

# **DATA SCIENCE**

## **10 WEEK PART TIME COURSE**

**Week 4 – Regularization**  
**Tuesday 13th June 2017**

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



## **DATA SCIENCE - Week 4 Day 1**

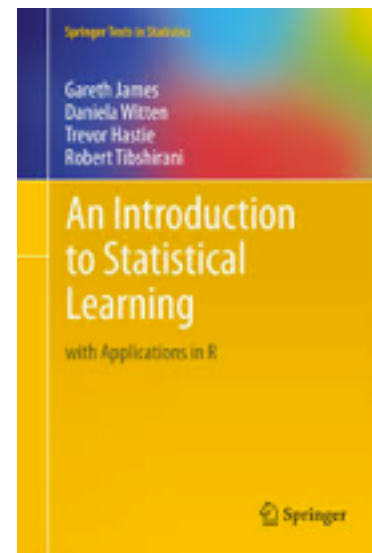
---

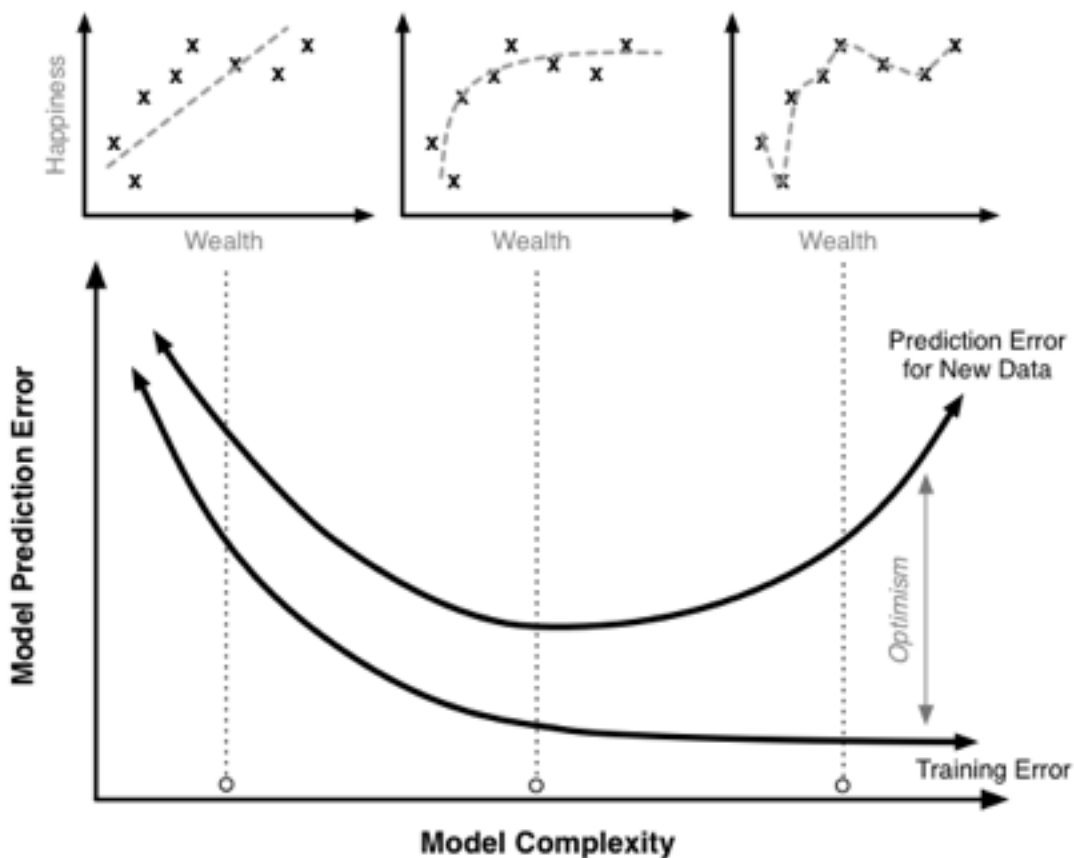
# **PRE-READING**

**Two parts of the Homework related to this lesson**

- **Homework 2 - Chapter 6 of Introduction to Statistical Learning, Linear Model Selection and Regularization**

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





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

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

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,

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

the  $\lambda$  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.

