MS&E 226: Fundamentals of Data Science

Lecture 4: Introduction to prediction

Ramesh Johari
rjohari@stanford.edu

Generalization

Where did our data come from?

Throughout the lecture:

- $\mathbf{Y}$ is the vector of n observed outcomes
- $\mathbf{X}$ is the corresponding matrix of covariates: n rows, with p covariates in each row

What process generated $\mathbf{X}$ *and* $\mathbf{Y}$*?*

Population vs. sample

The observed data we have, $\mathbf{Y}$ and $\mathbf{X}$, are referred to as the *sample*.

These came from some system or data-generating process, that we refer to as the *population*.

Think of surveys: we try to understand the broader population through a smaller sample.

The population model: A probabilistic view

How do we reason about the population? Using a *probabilistic model*.

- There is a probability distribution of $\vec{X} = (X_1, \ldots, X_p)$ in the population.
- And Y has a conditional probability distribution *given* $\vec{X}$

Together, these give a *joint* distribution over $\vec{X}$ and Y.

Example: the *linear normal population model*.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,$$

where $\vec{X}$ is jointly multivariate normal, and ε is normal with zero mean, independent of $\vec{X}$.

Example

Suppose in a population that father's heights are normally distributed with mean 69 inches and variance 4 inches.

Suppose that if a father has height $X = x$, his child's height is normally distributed with mean $40 + 0.4 \times x$ and variance 3 inches.

Then the population model is that:

$$Y = 40 + 0.4 \times X + \varepsilon$$

where $X \sim N(69, 4)$, $\varepsilon \sim N(0, 3)$, and X and ε are independent.

Generalization

The following reaction is quite common:

> *Wait, you're saying that the covariates and outcomes are* random*? Then why do I have a fixed dataset that I can see, that is definitively* not random*?*

The idea is that we use *the sample* (the dataset) we have to reason about *the population*.

This is called *generalization*.

Generalization

The first step to reasoning about the population is to build a *fitted model*: a function $\hat{f}$ that uses $\mathbf{X}$ and $\mathbf{Y}$ to capture the relationship between $\vec{X}$ and Y in the population:

$$Y \approx \hat{f}(\vec{X}).$$

A key example is the OLS approach to linear regression we have been studying:

- Given the data $\mathbf{X}$ and $\mathbf{Y}$, find coefficients $\hat{\boldsymbol{\beta}}$ such that $\|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$ is minimized.
- $\hat{f}(\vec{X}) = \sum_j \hat{\beta}_j X_j$.

Prediction and inference

What are statements we want to make using $\hat{f}$? They fall into two classes:

- *Prediction*. Given a new $\vec{X}$ that is observed, what is our "best" guess of the corresponding Y?
  $\implies$ *Predicting* that Y will be $\hat{f}(\vec{X})$.
- *Inference*. Describe the population model: the joint distribution of $\vec{X}$ and Y.
  $\implies$ *Interpreting* the structure of $\hat{f}$.

It may seem puzzling that these are different: can we make good predictions without good inference?

Example: Breast cancer risk and wealth

Consider the following story:

abc NEWS U.S. World Politics Entertainment Health Tech ...

Breast Cancer Risk Associated With Wealth

By JOY VICTORY • Dec. 1, 2005

Share with Facebook Share with Twitter

0 SHARES

Women who live in regions of the United States known as breast cancer "hot spots" may have an increased risk because of personal wealth and not pollution or electrical wires, researchers say.

Deborah Winn, a scientist with the National Institutes of Health, states in the December issue of the journal Nature Reviews Cancer that the most likely reason that women in certain communities -- such as Long Island or San Francisco -- have increased breast cancer risk is that those areas are populated by wealthy women. Winn's article analyzes a series of studies conducted by the Long Island Breast Cancer Study Project in New York.

These women tend to have children later, have fewer children, and are more likely to receive costly replacement hormone therapy -- all of which are linked to increased breast cancer risk.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.
- If we made everyone poorer, there would be fewer cases of breast cancer.

Example: Breast cancer risk and wealth

What can we say?

- Wealth is predictive of breast cancer.
- Breast cancer is predictive of wealth.
- The reason certain women have breast cancer is that they are wealthier.
- The reason certain women are wealthier is that they have breast cancer.
- If wealth increases, then incidence of breast cancer increases.
- If we made everyone poorer, there would be fewer cases of breast cancer.

