GENERAL ASSEMBLY

INTRO TO DATA SCIENCE

REGULARIZATION, DIMENSIONALITY REDUCTION

Recap

cross validation, feature subset selection

Regularization

ridge, lasso

Dimensionality reduction

PCA

Assessing Model Accuracy

What is a model? What is a "good" model?

"Essentially, all models are wrong, but some are useful"

-George Box

SOME IMPORTANT QUESTIONS

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the test data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

SOME IMPORTANT QUESTIONS

1. Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?
3. How well does the model fit the test data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?
5. Are there ways to *combine* or *transform* predictors to summarize their contribution to the response variable?

THREE METHOD CLASSES

1. Subset selection
2. Regularization
3. Dimension Reduction

THREE METHOD CLASSES

1. Subset selection
   - identify a subset of predictors that contribute to the response
2. Regularization
   - fit a model with all predictors, but *shrink* coefficients (***β***) towards zero relative to the least squares estimates
3. Dimensionality Reduction
   - compute ***linear combinations***, or ***projections***, of the variables and use these projections as predictors

REGULARIZATION

REGULARIZATION: RATIONALE

- ***Multicollinearity*** in predictors leads to high variance of estimator
- Number of observations should be >> number of predictors
- Prediction error increases as a function of number of predictors

REGULARIZATION: RATIONALE

- Avoid ***overfitting*** by not generating high coefficients for predictors that are sparse
- Stabilize the estimates, especially when there's ***collinearity*** in the data
- Deliberately introduce ***bias*** into the estimation of ***β*** in order to reduce the ***variance*** of the estimate.

Resulting estimators generally have lower mean squared error than the OLS estimates, particularly when ***multicollinearity*** is present.

REGULARIZATION

Regularization works by adding the penalty associated with the coefficient values to the error of the hypothesis. This way, an accurate hypothesis with unlikely coefficients would be penalized, while a somewhat less accurate but more conservative hypothesis with low coefficients would not be penalized as much.

It may not be immediately obvious why such a constraint should improve the fit, but it turns out that shrinking the coefficient estimates can significantly reduce their variance...

REGULARIZATION

- Instead of minimizing RSS, minimize
  RSS + (* penalty on parameters)

- $\sum |\beta_i|$
  Lasso, L1-norm
- $\sum \beta_i^2$
  Ridge, L2-norm

RIDGE REGRESSION (L2 NORM)

- Recall that the least squares fitting procedure estimates $\boldsymbol{\beta}$ using the values that minimize:

$$min(||y - \beta x||^2)$$

- In contrast, the ridge regression coefficient estimates $\boldsymbol{\beta}$ are the values that minimize:

$$min(||y - \beta x||^2 + \lambda ||\boldsymbol{\beta}||^2)$$

where $\lambda \geq 0$ is a tuning parameter, to be determined separately

RIDGE REGRESSION (L2)

- Alleviate ***multicollinearity*** among regression predictors
- Penalize large values of the parameters in the quantity we're seeking to minimize
- The term, $\lambda ||\beta||^2$, called a shrinkage penalty, is small when β_1, ..., β_p are close to zero, and so it has the effect of shrinking the estimates of β towards zero.
- The tuning parameter λ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Selecting a good value for λ is critical; cross-validation is used for this.

LASSO REGRESSION (L1)

- Ridge regression does have one obvious disadvantage: unlike subset selection, which will generally select models that involve just a subset of the variables, ridge regression will include all predictors in the final model
- The Lasso is a relatively recent alternative to ridge regression that overcomes this disadvantage. The lasso coefficients minimize the quantity:

$$min(||y - \beta x||^2 + \square|\boldsymbol{\beta}|)$$

- LASSO: **L**east **A**bsolute **S**hrinkage and **S**election **O**perator)

