Clustering and Prediction

Probability and Statistics for Data Science

CSE594 - Spring 2016

But first,

One final useful statistical technique from Part II

Confidence Intervals

Motivation: p-values tell a nice succinct story but neglect a lot of information.

Estimating a point, approximated as normal (e.g. error or mean)

$$\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i. \qquad\qquad \mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \qquad\qquad \left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

find CI% based on standard normal distribution (i.e. CI% = 95, z = 1.96)

Resampling Techniques Revisited

The bootstrap

- What if we don't know the distribution?

Resampling Techniques Revisited

The bootstrap

- What if we don't know the distribution?
- *Resample* many potential distributions based on the observed data and find the range that CI% of the data fall in (e.g. mean).

Resample: for each *i* in *n* observations, put all observations in a hat and draw one (all observations are equally likely).

Clustering and Prediction

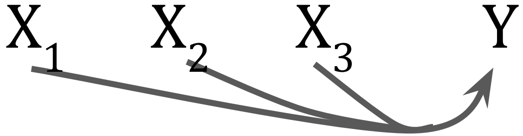
(now back to our regularly scheduled program)

I. Probability Theory

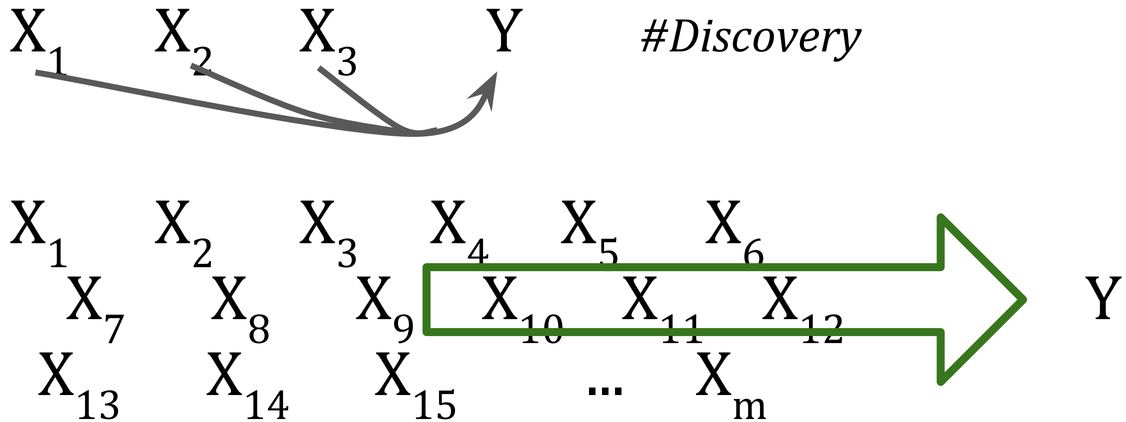
II. Discovery: Quantitative Research Methods

III. Clustering and Prediction

(now back to our regularly scheduled program)



Clustering and Prediction



Clustering and Prediction

$X_1 \quad X_2 \quad X_3 \quad \rightarrow \quad Y$

#Discovery

M < ~5 or m << n (much less)

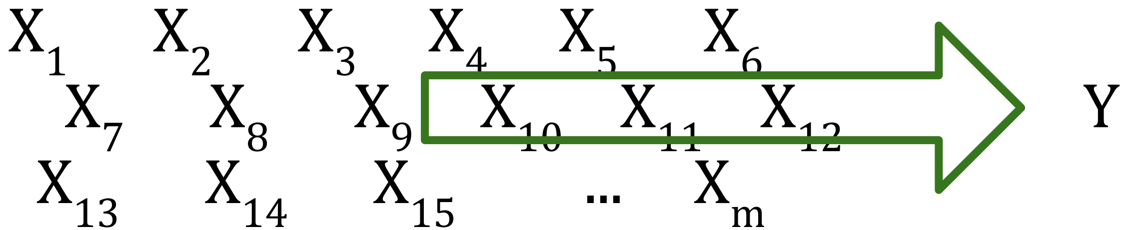
$X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6$

