BBS741 Project #2

Deadline: November 12, 2018 @ 11:59 pm
Email to BBS741.2018@gmail.com
with LastName_FirstName_Project1 as subject line

Part 0 (0 points)

We will be using two python packages for this project:

1. The `statsmodels` python package (which you will need to download and install)
2. The `sklearn.linear_model` package (is available through Anaconda)
    a. LinearRegression
    b. LassoCV

Overview of data:

As detailed in class, we have a matrix containing transcription factor (TF) ChIP-seq signal for 127 TFs at 10,407 ccREs on chromosome 1. In the second column of the matrix, we have the signal for CTCF. In columns 2-128 we have the signal for 126 other TFs. Our goal is to predict CTCF signal at ccREs active GM12878 cells using other TFs.

Part I - Determining coefficient (25 points)

Our first step is to determine which TFs are most predictive for CTCF binding. Build an OLS regression model that models CTCF signal (y) as a linear combination of the 126 TF signals (x's).

What is the R^2 value of this model?

What are the five most important features of the model?

What is the biological significance of these TFs? Would we expect them to predict CTCF binding?

HINT: Use the OLS function as part of the statsmodel package with the summary function to get details about the model.

Part II - Regularization and cross validation (15 points)

We will now use a LASSO regression model with 10 fold cross validation.
HINT: LassoCV will implement this is one step

What is the R^2 value of this model?
How many features were dropped in this model (i.e. how many of the original 126 now have a coefficient of zero?)

Part III - Minimal model (10 points)

Now let's say we want a model that uses fewer input features since we do not have all of these TFs surveyed in every cell type. Select the top 10 features from your original OLS analysis, and construct a new minimal OLS model (model #3).
HINT: You may either implement this with the statsmodel OLS function or scikit learn LinearRegression.

What is the R^2 value of this new ten feature model?

Part IV- Estimating confidence interval of R^2 (25 points)

Using the bootstrap method, estimate the 95th confidence interval of R^2 for model #3. This part will require you to write code that will:

1. Randomly select 10,407 ccREs with resampling
2. Calculate R^2 for the ten feature linear regression model using these data points
3. Repeat process 10,000 times, calculating the R^2 each time
4. Calculate 5th and 95th percentiles → your script should output these values

Make a histogram of the R^2 values from the 10,000 iterations with lines showing the 5th and 95th percentiles. You may use any graphing software you like (i.e. python, R, matlab) but make sure to include your code.

Part V- Applying models to other chromosomes (25 points)

Using the three models (full OLS, LASSO, and ten feature model) which you created in parts I, II, and III, predict the CTCF signals at ccREs on chromosome 2 (15 pts).

Which model has the best performance, (i.e. which has smallest mean squared error)?

Grading Rubric

Submissions received after **November 12, 2018 @ 11:59 pm** will earn ½ credit

Part I (25 pts)	Script correctly implements OLS model	10 pts
	Script outputs correct R^2	5 pts
	Student identifies five most important features	5 pts
	Correct biological interpretation of features	5 pts

Part II (15 pts)	Script correctly implements LASSO model with 10 fold CV	10 pts
	Script outputs correct R^2	5 pts
Part III (10 pts)	Script correctly implements ten feature model	5 pts
	Script outputs correct R^2	5 pts
Part IV (15 pts)	Script correctly implements Bootstrapping to estimate R^2f and outputs confidence interval	15 pts
	Correct histogram of R^2 values with labeled 5th and 95th percentiles	10 pts
Part V (25 pts)	Script correctly applies three models to chr2 test set	15 pts
	Script correctly calculates MSE and identifies model with lowest MSE	10 pts
Total		**100 pts**