STAT 215A Fall 2020
Week 9

James Duncan, OH: M, Th 2-4pm

Announcements

- Lab 3 due this **Thursday 10/22 at 11:59pm**
- Midterm: 10/29
  - Bring your questions to class on 10/27
- Class **Friday** will be from **1:00-2:30pm**
  - Please come if you can, but it will be recorded as usual
- Happy World Statistics Day!

Lab 3: Stability of K-means + Computability

- How's it going?
- Ben-Hur, et al. notes that similarity can be computed in $O(k_1 k_2 n)$
  - This should be your goal
- You can do better than the Figure 3 in Ben-Hur
- `foreach`
  - You can use the `doRNG` package to make your results reproducible
  - If you're having issues with `foreach`, try `future` or `parallel`
- Remember, no need to push a blinded version
- Make sure your `lab3` folder is well-organized and **only** contains the files I would need to reproduce your results!

Outline for today

- Regularization Pt. 1: Ridge
- Practice midterm solutions

Regularization Part I: Ridge

Thanks to Tiffany Tang for sharing her slides

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

- x_0 : observed “predictors” for a sample (assume fixed, non-random)
-

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

- x_0 : observed "predictors" for a sample (assume fixed, non-random)
- $y_0 = f(x_0) + \varepsilon$: the "response" for x_0
  - $f(x_0)$: the "true" value of f at x_0
  - $\varepsilon \sim (0, \sigma^2)$: additive random error
-

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

- x_0 : observed "predictors" for a sample (assume fixed, non-random)
- $y_0 = f(x_0) + \varepsilon$: the "response" for x_0
  - $f(x_0)$: the "true" value of f at x_0
  - $\varepsilon \sim (0, \sigma^2)$: additive random error
- $\hat{y}_0 = \hat{f}(x_0)$: the predicted response for x_0

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

- x_0 : observed "predictors" for a sample (assume fixed, non-random)
- $y_0 = f(x_0) + \varepsilon$: the "response" for x_0
  - $f(x_0)$: the "true" value of f at x_0
  - $\varepsilon \sim (0, \sigma^2)$: additive random error
- $\hat{y}_0 = \hat{f}(x_0)$: the predicted response for x_0

Expected prediction error (EPE):

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \underbrace{\mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}^2(\hat{f}(x_0))}_{\textbf{Mean squared error}}$$

Expected prediction error

Goal: Evaluate the *error* in estimating an unknown function f by the estimator $\hat{f}$

- x_0 : observed "predictors" for a sample (assume fixed, non-random)
- $y_0 = f(x_0) + \varepsilon$: the "response" for x_0
    - $f(x_0)$: the "true" value of f at x_0
    - $\varepsilon \sim (0, \sigma^2)$: additive random error
- $\hat{y}_0 = \hat{f}(x_0)$: the predicted response for x_0

Expected prediction error (EPE):

$$\mathbb{E}[y_0 - \hat{f}(x_0)]^2 = \underbrace{\sigma^2}_{\text{"irreducible" error}} + \underbrace{\mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}^2(\hat{f}(x_0))}_{\textbf{Mean squared error}}$$

The Bias-Variance Tradeoff

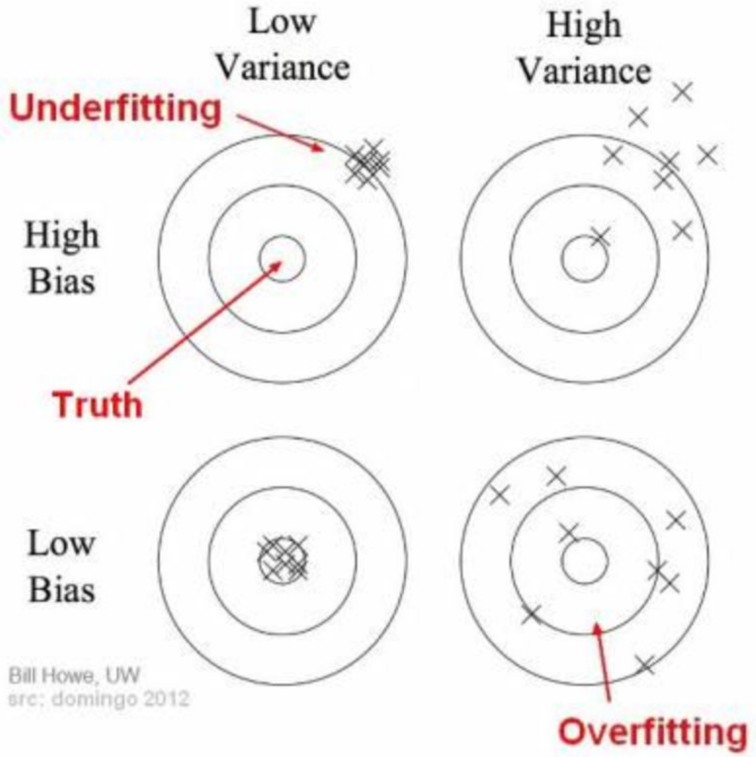
$$\mathrm{MSE} = \mathrm{Var}(\hat{f}(x_0)) + \mathrm{Bias}^2(\hat{f}(x_0))$$

