Machine Learning for Business Analytics

Lecture 03

Recap of lecture 02

- Overfitting (low train MSE, high test MSE) and underfitting (high train and test MSE)

- The bias-variance tradeoff in machine learning
  - What is bias? Being consistently off target
  - What is variance? Making prediction that vary a lot depending on the training sample.

- How do bias and variance relate MSE?
  - MSE = Variance + Bias^2 + Irreducible Error

- Exploring the bias-variance tradeoff in practice
  - Recall the plot with 5 x 2 datasets

- Introduction to model selection
  - Forward and backward selection, more today.

Today, lecture 03

- Model selection
- Forward and backward selection
- Lasso regression
- Ridge regression

Navigating the bias-variance trade-off

- How do we balance model complexity and prediction accuracy?
- How about this:
  - For every possible combination of predictors, fit a linear regression
  - Compute test MSE
  - Pick the best model

- Would this work?
- Is it practical?

Forward selection

Simple idea

1. Start with a model that only contains an intercept
2. Fit *p* linear regressions, one for each predictor X_p, and compute $MSE(p)$
3. Add the predictor X_p with the smallest $MSE(p)$ to the model
4. Repeat

Backward selection

Similar idea

1. Start with a model that only contains ALL variables
2. Fit p linear regressions, removing one predictor X_p each time, and compute $MSE(p)$
3. Choose the model with smallest $MSE(p)$
4. Repeat

Navigating the bias-variance trade-off

- Can we do anything other than variable selection?

- Recall that so far we have been fitting least squares which minimizes the loss function

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

- What if we selected a different loss function?
- Why would we do such a thing?

Regularization

- Key idea in modern analytics
- *Automatically* constrain model to subset of variables
- To do so we minimize a different object function than MSE
- Why would minimizing something other than MSE (on the training data) work? We will still evaluate our predictions using MSE on the test data after all.
- Essentially, we are trading off some bias for some variance. Hopefully, the trade-off is good: we sacrifice a bit in terms of bias to gain a lot in terms of variance.

