GENERAL ASSEMBLY

DATA SCIENCE
11 WEEK PART TIME COURSE

Week 4 - Regularization
Wednesday 13th April 2016

AGENDA

1. Motivation / Review
2. What is Regularization?
3. Why use Regularization
4. Lab
5. Discussion

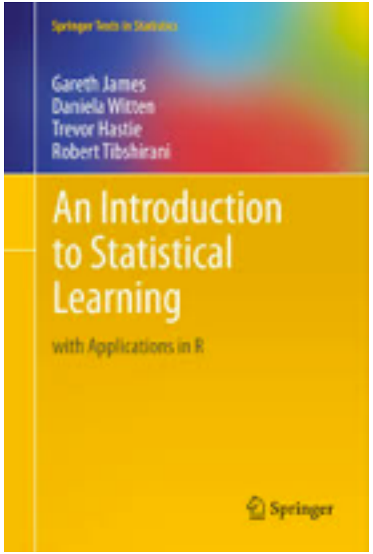

Produse
INTERNATIONALE
TABASCO
SUSHI RICE

HOMEWORK

Two parts of the Homework related to this lesson

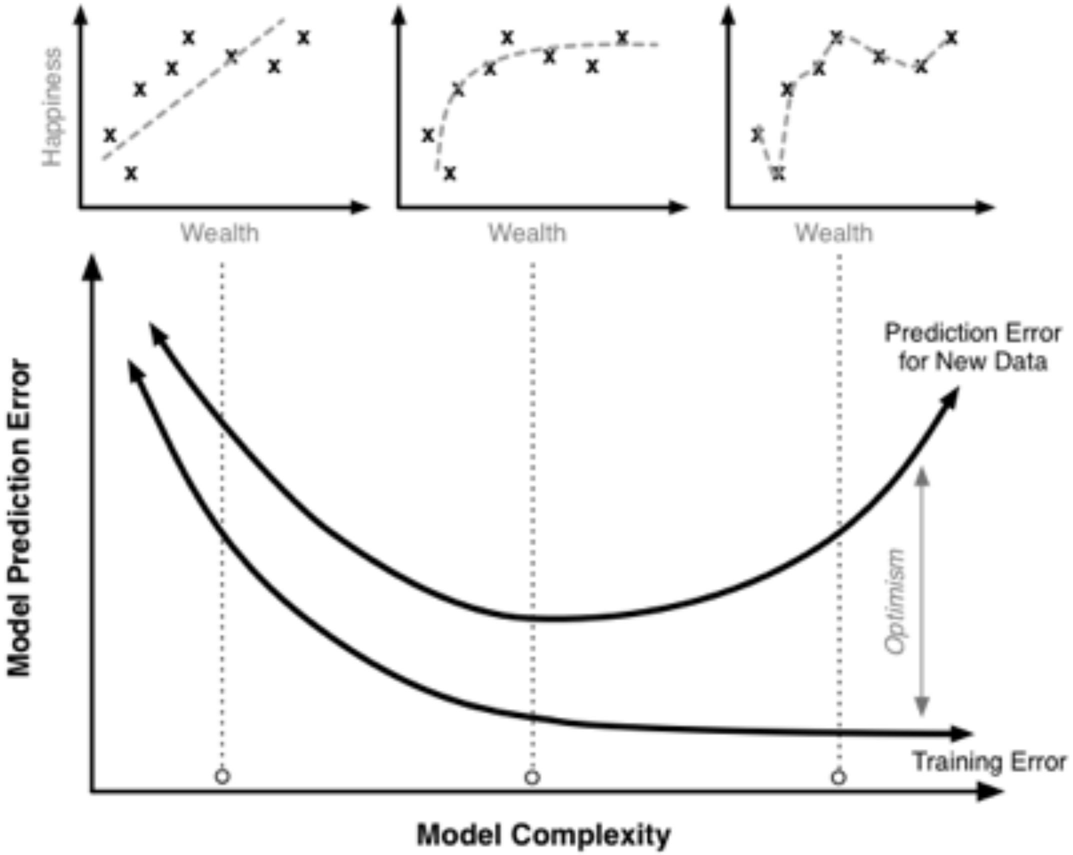
- Homework 2 - Chapter 6 of Introduction to Statistical Learning, Linear Model Selection and Regularization
- Task list - Data Robot Article, Regularized Linear Regression with scikit-learn

- Describe 3 ways we can select what features to use in a model?
- Why would we use regularization?



DataRobot

BIAS – VARIANCE TRADEOFF



We could fit a separate linear regression model for every combination of our features.

But what happens when we have a large number of features?

Computation time becomes a factor and we also need to consider that as we include more features we are increasing the chance we include a variable that doesn't add any predictive power for future data.