- **Bias:** On average, how wrong is your prediction?

$$\mathrm{Bias}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)) - f(x)$$

- **Variance:** If you obtain a new, but similar dataset, how much does this change your predictions?

$$\mathrm{Var}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)^2) - \mathbb{E}(\hat{f}(x))^2$$



The Bias-Variance Tradeoff

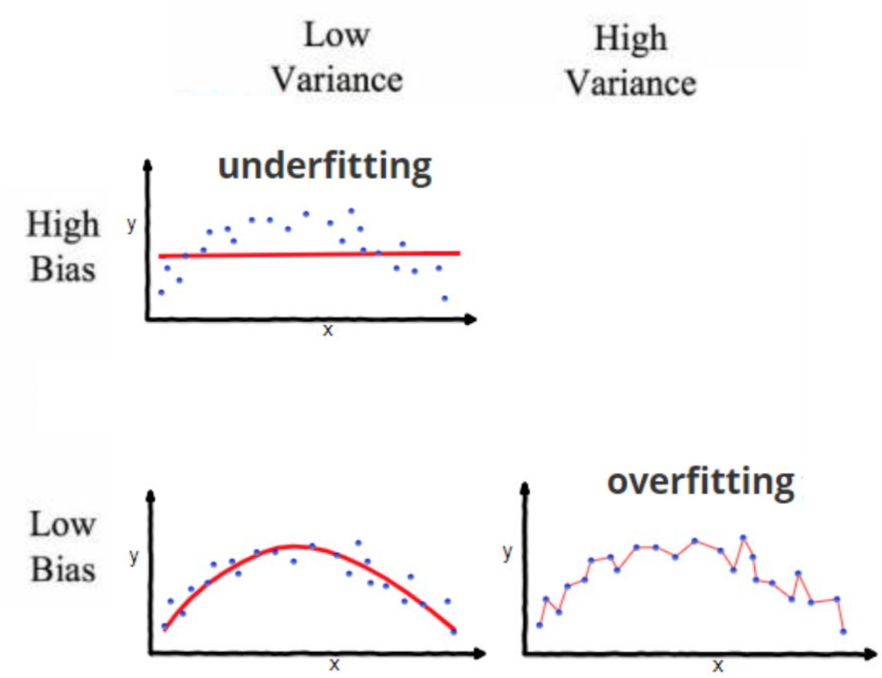
$$\text{MSE} = \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0))$$

- **Bias:** On average, how wrong is your prediction?