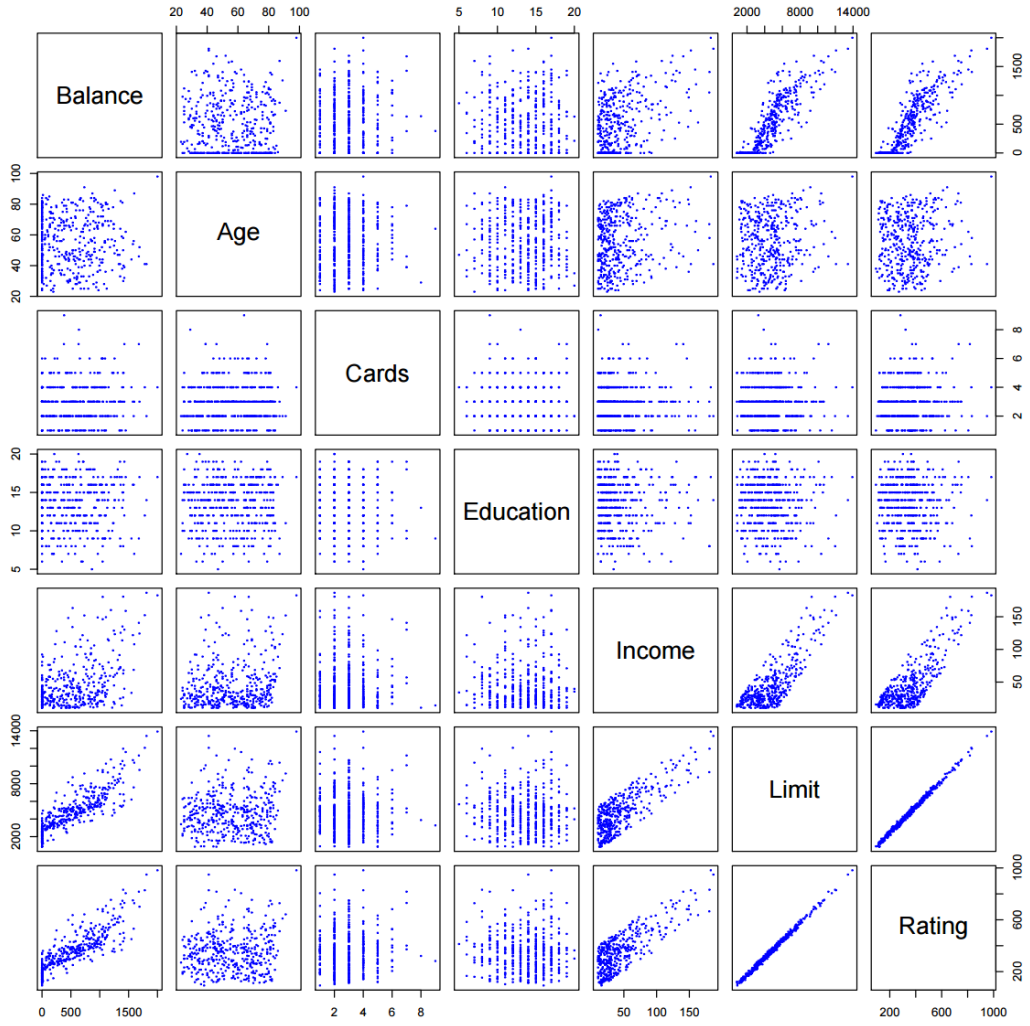
Regularization: Ridge regression

- Ridge regression minimizes the objective function

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2}_{\text{MSE}} + \lambda \sum_{j=1}^{p} \beta_j^2$$

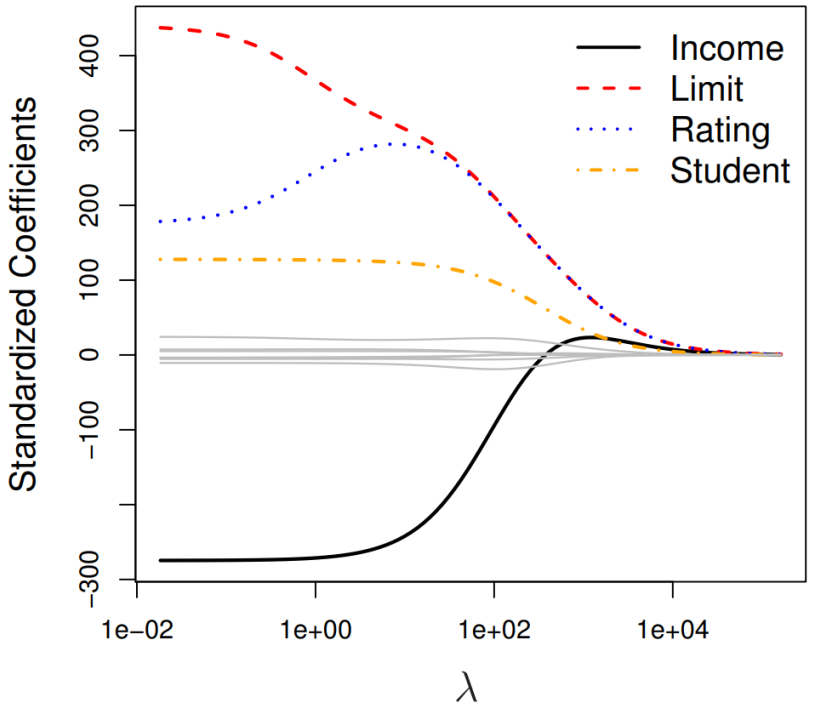
- MSE is the usual sum of squared residuals
- The second term penalizes large coefficients – known as shrinkage
- λ is a tuning parameter supplied by the user, or selected via cross-validation (more in the book)
- The tuning parameter controls the relative impact of these two terms – higher λ will shrink the coefficients more