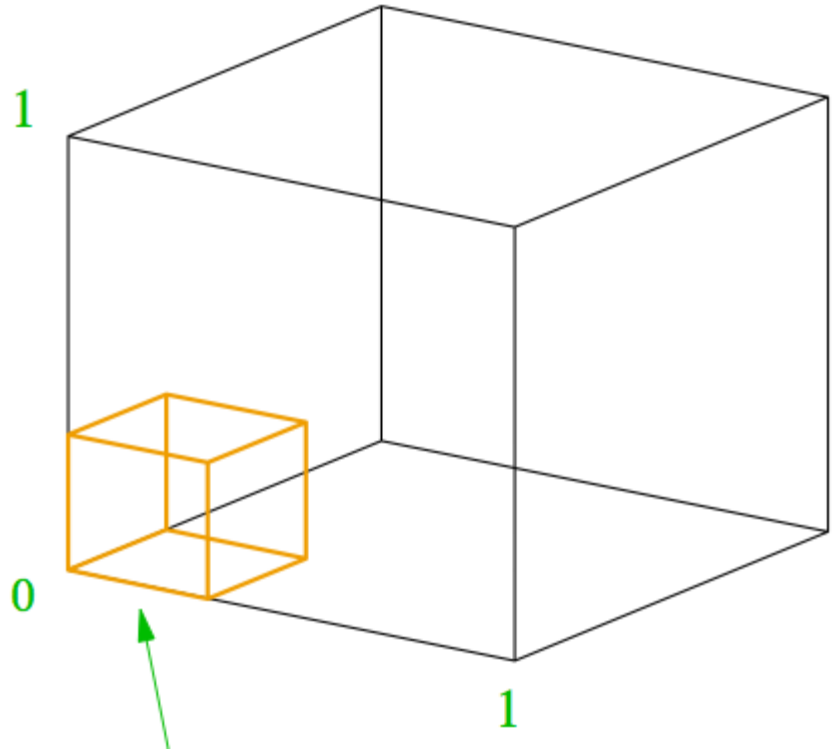
LASSO REGRESSION (L1)

- As with ridge regression, the lasso shrinks the coefficient estimates towards zero
- However, in the case of the lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter **λ** is sufficiently large
- Hence, much like best subset selection, the lasso performs variable selection
- We say that the lasso yields sparse models — that is, models that involve only a subset of the variables
- As in ridge regression, selecting a good value of λ for the lasso is critical; cross-validation is again the method of choice

DIMENSIONALITY REDUCTION

CURSE OF DIMENSIONALITY

CLUSTERING METHODS
BREAK DOWN
OUR FEATURE SPACE
BECOMES VASTLY
LARGER THAN OUR
AVAILABLE SAMPLE
COMPUTATIONALLY
EXPENSIVE



DIMENSIONALITY REDUCTION: COVARIANCE

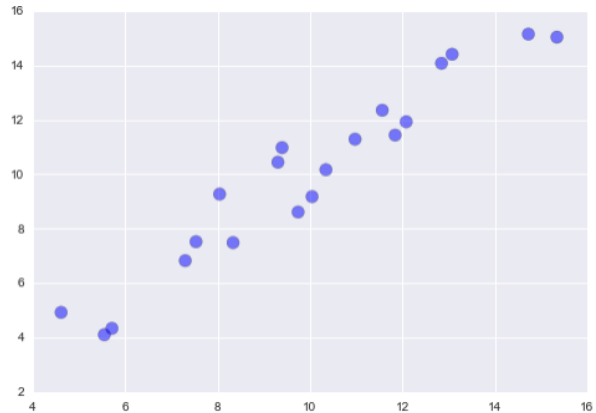
VARIANCE IS ONE DIMENSIONAL

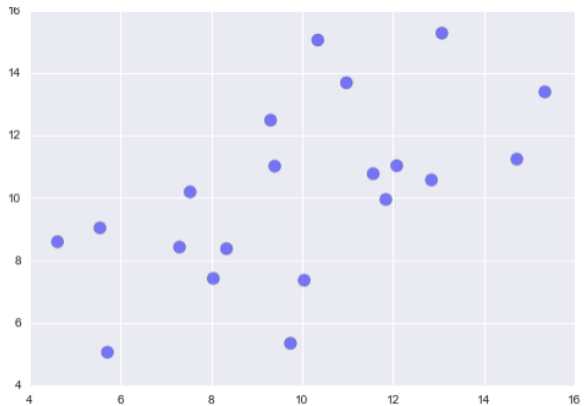
$$var(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

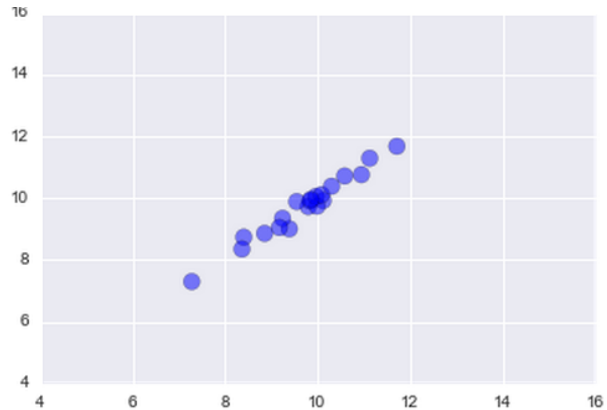
COVARIANCE IS ALWAYS MEASURED BETWEEN TWO DIMENSIONS