$X_7 \quad X_8 \quad X_9 \quad X_{10} \quad X_{11} \quad X_{12} \quad \Rightarrow \quad Y$

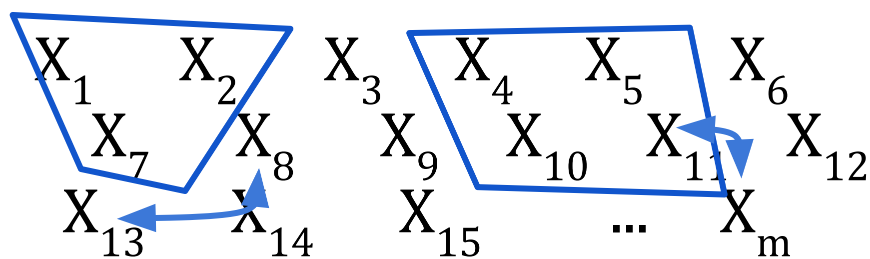
$X_{13} \quad X_{14} \quad X_{15} \quad \dots \quad X_m$

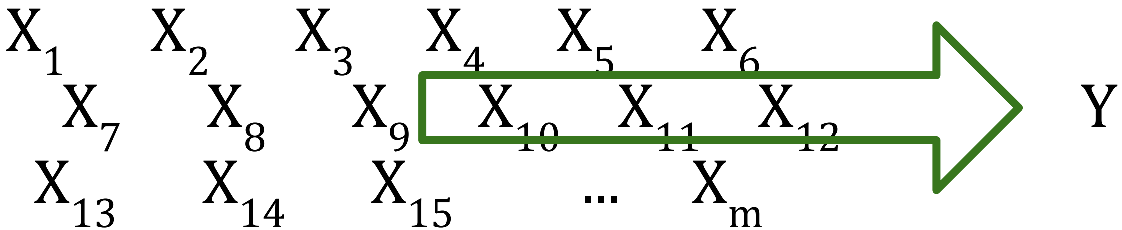
M > ~100 or m □ n or m >> n

Clustering and Prediction

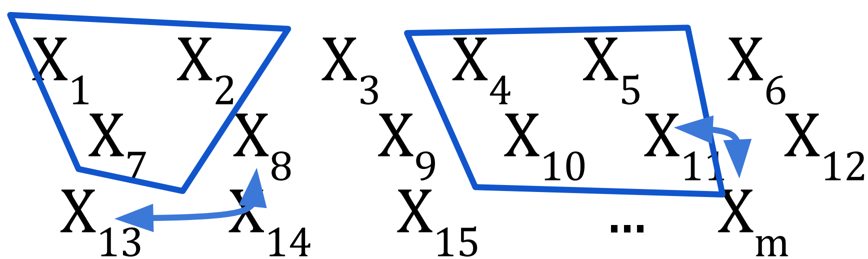


Clustering and Prediction

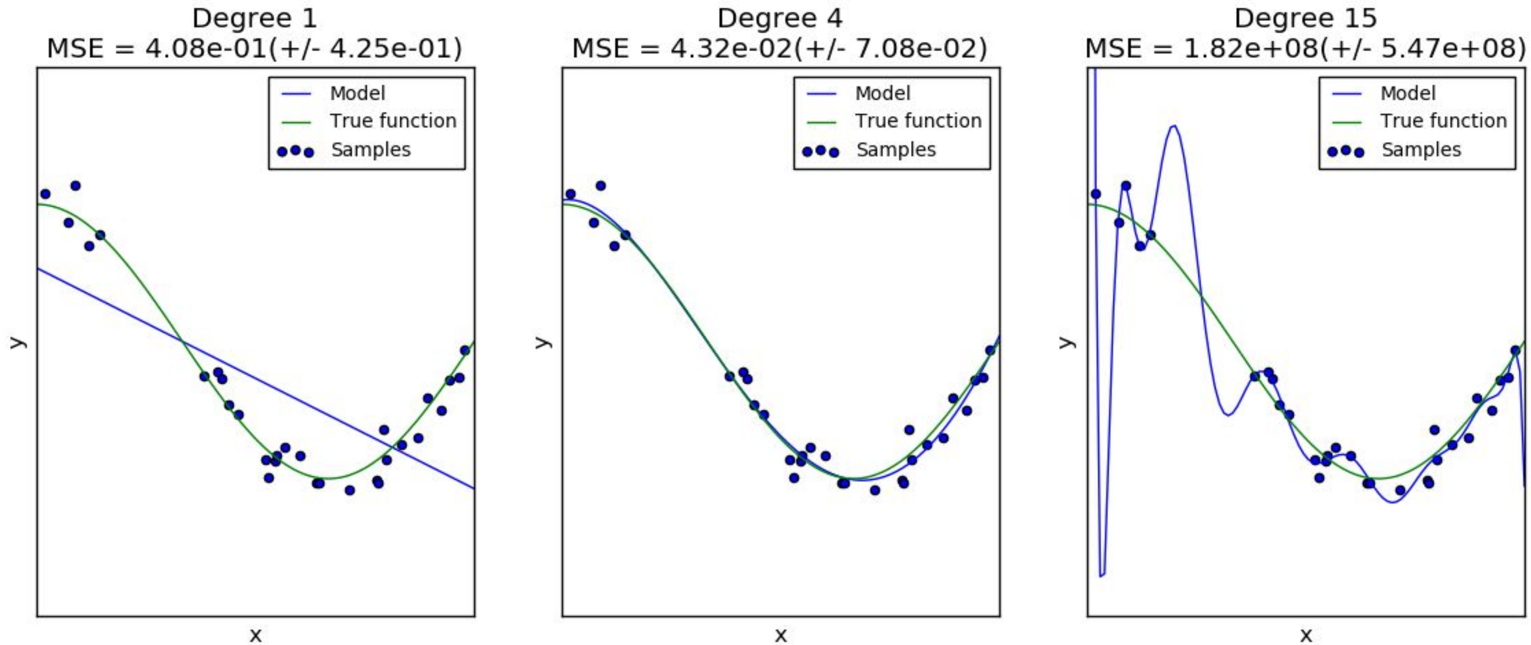




Clustering and Prediction



Overfitting (1-d example)