$$\text{Bias}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)) - f(x)$$

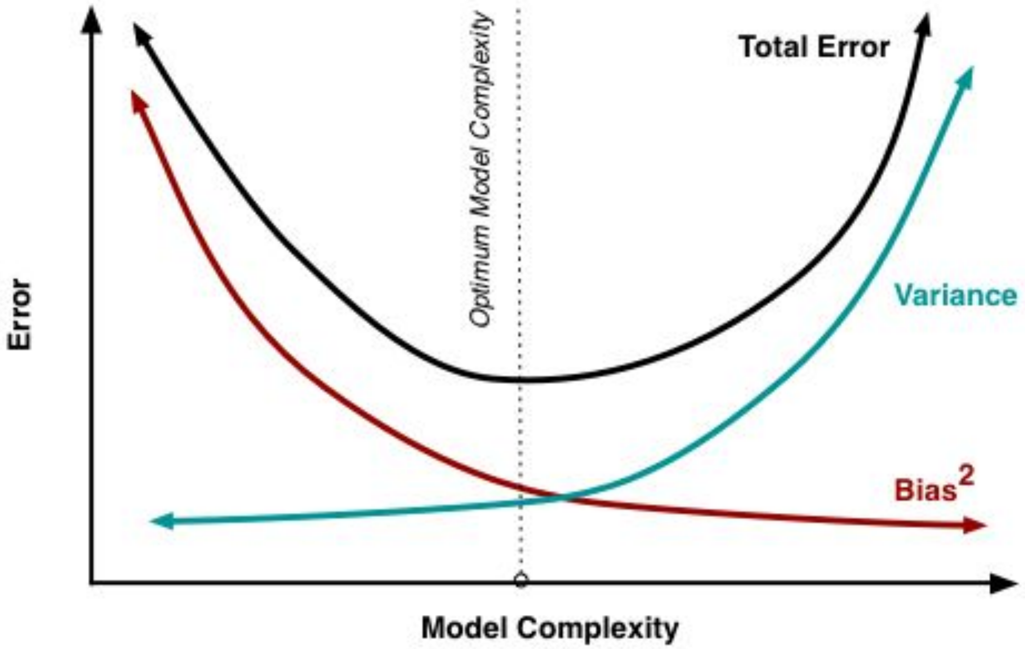
- **Variance:** If you obtain a new, but similar dataset, how much does this change your predictions?

$$\text{Var}(\hat{f}(x)) = \mathbb{E}(\hat{f}(x)^2) - \mathbb{E}(\hat{f}(x))^2$$



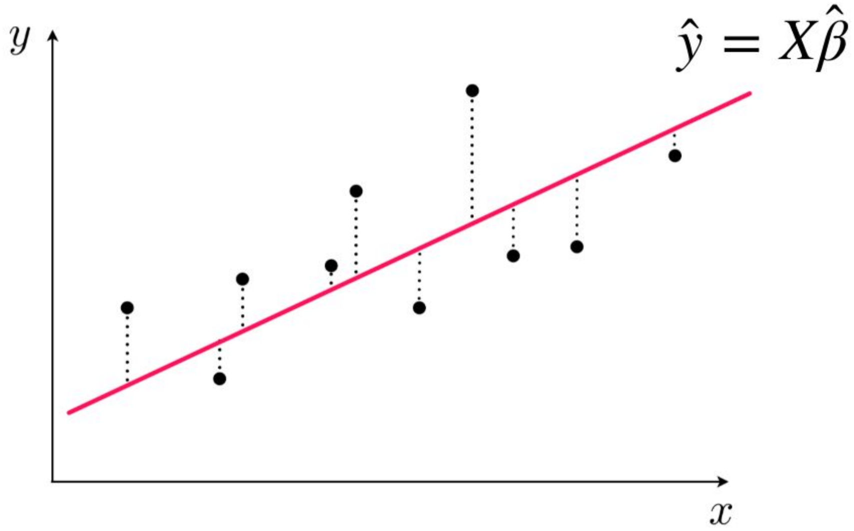
The Bias-Variance Tradeoff

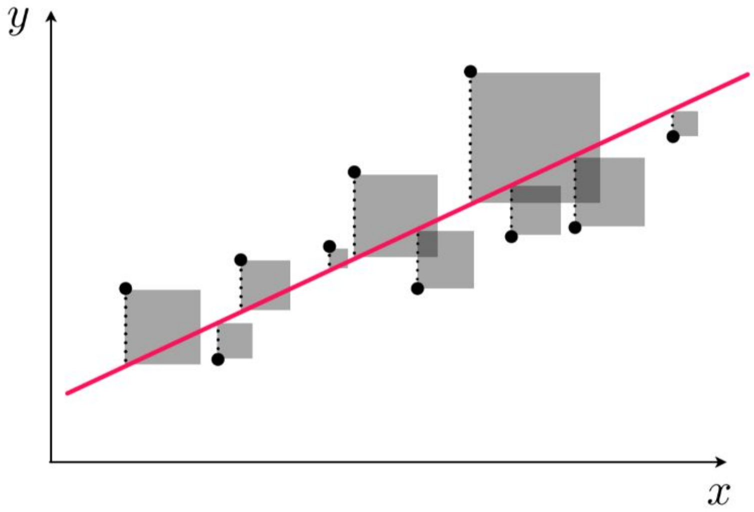
$$\text{MSE} = \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0))$$



Recall Ordinary Least Squares

$$\hat{\boldsymbol{\beta}}_{OLS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; \|\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 \qquad \hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$





OLS

Advantages

- Simple
- Closed-form solution
- Interpretable (?)
- Under some modeling assumptions, OLS has some desirable properties

OLS + Model

Assume $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$.