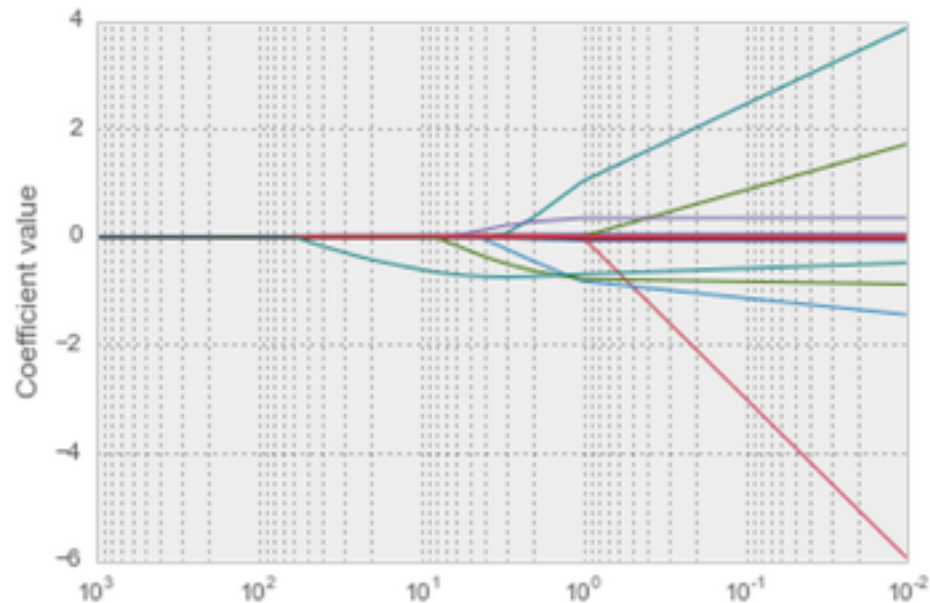
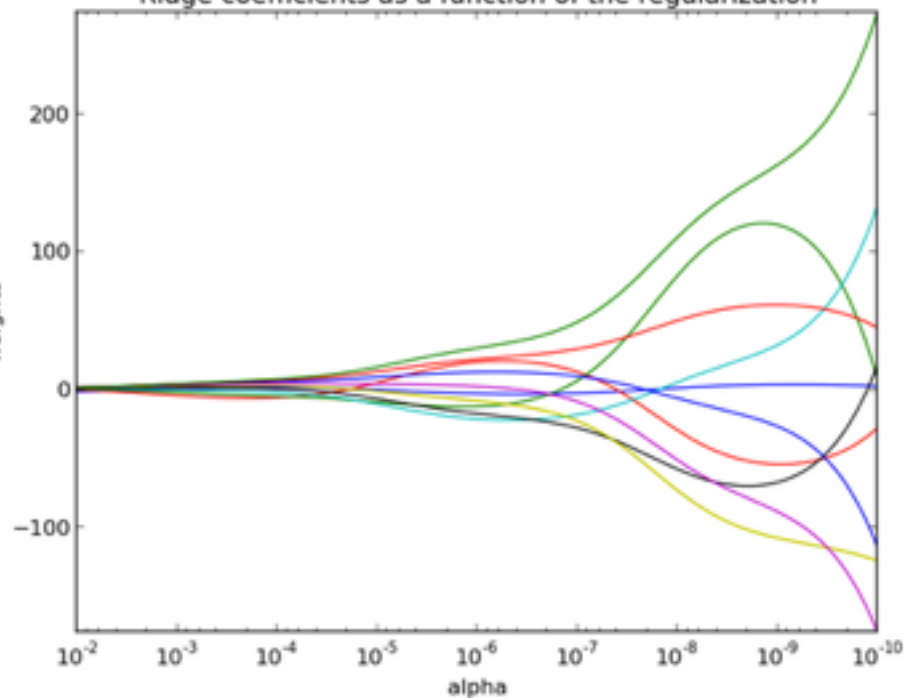
Lasso Regression is similar to Ridge Regression, except we have the absolute value of beta in our penalty term,

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

the  $\lambda$  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 coefficients as a function of the regularization



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.