$$cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

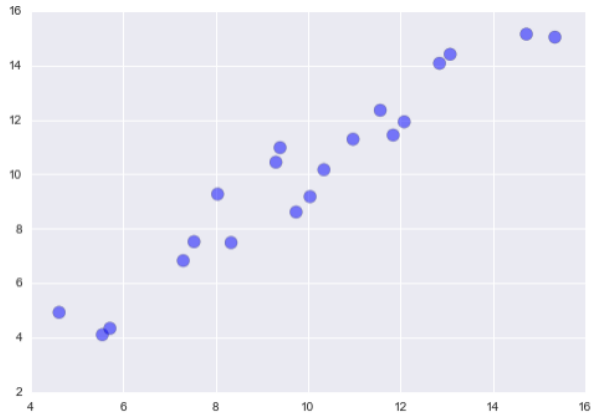
COVARIANCE

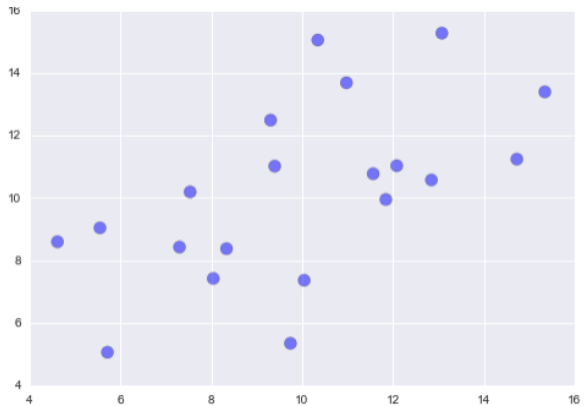

16
14
12
10
8
6
4
2
4
6
8
10
12
14
16

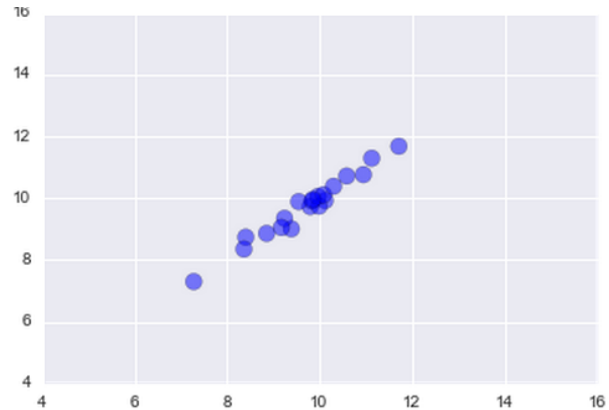

16
14
12
10
8
6
4
4
6
8
10
12
14
16


16
14
12
10
8
6
4
4
6
8
10
12
14
16

COVARIANCE

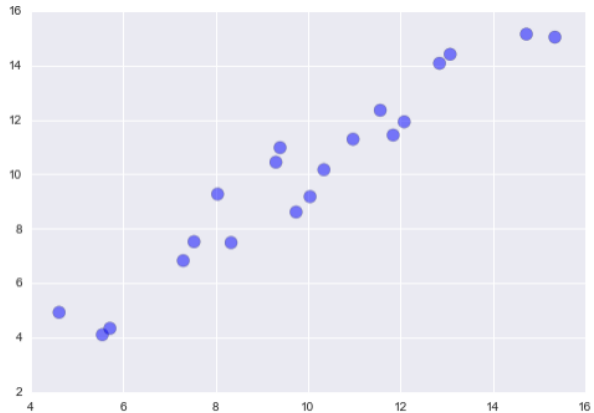




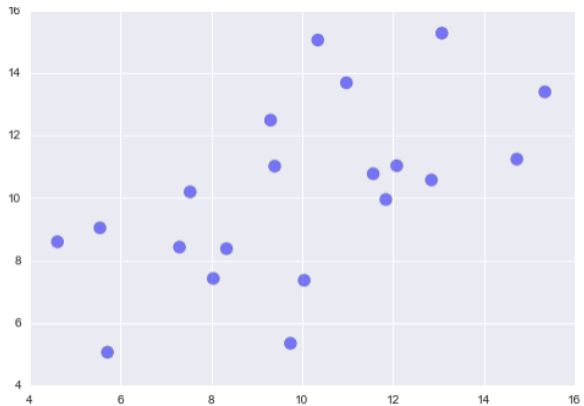


cov(x, y) = 9.8

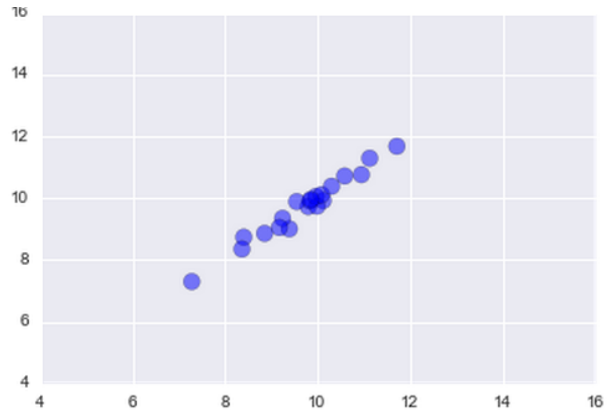
COVARIANCE



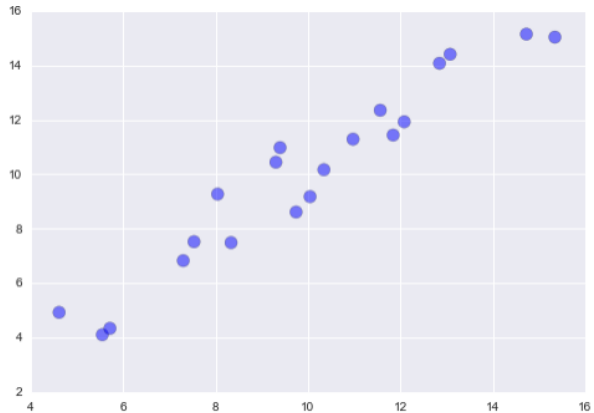
cov(x, y) = 9.8



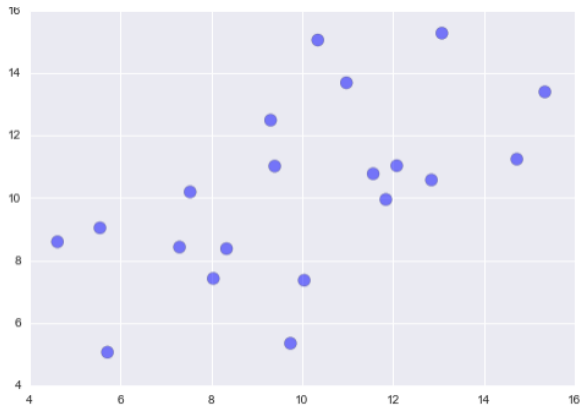
cov(x, y) = 4.9



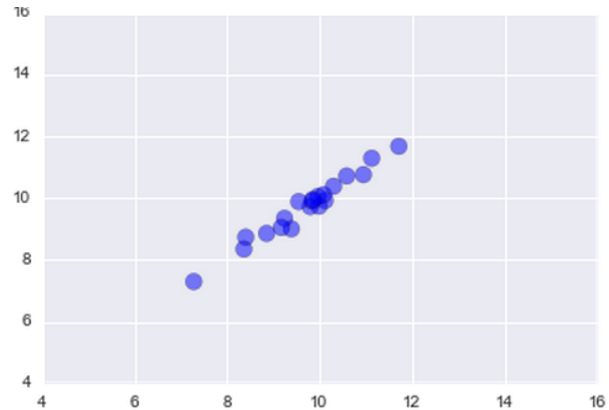
COVARIANCE



cov(x, y) = 9.8

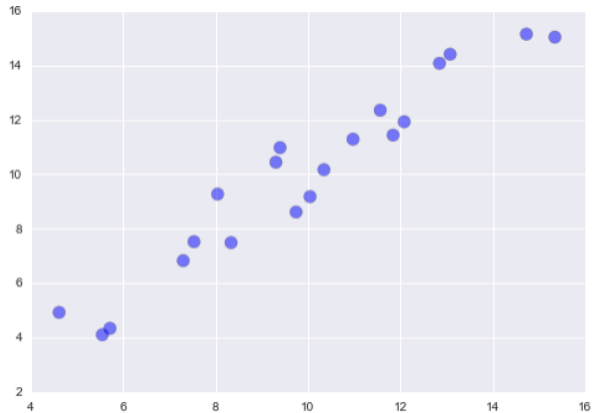


cov(x, y) = 4.9

