrow or from peripheral blood.
- `T.B.cell`: ALL arises from two different types of lymphocytes (T-cell and B-cell). This specifies which for the ALL patients; it is `NA` for the AML samples.
- `FAB`: FAB classification.
- `Date`: The date the sample was obtained.
- `Gender`: The gender of the patient the sample was obtained from.
- `pctBlasts`: An estimate of the percentage of blasts.
- `Treatment`: For the AML patient and indicator of whether the treatment was a success.
- `PS`: Prediction Strength.
- `Source`: The institution that provided the samples.

Source

http://www-genome.wi.mit.edu/mpr/data_set_ALL_AML.html, after some anonymous Bioconductor massaging

References

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, Science, 531-537, 1999, T. R. Golub and D. K. Slonim and P. Tamayo and C. Huard and M. Gaasenbeek and J. P. Mesirov and H. Coller and M.L. Loh and J. R. Downing and M. A. Caligiuri and C. D. Bloomfield and E. S. Lander

Identify the genes that are differently expressed (and thus help identify the type of diabetes) with both FDR and LASSO Techniques. Further references can be found at http://cs229.stanford.edu/proj2019spr/report/61.pdf