1. re-name your labs with lab\_name.<yourname>.ipynb (to prevent a conflict)
2. cd <path to the root of your SYD\_DAT\_8 local repo>
3. commit your changes ahead of sync
  - git status
  - git add .
  - git commit -m "descriptive label for the commit"
  - git status
4. download new material from official course repo (upstream) and merge it
  - git checkout master (ensures you are in the master branch)
  - git fetch upstream
  - git merge upstream/master





**DATA SCIENCE PART TIME COURSE**

**LAB**



Many variables are included so that algorithms that select or learn weights for attributes could be tested. However, clearly unrelated attributes were not included; attributes were picked if there was any plausible connection to crime (N=122), plus the attribute to be predicted (Per Capita Violent Crimes). The variables included in the dataset involve the community, such as the percent of the population considered urban, and the median family income, and involving law enforcement, such as per capita number of police officers, and percent of officers assigned to drug units.

Data is described below based on original values. All numeric data was normalized into the decimal range 0.00-1.00 using an Unsupervised, equal-interval binning method. Attributes retain their distribution and skew (hence for example the population attribute has a mean value of 0.06 because most communities are small). E.g. An attribute described as 'mean people per household' is actually the normalized (0-1) version of that value

Attribute Information: (122 predictive, 5 non-predictive, 1 goal)

- state: US state (by number) - not counted as predictive above, but if considered, should be considered nominal (nominal)
- county: numeric code for county - not predictive, and many missing values (numeric)
- community: numeric code for community - not predictive and many missing values (numeric)
- communityname: community name - not predictive - for information only (string)
- fold: fold number for non-random 10 fold cross validation, potentially useful for debugging, paired tests - not predictive (numeric)
- population: population for community: (numeric - decimal)
- householdsize: mean people per household (numeric - decimal)
- racePctBlack: percentage of population that is african american (numeric - decimal)
- racePctWhite: percentage of population that is caucasian (numeric - decimal)
- racePctAsian: percentage of population that is of asian heritage (numeric - decimal)
- racePctHispanic: percentage of population that is of hispanic heritage (numeric - decimal)
- agePct12t21: percentage of population that is 12-21 in age (numeric - decimal)
- agePct12t29: percentage of population that is 12-29 in age (numeric - decimal)
- agePct16t24: percentage of population that is 16-24 in age (numeric - decimal)
- agePct65up: percentage of population that is 65 and over in age (numeric - decimal)
- numbUrban: number of people living in areas classified as urban (numeric - decimal)

- pctUrban: percentage of people living in areas classified as urban (numeric - decimal)
- medIncome: median household income (numeric - decimal)
- pctWWage: percentage of households with wage or salary income in 1989 (numeric - decimal)
- pctWFarmSelf: percentage of households with farm or self employment income in 1989 (numeric - decimal)
- pctWInvInc: percentage of households with investment / rent income in 1989 (numeric - decimal)
- pctWSocSec: percentage of households with social security income in 1989 (numeric - decimal)
- pctWPubAsst: percentage of households with public assistance income in 1989 (numeric - decimal)
- pctWRetire: percentage of households with retirement income in 1989 (numeric - decimal)
- medFamInc: median family income (differs from household income for non-family households) (numeric - decimal)
- perCapInc: per capita income (numeric - decimal)
- whitePerCap: per capita income for caucasians (numeric - decimal)
- blackPerCap: per capita income for african americans (numeric - decimal)
- indianPerCap: per capita income for native americans (numeric - decimal)
- AsianPerCap: per capita income for people with asian heritage (numeric - decimal)
- OtherPerCap: per capita income for people with 'other' heritage (numeric - decimal)
- HispPerCap: per capita income for people with hispanic heritage (numeric - decimal)
- NumUnderPov: number of people under the poverty level (numeric - decimal)