Underfit

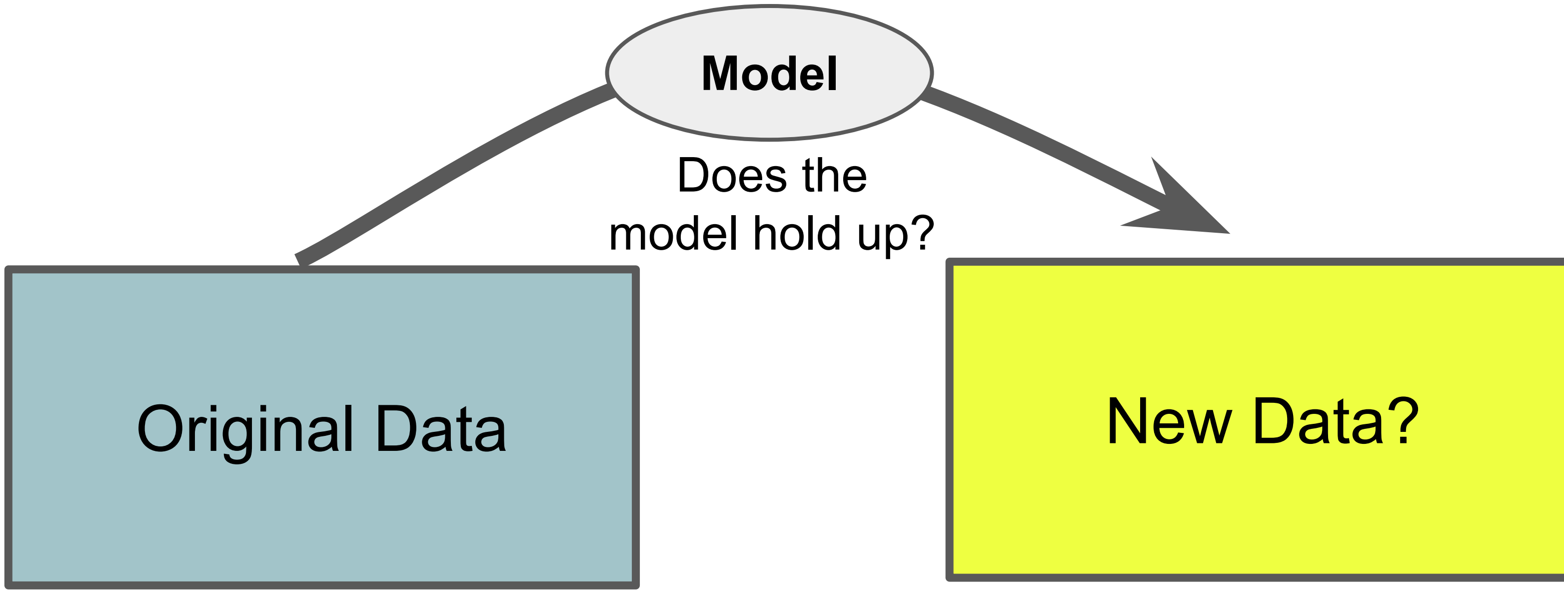
High Bias

Overfit

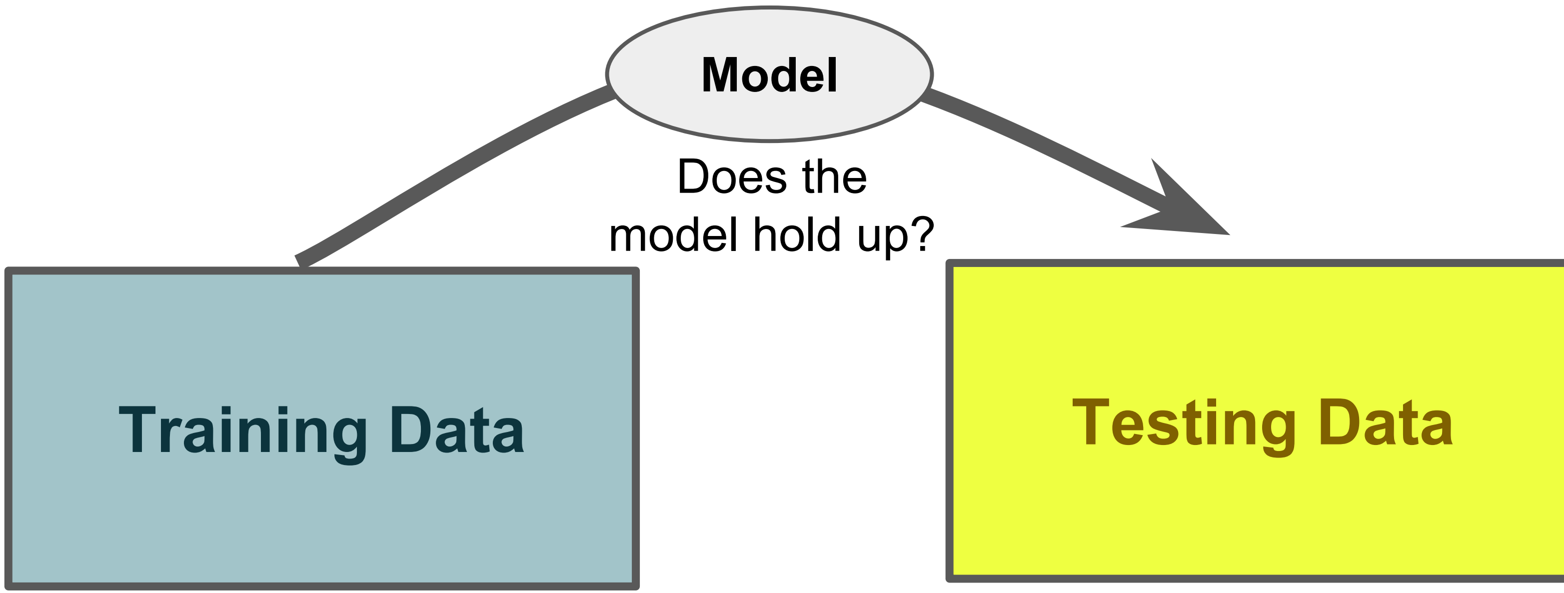
High Variance

(image credit: Scikit-learn; in practice data are rarely this clear)