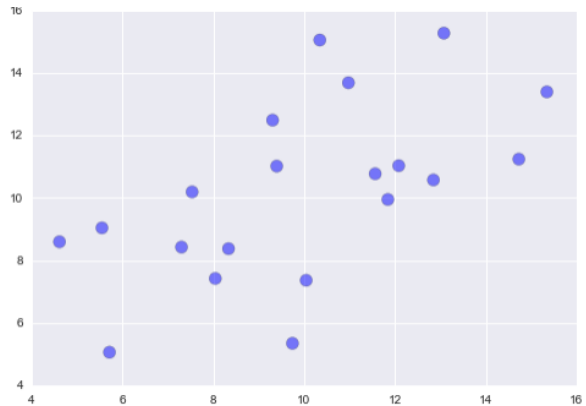


cov(x, y) = 1.0

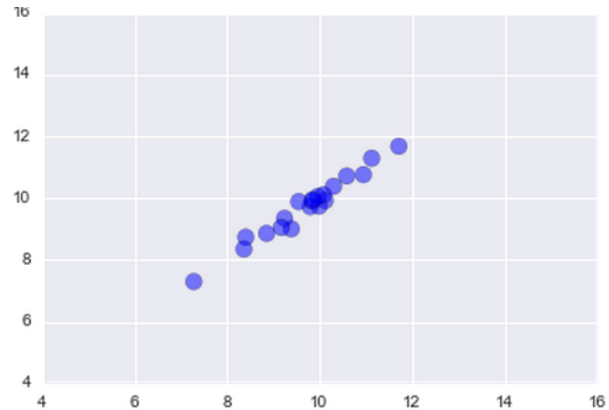
COVARIANCE



cov(x, y) = 9.8



cov(x, y) = 4.9



cov(x, y) = 1.0

COVARIANCE MEASURES THE DEGREE TO WHICH TWO VARIABLES ARE LINEARLY ASSOCIATED

DIMENSIONALITY REDUCTION: PCA

A technique useful for the compression and classification of data. The purpose is to reduce the dimensionality of a dataset by finding a new set of variables, smaller than the original set, that still retains most of the information of the original set.

The ***principal component*** is the one that maximizes the variance of the projected data.

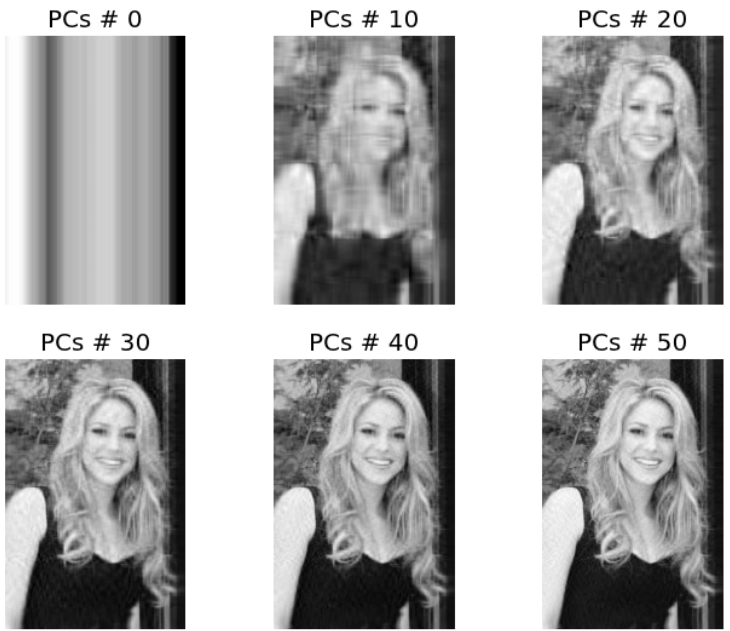
DIMENSIONALITY REDUCTION: PCA

A TECHNIQUE USEFUL FOR THE COMPRESSION AND CLASSIFICATION OF DATA. THE PURPOSE IS TO REDUCE THE DIMENSIONALITY OF A DATA SET BY FINDING A NEW SET OF VARIABLES, SMALLER THAN THE ORIGINAL VARIABLES, THAT NONETHELESS RETAINS MOST OF THE SAMPLE'S INFORMATION.

INFORMATION = VARIANCE

ADVANTAGES OF PCA

- **ALLOWS US TO COMPRESS DATA WHILE LOSING AS LITTLE OF THE VARIANCE AS POSSIBLE**



DISADVANTAGES OF PCA

- Covariance of 0 does not mean your data is independent

- Better methods for nonlinear data
- Not the best method for parsing hidden or latent factors

OTHER DIM REDUCTION ALGORITHMS

- Singular Value Decomposition (SVD)
- Factor Analysis
- Topological Data Analysis
- Discriminant Analysis
- Nonlinear dimensionality reduction
  - Principal Curves/Manifolds