- PctPopUnderPov: percentage of people under the poverty level (numeric - decimal)
- PctLess9thGrade: percentage of people 25 and over with less than a 9th grade education (numeric - decimal)
- PctNotHSGrad: percentage of people 25 and over that are not high school graduates (numeric - decimal)
- PctBSorMore: percentage of people 25 and over with a bachelors degree or higher education (numeric - decimal)
- PctUnemployed: percentage of people 16 and over, in the labor force, and unemployed (numeric - decimal)
- PctEmploy: percentage of people 16 and over who are employed (numeric - decimal)
- PctEmplManu: percentage of people 16 and over who are employed in manufacturing (numeric - decimal)
- PctEmplProfServ: percentage of people 16 and over who are employed in professional services (numeric - decimal)
- PctOccupManu: percentage of people 16 and over who are employed in manufacturing (numeric - decimal) #####
- PctOccupMgmtProf: percentage of people 16 and over who are employed in management or professional occupations (numeric - decimal)
- MalePctDivorce: percentage of males who are divorced (numeric - decimal)
- MalePctNevMarr: percentage of males who have never married (numeric - decimal)
- FemalePctDiv: percentage of females who are divorced (numeric - decimal)
- TotalPctDiv: percentage of population who are divorced (numeric - decimal)
- PersPerFam: mean number of people per family (numeric - decimal)
- PctFam2Par: percentage of families (with kids) that are headed by two parents (numeric - decimal)

- PctKids2Par: percentage of kids in family housing with two parents (numeric - decimal)
- PctYoungKids2Par: percent of kids 4 and under in two parent households (numeric - decimal)
- PctTeen2Par: percent of kids age 12-17 in two parent households (numeric - decimal)
- PctWorkMomYoungKids: percentage of moms of kids 6 and under in labor force (numeric - decimal)
- PctWorkMom: percentage of moms of kids under 18 in labor force (numeric - decimal)
- NumIlleg: number of kids born to never married (numeric - decimal)
- PctIlleg: percentage of kids born to never married (numeric - decimal)
- NumImmig: total number of people known to be foreign born (numeric - decimal)
- PctImmigRecent: percentage of \_immigrants\_ who immigrated within last 3 years (numeric - decimal)
- PctImmigRec5: percentage of \_immigrants\_ who immigrated within last 5 years (numeric - decimal)
- PctImmigRec8: percentage of \_immigrants\_ who immigrated within last 8 years (numeric - decimal)
- PctImmigRec10: percentage of \_immigrants\_ who immigrated within last 10 years (numeric - decimal)
- PctRecentImmig: percent of \_population\_ who have immigrated within the last 3 years (numeric - decimal)
- PctRecImmig5: percent of \_population\_ who have immigrated within the last 5 years (numeric - decimal)
- PctRecImmig8: percent of \_population\_ who have immigrated within the last 8 years (numeric - decimal)
- PctRecImmig10: percent of \_population\_ who have immigrated within the last 10 years (numeric - decimal)

- PctSpeakEnglOnly: percent of people who speak only English (numeric - decimal)
- PctNotSpeakEnglWell: percent of people who do not speak English well (numeric - decimal)
- PctLargHouseFam: percent of family households that are large (6 or more) (numeric - decimal)
- PctLargHouseOccup: percent of all occupied households that are large (6 or more people) (numeric - decimal)
- PersPerOccupHous: mean persons per household (numeric - decimal)
- PersPerOwnOccHous: mean persons per owner occupied household (numeric - decimal)
- PersPerRentOccHous: mean persons per rental household (numeric - decimal)
- PctPersOwnOccup: percent of people in owner occupied households (numeric - decimal)
- PctPersDenseHous: percent of persons in dense housing (more than 1 person per room) (numeric - decimal)
- PctHousLess3BR: percent of housing units with less than 3 bedrooms (numeric - decimal)
- MedNumBR: median number of bedrooms (numeric - decimal)
- HousVacant: number of vacant households (numeric - decimal)
- PctHousOccup: percent of housing occupied (numeric - decimal)
- PctHousOwnOcc: percent of households owner occupied (numeric - decimal)
- PctVacantBoarded: percent of vacant housing that is boarded up (numeric - decimal)
- PctVacMore6Mos: percent of vacant housing that has been vacant more than 6 months (numeric - decimal)