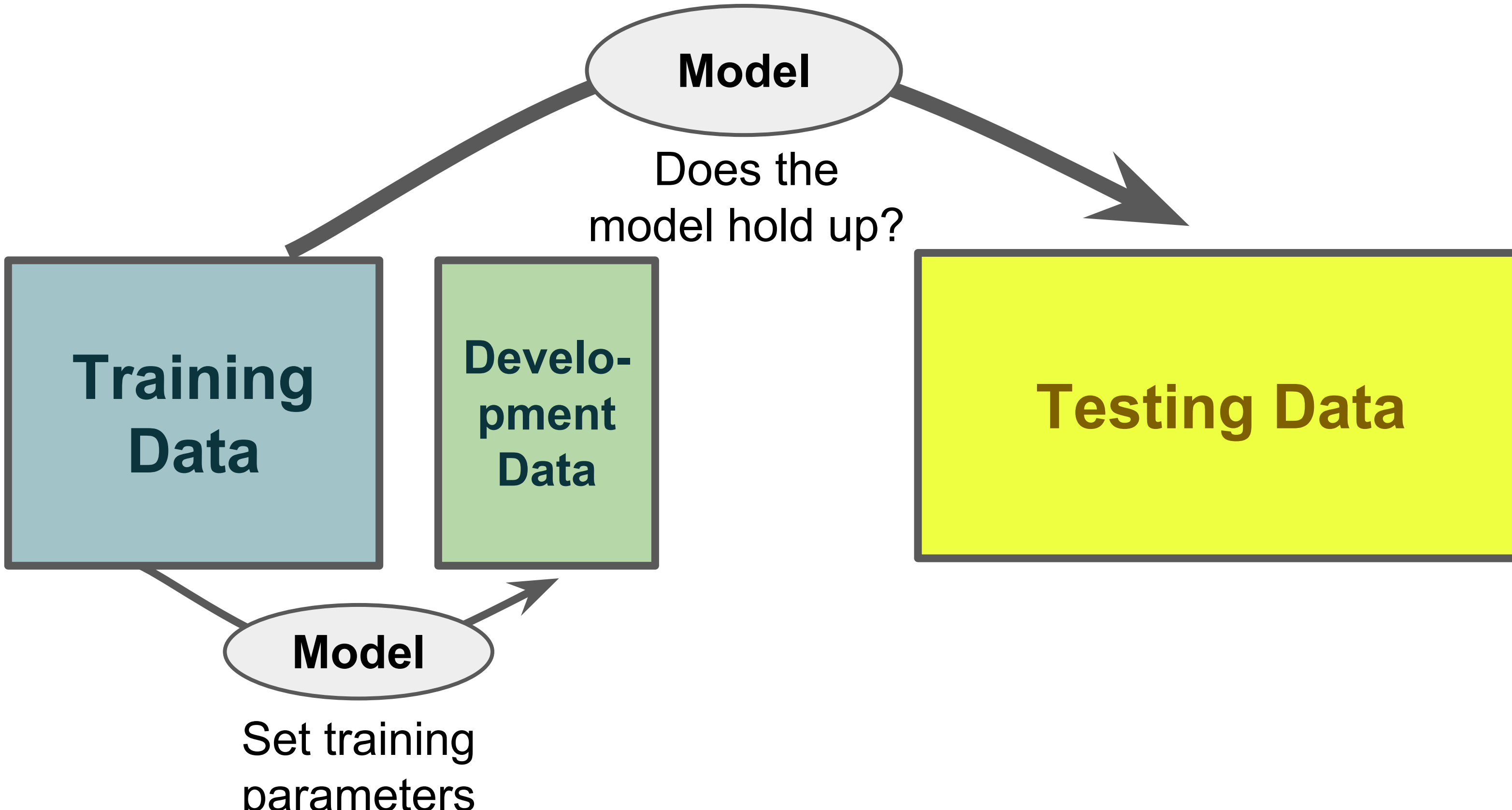
Common Goal: Generalize to new data



Common Goal: Generalize to new data



Common Goal: Generalize to new data



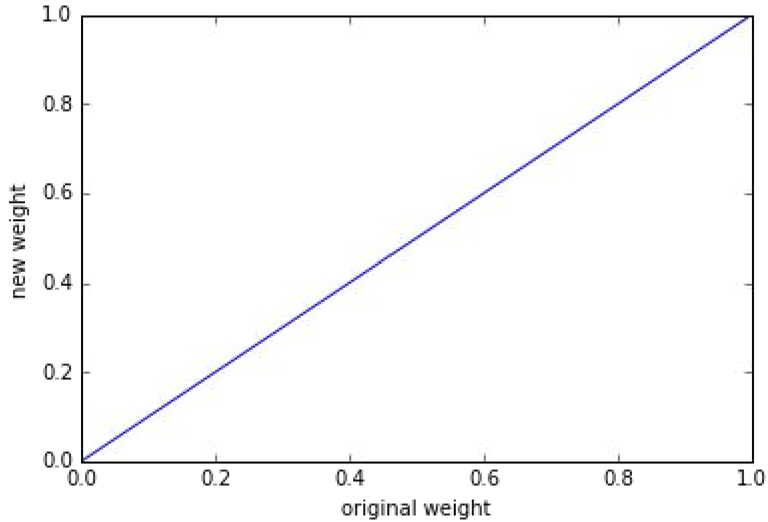
Feature Selection / Subset Selection

Forward Stepwise Selection:

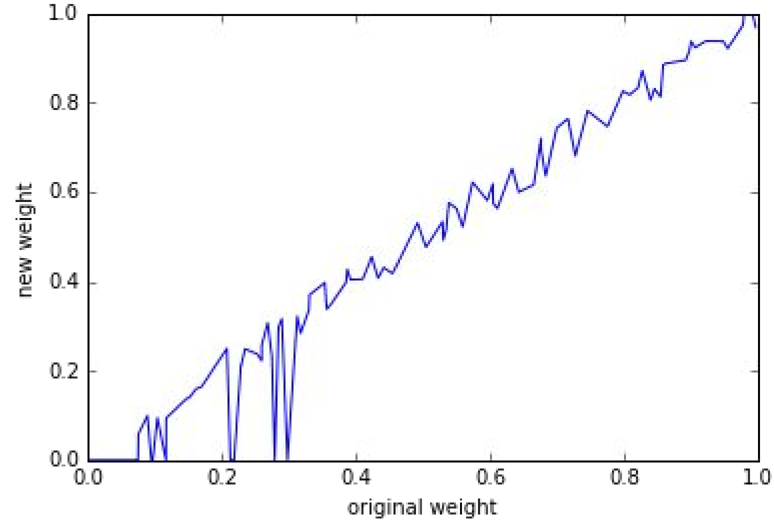
- start with current_model just has the intercept (mean)
  remaining_predictors = all_predictors
- for i in range(k)

```
#find best p to add to current_model:
for p in remaining_prepdictors
        refit current_model with p
#add best p, based on RSS_p to current_model
#remove p from remaining predictors
```

Regularization (Shrinkage)



No selection (weight=beta)



forward stepwise

Why just keep or discard features?