Ridge example: predicting credit card balances



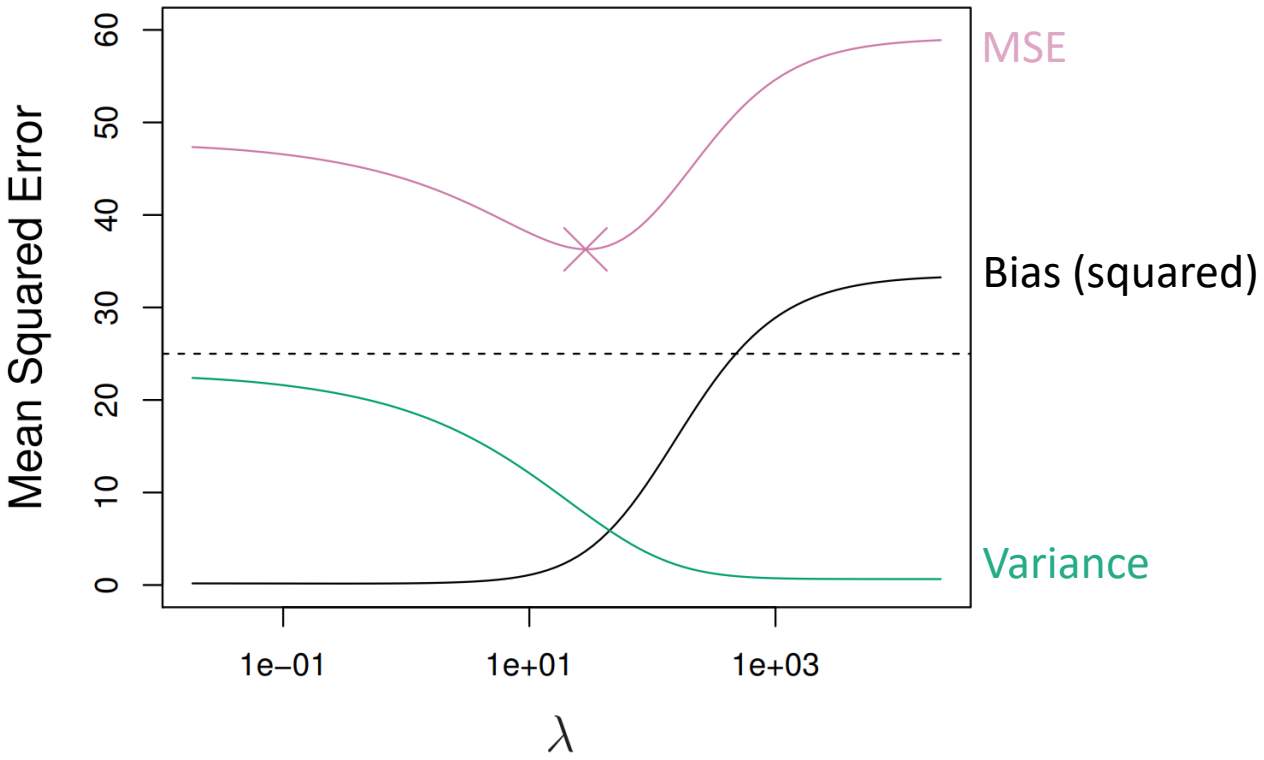
Source: Tibshirani et al.

Ridge example: predicting credit card balances



Source: Tibshirani et al.

Why can ridge regression be better than OLS?



Recall that MSE = Variance + Bias + Irreducible error

Source: Tibshirani et al.

The Lasso

- Least Absolute Shrinkage and Selection Operator

- What is a problem with ridge regression?