- MedYrHousBuilt: median year housing units built (numeric - decimal)
- PctHousNoPhone: percent of occupied housing units without phone (in 1990, this was rare!) (numeric - decimal)
- PctWOFullPlumb: percent of housing without complete plumbing facilities (numeric - decimal)
- OwnOccLowQuart: owner occupied housing - lower quartile value (numeric - decimal)
- OwnOccMedVal: owner occupied housing - median value (numeric - decimal)
- OwnOccHiQuart: owner occupied housing - upper quartile value (numeric - decimal)
- RentLowQ: rental housing - lower quartile rent (numeric - decimal)
- RentMedian: rental housing - median rent (Census variable H32B from file STF1A) (numeric - decimal)
- RentHighQ: rental housing - upper quartile rent (numeric - decimal)
- MedRent: median gross rent (Census variable H43A from file STF3A - includes utilities) (numeric - decimal)
- MedRentPctHousInc: median gross rent as a percentage of household income (numeric - decimal)
- MedOwnCostPctInc: median owners cost as a percentage of household income - for owners with a mortgage (numeric - decimal)
- MedOwnCostPctIncNoMtg: median owners cost as a percentage of household income - for owners without a mortgage (numeric - decimal)
- NumInShelters: number of people in homeless shelters (numeric - decimal)
- NumStreet: number of homeless people counted in the street (numeric - decimal)
- PctForeignBorn: percent of people foreign born (numeric - decimal)

- PctBornSameState: percent of people born in the same state as currently living (numeric - decimal)
- PctSameHouse85: percent of people living in the same house as in 1985 (5 years before) (numeric - decimal)
- PctSameCity85: percent of people living in the same city as in 1985 (5 years before) (numeric - decimal)
- PctSameState85: percent of people living in the same state as in 1985 (5 years before) (numeric - decimal)
- LemasSwornFT: number of sworn full time police officers (numeric - decimal)
- LemasSwFTPerPop: sworn full time police officers per 100K population (numeric - decimal)
- LemasSwFTFieldOps: number of sworn full time police officers in field operations (on the street as opposed to administrative etc) (numeric - decimal)
- LemasSwFTFieldPerPop: sworn full time police officers in field operations (on the street as opposed to administrative etc) per 100K population (numeric - decimal)
- LemasTotalReq: total requests for police (numeric - decimal)
- LemasTotReqPerPop: total requests for police per 100K population (numeric - decimal)
- PolicReqPerOffic: total requests for police per police officer (numeric - decimal)
- PolicPerPop: police officers per 100K population (numeric - decimal)
- RacialMatchCommPol: a measure of the racial match between the community and the police force. High values indicate proportions in community and police force are similar (numeric - decimal)
- PctPolicWhite: percent of police that are caucasian (numeric - decimal)

- PctPolicBlack: percent of police that are african american (numeric - decimal)
- PctPolicHisp: percent of police that are hispanic (numeric - decimal)
- PctPolicAsian: percent of police that are asian (numeric - decimal)
- PctPolicMinor: percent of police that are minority of any kind (numeric - decimal)
- OfficAssgnDrugUnits: number of officers assigned to special drug units (numeric - decimal)
- NumKindsDrugsSeiz: number of different kinds of drugs seized (numeric - decimal)
- PolicAveOTWorked: police average overtime worked (numeric - decimal)
- LandArea: land area in square miles (numeric - decimal)
- PopDens: population density in persons per square mile (numeric - decimal)
- PctUsePubTrans: percent of people using public transit for commuting (numeric - decimal)
- PolicCars: number of police cars (numeric - decimal)
- PolicOperBudg: police operating budget (numeric - decimal)
- LemasPctPolicOnPatr: percent of sworn full time police officers on patrol (numeric - decimal)
- LemasGangUnitDeploy: gang unit deployed (numeric - decimal - but really ordinal - 0 means NO, 1 means YES, 0.5 means Part Time)
- LemasPctOfficDrugUn: percent of officers assigned to drug units (numeric - decimal)
- PolicBudgPerPop: police operating budget per population (numeric - decimal)
- ViolentCrimesPerPop: total number of violent crimes per 100K popuation (numeric - decimal) GOAL attribute (to be predicted)

# **DISCUSSION TIME**

- **Review of last week**
- **Further Reading for Logistic Regression**
- **Check in with homework/course project**

## **WEEK 4 - Review**

---

# **DISCUSSION TIME**

- **Logistic Regression**
- **Test, Training and Validation data splits**
- **Cross-Validation**
- **Bias-Variance Trade-Off**

# **PRE-READING**

**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**