Regularization (L2, Ridge Regression)

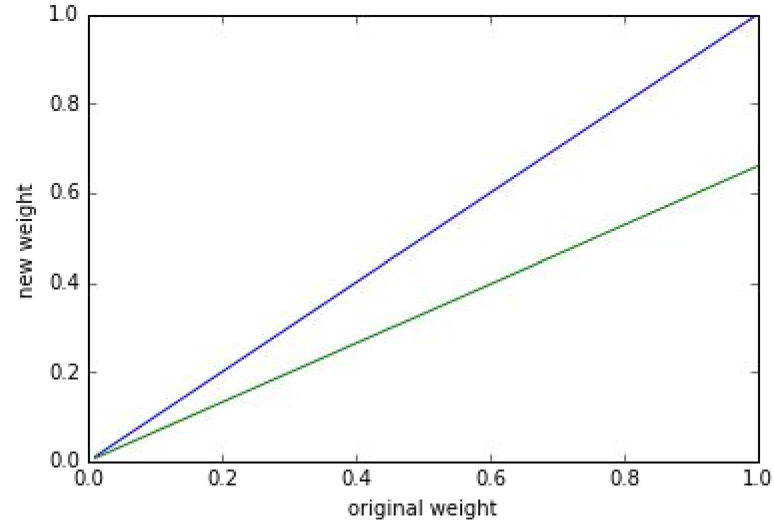
Idea: Impose a penalty on size of weights:

Ordinary least squares objective:

$$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2\}$$

Ridge regression:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2\}$$



Regularization (L2, Ridge Regression)

Idea: Impose a penalty on size of weights:

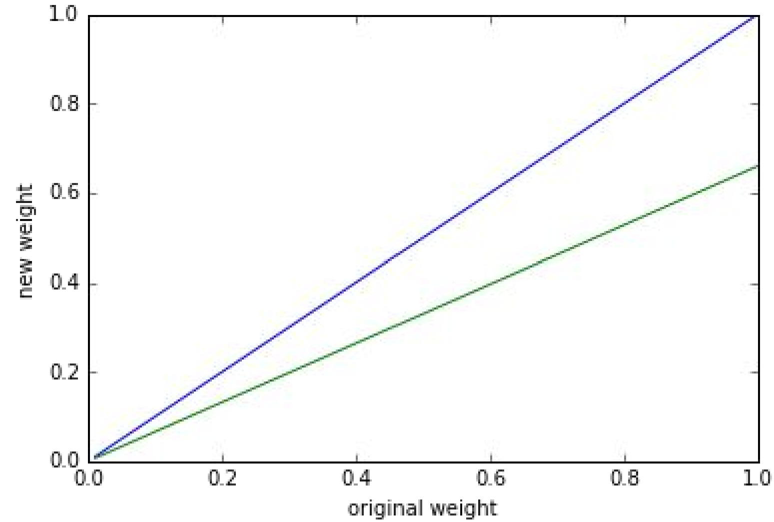
Ordinary least squares objective:

$$\hat{\beta} = argmin_\beta \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 \}$$

Ridge regression:

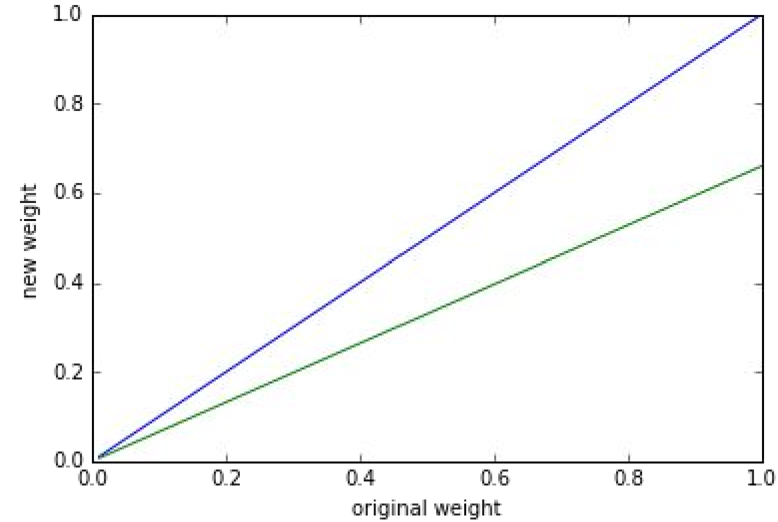
$$\hat{\beta}^{ridge} = argmin_\beta \{ \sum_{i=1}^{N} (y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \}$$

$$\lambda ||\beta||_2^2$$



Regularization (L2, Ridge Regression)

Idea: Impose a penalty on size of weights:



Ordinary least squares objective:

$$\hat{\beta} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2\}$$

Ridge regression:

$$\hat{\beta}^{ridge} = argmin_{\beta}\{\sum_{i=1}^{N}(y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{m}\beta_j^2\}$$

$$\lambda||\beta||_2^2$$

In Matrix Form:

$$\mathrm{RSS}(\lambda) = (y - X\beta)^T(y - X\beta) + \lambda\beta^T\beta$$

$$\hat{\beta}^{ridge} = (X^TX + \lambda I)^{-1}X^Ty$$

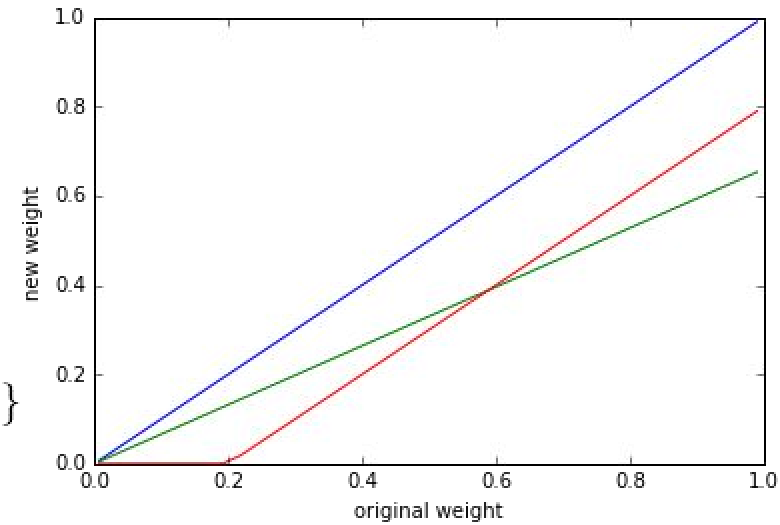
I: m x m identity matrix

Regularization (L1, The "Lasso")

Idea: Impose a penalty and zero-out
some weights

The Lasso Objective:

$$\hat{\beta}^{lasso} = argmin_{\beta}\{\frac{1}{2}\sum_{i=1}^{N}(Y_i - \sum_{j=1}^{m} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{m} |\beta_j|\}$$



$$\lambda||\beta||_1$$

No closed form matrix solution, but
often solved with coordinate descent.

Application: m ≅ n or m >> n

Regularization (L1L2, “Elastic Net”)

Regularized Logistic Regression

NFold Cross-Validation

Goal: Decent estimate of model accuracy

Common Goal: Generalize to new data