The Lasso

- What is the problem with ridge regression?
- Ridge estimates include all variables even though some of them end up with very small coefficients
- Ridge is not parsimonious

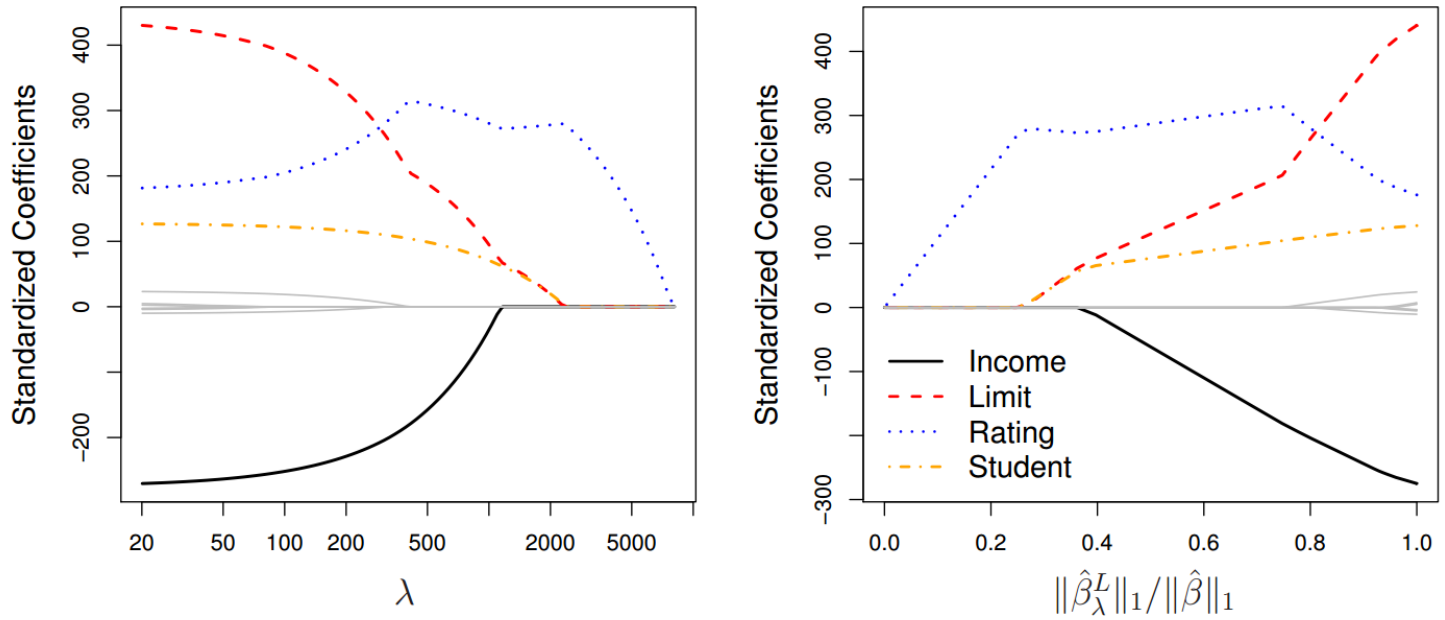
The Lasso

- The Lasso is a recent advance is statistical learning
- Select model parameters to minimize

$$MSE + \lambda \sum_{j=1}^{p} |\beta_j|$$

- The penalty term now includes the absolute values of the coefficients instead of their squares. This is known as the L1 norm.
- Like Ridge, Lasso shrinks coefficients towards zero
- Unlike Ridge, the Lasso sets some coefficients to zero exactly!
- Therefore it yields sparse, parsimonious models and can be used to do variable selection.

Lasso on the credit card data



The notation $||\beta||_1$ denotes the L1 norm of a vector, and is defined as $||\beta||_1 = \sum_j |\beta_j|$.

It captures how big the model parameters are when considered together.

Lasso or ridge?

- It depends on your dataset
- Lasso truly assumes that many coviarates have an exact ZERO effect
- If that's the case for you data, Lasso will be better than ridge regression

END

- End