- A tuning parameter lambda (or sometimes alpha) imposes a penalty on the size of coefficients.
- Instead of minimizing the "loss function" (mean squared error), it minimizes the "loss plus penalty".
- A tiny alpha imposes no penalty on the coefficient size, and is equivalent to a normal linear model.
- Increasing the alpha penalizes the coefficients and shrinks them toward zero.

Recall from Week 2 that the least squares procedure estimates coefficients that minimise

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

Regularization (or Shrinkage) is a way to constrain the estimates of beta to be close or equal to zero.

Ridge Regression is similar to least squares, except we include a penalty term,

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\beta_j^2=\mathrm{RSS}+\lambda\sum_{j=1}^{p}\beta_j^2,$$

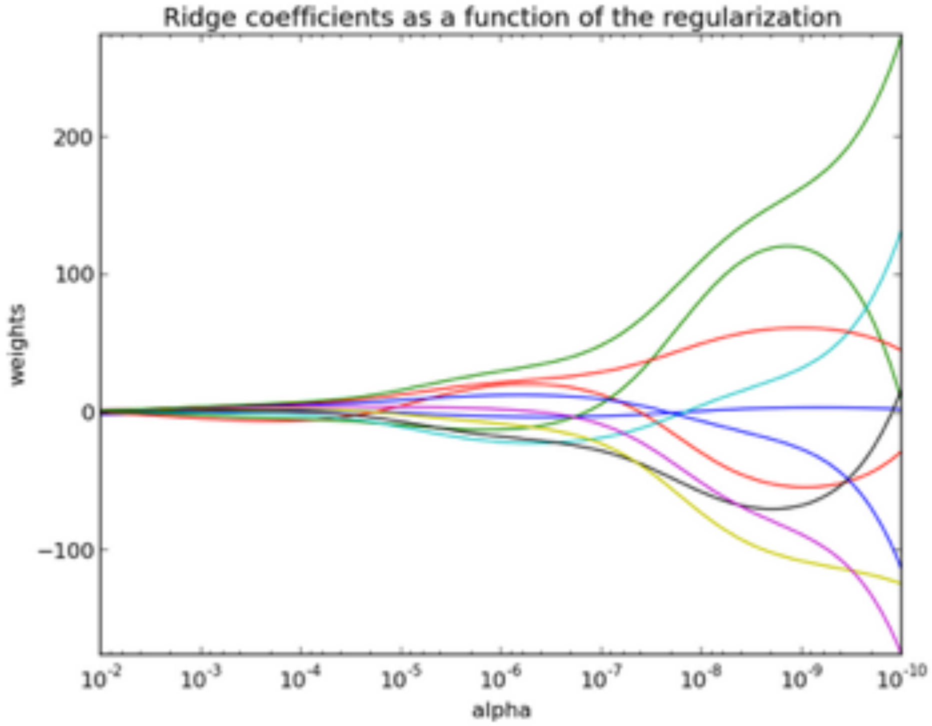
the λ term is a tuning parameter. When it is zero we get least squares, as it increases the term, $\lambda\sum_{j=1}^{p}\beta_j^2$ (the shrinkage penalty) has more of an impact and the coefficients will *approach* zero.

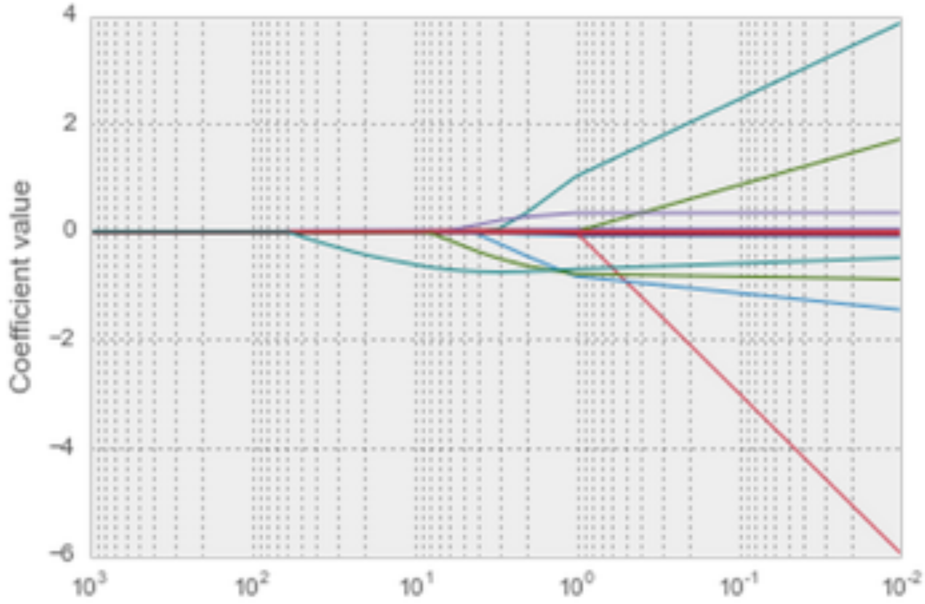
Lasso Regression is similar to Ridge Regression, except we have the absolute value of beta in our penalty term,

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

the λ term is a tuning parameter. When it is zero we get least squares, as it increases the term, $\lambda\sum_{j=1}^{p}|\beta_j|$ (the shrinkage penalty) has more of an impact and the coefficients will **equal** zero.

