Homework #3

October 19, 2017

This is due at the beginning of class on **Tuesday, October 31**. Working in groups is fine but the write-up must be entirely your own and you should *list all collaborators* on the cover sheet of your submission.

1. For an $m \times n$ matrix A verify that
   (a) A^TA and AA^T have real eigenvalues
   (b) A^TA and AA^T are diagonalizable (recall this means that a matrix is diagonalizable if there's a basis of eigenvectors of the matrix).
2. In class we claimed that $\hat{\theta}_{ridge} = (X^TX + \lambda I)^{-1}X^T\mathbf{y}$ where X is the design matrix and $\mathbf{y}$ is the vector of responses. Justify this claim.
3. Prove that $\|\hat{\theta}_{ridge}\|$ increases as $\lambda \searrow 0$. Is this true for $\|\hat{\theta}_{lasso}\|$ as well?
4. The median of a random variable X is the number m such that
$$\mathbb{P}(X > m) = \frac{1}{2}$$
Consider a sample of N iid uniformly distributed samples from B_d, the unit ball in d dimensions. Prove that the median distance from the origin to the closest point is given by
$$m_{N,d} = (1 - (\frac{1}{2})^{\frac{1}{N}})^{\frac{1}{d}}$$
Use this to show that when you have 500 uniformly drawn samples from a 10-dimensional unit ball, more than half are closer to the boundary than to the origin.
5. Before beginning, read about implementation of the regularized regression methods in the scikit-learn Python module
   `http://scikit-learn.org/stable/modules/linear_model.html`
   As before, this is to be done in the Python programming language. Print and attach both your sample code and plots. As in the last assignment, select a stock of your choice and download a list of at least 100 quotes. Transform the quotes into returns via
$$\mu_{t_i} = \log(\frac{S_{t_{i+1}}}{S_{t_i}}), \qquad i = 1, ..., N \geq 100$$

Separate out from your dataset 20% of the data to be used as a validation/testing set which you are not to use in the regression. Using the built-ins perform linear regression as well as regularized regression using ridge and LASSO for at least 5 different choices of the regularization parameter. Compare accuracy for each of the models on the validation/test set.