- Then $\hat{\beta}_{\text{OLS}}$ is an unbiased estimator for β

OLS + Model

Assume $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$.

- Then $\hat{\beta}_{\text{OLS}}$ is an unbiased estimator for β
- If we also assume that $\text{Var}(\varepsilon) = \sigma^2 \text{I}$ then OLS is BLUE
    - Best linear unbiased estimator, i.e. out of all linear unbiased estimators $\hat{\beta}_{\text{OLS}}$ has minimum variance (Gauss-Markov)

OLS + Model

Assume $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$.

- Then $\hat{\beta}_{\text{OLS}}$ is an unbiased estimator for β
- If we also assume that $\text{Var}(\varepsilon) = \sigma^2 \text{I}$ then OLS is BLUE
  - Best linear unbiased estimator, i.e. out of all linear unbiased estimators $\hat{\beta}_{\text{OLS}}$ has minimum variance (Gauss-Markov)
- Say we have a new point x_0 and prediction $\hat{y}_0 = x_0^\top \hat{\beta}$.

OLS + Model

Assume $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$.

- Then $\hat{\beta}_{\text{OLS}}$ is an unbiased estimator for β
- If we also assume that $\text{Var}(\varepsilon) = \sigma^2 \text{I}$ then OLS is BLUE
  - Best linear unbiased estimator, i.e. out of all linear unbiased estimators $\hat{\beta}_{\text{OLS}}$ has minimum variance (Gauss-Markov)
- Say we have a new point x_0 and prediction $\hat{y}_0 = x_0^\top \hat{\beta}$.
  - $\text{Bias}(\hat{y}_0) = 0$

OLS + Model

Assume $y = X\beta + \varepsilon$ with $\mathbb{E}(\varepsilon) = 0$.

- Then $\hat{\beta}_{\text{OLS}}$ is an unbiased estimator for β
- If we also assume that $\text{Var}(\varepsilon) = \sigma^2 \text{I}$ then OLS is BLUE
  - Best linear unbiased estimator, i.e. out of all linear unbiased estimators $\hat{\beta}_{\text{OLS}}$ has minimum variance (Gauss-Markov)
- Say we have a new point x_0 and prediction $\hat{y}_0 = x_0^\top \hat{\beta}$.
  - $\text{Bias}(\hat{y}_0) = 0$
  - For large n, and assuming $\mathbb{E}(X) = 0$:

$$\mathbb{E}_{x_0} \mathbb{E}(y_0 - \hat{y}_0)^2 \approx \sigma^2 + \sigma^2 (p/n)$$

See *The Elements of Statistical Learning*, Section 2.5 for details

OLS

Advantages

- Simple
- Closed-form solution
- Interpretable (?)
- Under some modeling assumptions, OLS has some desirable properties

Disadvantages

- Too simple
- No bias (assuming correctly specified), all variance... can lead to overfitting
- When $p > n$:
    - $X^\top X$ singular

A possible solution

$$\text{MSE} = \text{Var}(\hat{f}(x_0)) + \text{Bias}^2(\hat{f}(x_0))$$

- We could try to sacrifice some of the bias to reduce the variance
- One way to reduce the variance of your predictions is to restrict the parameter space in the optimization $\underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$
- In the linear setting, this motivates regularized linear regression methods such as ridge, Lasso, and elastic net

Ridge regression

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}_{\text{Loss function}} \quad \text{subject to} \quad \overbrace{||\boldsymbol{\beta}||_2^2 \leq \tau}^{\text{constraint}}$$

or equivalently

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \overbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}^{\text{Loss function}} + \overbrace{\lambda||\boldsymbol{\beta}||_2^2}^{\text{penalty}}$$