RIDGE VS LASSO





Lasso regularization is useful if we believe many features are irrelevant, since a feature with a zero coefficient is essentially removed from the model. Thus, it is a useful technique for feature selection.


Produse
INTERNATIONALE
TABASCO
SUSHI RICE

DATA SCIENCE PART TIME COURSE

LAB

git remote -v

git remote add upstream https://github.com/ihansel/SYD_DAT_3.git

git remote -v

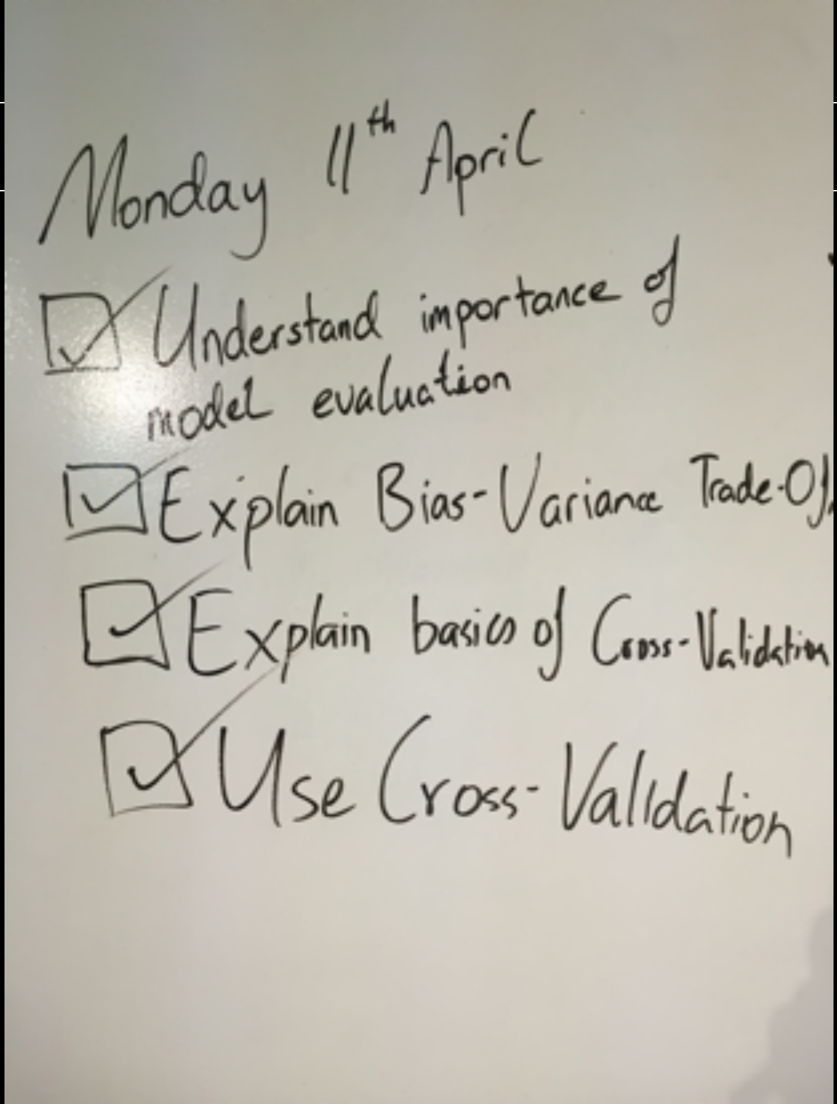
git fetch upstream

git checkout master

git merge upstream/master

OR git reset –hard upstream/master

WEEK 4


Monday 11th April
Understand importance of model evaluation
Explain Bias-Variance Trade-Of
Explain basics of Cross-Validation
Use Cross-Validation

READINGS

Read the following before class on Monday

- Clustering Methods in Introduction to Statistical Learning, Chapter 10.3 (15 pages)
- Python Notebook on Clustering http://nbviewer.ipython.org/github/nborwankar/LearnDataScience/blob/master/notebooks/D1.%20K-Means%20Clustering%20-%20Overview.ipynb

DISCUSSION TIME

Free scope. Anything you would like to talk about? Can be anything, e.g.

- Software
- News Articles
- Things you'd like to cover in the course
- Things you've been thinking about trying out