Ridge regression

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}_{\text{Loss function}} \quad \text{subject to} \quad \overbrace{||\boldsymbol{\beta}||_2^2 \leq \tau}^{\text{constraint}}$$

or equivalently

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \overbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}^{\text{Loss function}} + \overbrace{\lambda||\boldsymbol{\beta}||_2^2}^{\text{penalty}}$$

$\lambda \geq 0$ is a tuning or penalty parameter and regulates how much *shrinkage* we introduce:

- $\lambda \to 0 \implies \hat{\beta}^r \to \text{OLS}$

Ridge regression

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}_{\text{Loss function}} \quad \text{subject to} \quad \overbrace{||\boldsymbol{\beta}||_2^2 \leq \tau}^{\text{constraint}}$$

or equivalently

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \overbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}^{\text{Loss function}} + \overbrace{\lambda||\boldsymbol{\beta}||_2^2}^{\text{penalty}}$$

$\lambda \geq 0$ is a tuning or penalty parameter and regulates how much *shrinkage* we introduce:

- $\lambda \to 0 \implies \hat{\beta}^r \to \text{OLS}$
- $\lambda \to \infty \implies \hat{\beta}^r \to 0$

Ridge regression

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \underbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}_{\text{Loss function}} \quad \text{subject to} \quad \overbrace{||\boldsymbol{\beta}||_2^2 \le \tau}^{\text{constraint}}$$

or equivalently

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \overbrace{||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2}^{\text{Loss function}} + \overbrace{\lambda||\boldsymbol{\beta}||_2^2}^{\text{penalty}}$$

$\lambda \geq 0$ is a tuning or penalty parameter and regulates how much *shrinkage* we introduce:

- $\lambda \to 0 \implies \hat{\beta}^r \to \text{OLS}$
- $\lambda \to \infty \implies \hat{\beta}^r \to 0$
- Usually try to find an intermediate value that provides some shrinkage
- Can choose λ via CV

Why ridge?

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; ||\,\mathbf{y} - \mathbf{X}\,\boldsymbol{\beta}\,||_2^2 + \lambda||\,\boldsymbol{\beta}\,||_2^2$$

1. The solution exists and is unique even when $p > n$ (unlike OLS)

$$\hat{\boldsymbol{\beta}}^r = (\mathbf{X}^\top\,\mathbf{X} + \lambda\,\mathbf{I})^{-1}\,\mathbf{X}^\top\,\mathbf{y}$$

   - Exercise: show this

Why ridge?

$$\hat{\boldsymbol{\beta}}^{r} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \; ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2$$

1. The solution exists and is unique even when $p > n$ (unlike OLS)

$$\hat{\boldsymbol{\beta}}^{r} = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

   - Exercise: show this

2. Assume the setup from before: $y = X\beta + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 \mathrm{I}$

   There is **always** a value of λ where the ridge MSE is less than the OLS MSE:

$$\mathrm{MSE}\left(\mathbf{X}\hat{\boldsymbol{\beta}}^{r}(\lambda)\right) < \mathrm{MSE}\left(\mathbf{X}\hat{\boldsymbol{\beta}}^{OLS}\right)$$

Why ridge?

$$\hat{\boldsymbol{\beta}}^r = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ ||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_2^2$$

1. The solution exists and is unique even when $p > n$ (unlike OLS)

$$\hat{\boldsymbol{\beta}}^r = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}$$

   - Exercise: show this

2. Assume the setup from before: $y = X\beta + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2 \mathrm{I}$

   There is **always** a value of λ where the ridge MSE is less than the OLS MSE:

$$\mathrm{MSE}\left(\mathbf{X}\hat{\boldsymbol{\beta}}^r(\lambda)\right) < \mathrm{MSE}\left(\mathbf{X}\hat{\boldsymbol{\beta}}^{OLS}\right)$$

3. Handles correlated features well -- features that are highly correlated tend to get shrunken together, i.e. they are given equal contribution to the linear model

Ridge in practice

- Don't want to penalize the intercept
  - Before applying regularization, center columns of X and y
- Most of the time, should also scale columns of X so that we don't penalize some coefficients more than others simply because of different scales
- These practical guidelines apply to all regularization methods

Regularization Part 2: Lasso

Next time...

Practice midterm solutions