Moral:

Prediction relies on correlation, not causation.

Example: Education and income

David Card, in his paper "The Causal Effect of Education on Earnings":

> *In the absence of experimental evidence, it is very difficult to know whether the higher earnings observed for better educated workers are caused by their higher education, or whether individuals with greater earning capacity have chosen to acquire more schooling.*

Example: Internet marketing

Suppose a customer sees multiple channels of advertising from you: a social media ad, a display ad, a promoted tweet, e-mail ad, etc..

At the time of placing ads, you have demographic information about the customer.

- *Prediction* asks: Will this customer purchase or not? How much is this customer going to spend?
- *Inference* asks: Which campaign is most responsible for the customer's spend?

Often you can make great predictions, even if you cannot infer the value of the different campaigns.[1]

[1] The latter problem is the *attribution* problem.

Prediction

The prediction problem

In this part of the class we focus only on the prediction problem:

Given data $\mathbf{X}$ *and* $\mathbf{Y}$*, construct a fitted model* $\hat{f}$ *so that given a new covariate vector* $\vec{X}$ *from the population, the* prediction error *between* $\hat{f}(\vec{X})$ *and the corresponding* Y *is minimized.*

How do we measure prediction error?

Classification vs. regression

Two broad classes of problems:

1. *Regression*: Y is a continuous variable (numeric). Examples:
   - Predict wealth given demographic factors
   - Predict customer spend given profile
   - Predict earthquake magnitude given seismic characteristics
   - Predict level of antigen given biological markers
2. *Classification*: Y is a categorical variable (factor). Examples:
   - Is this e-mail spam or not?
   - What zip code does this handwriting correspond to?
   - Is this customer going to buy an item or not?
   - Does this patient have the disease or not?

Prediction error

Measurement of prediction error depends on the type of prediction problem.

For regression, examples of prediction error measures include:

- *Squared error* $(Y - \hat{f}(\vec{X}))^2$;
- *Absolute deviation* $|Y - \hat{f}(\vec{X})|$.

For classification, a common example of prediction error is *0-1 loss*: the error is 1 if $Y \neq \hat{f}(\vec{X})$, and 0 otherwise.

Prediction error

For now we will focus on *regression* with *squared error* as our measure of prediction error.

Suppose we are given data $\mathbf{X}$ and $\mathbf{Y}$. What should we aim to do? Minimize the *generalization* error (or *test* error):

known

$$\mathbb{E}_{\vec{X},Y}[(Y - \hat{f}(\vec{X}))^2 | \mathbf{X}, \mathbf{Y}].$$

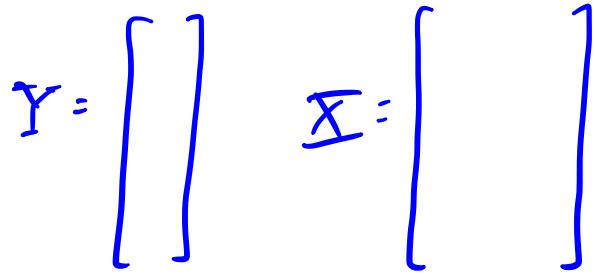
I.e.: "Minimize prediction error on new data."

Note that in this definition we condition on $\mathbf{X}$ and $\mathbf{Y}$: the data is given.

The only *randomness* is in the new sample $\vec{X}$ and Y, as denoted by the subscripts on the expectation.[2]

[2] There are other forms of generalization error; e.g., you might also assume the new $\vec{X}$ is also known. We will return to this later.

Training vs. validation vs. testing



With enough data, we can build effective predictive models as follows:

1. Separate data into three groups: *training*, *validation*, and *test*.
2. Use training data to fit different models ($\hat{f}$'s).
3. Use validation data to estimate generalization error of the different models, and pick the best one.
4. Use test data to assess performance of the chosen model.

Question: Why do we need to separate validation data and test data?

Validation

The validation step estimates the generalization error of the different models, and chooses the best one.
Formally:

- Suppose samples $(\tilde{\mathbf{X}}_1, \tilde{Y}_1), \ldots, (\tilde{\mathbf{X}}_k, \tilde{Y}_k)$ in the validation set.
- For each fitted model $\hat{f}$, estimate the generalization error as follows:

$$\frac{1}{k}\sum_{i=1}^{k}(\tilde{Y}_i - \hat{f}(\tilde{\mathbf{X}}_i))^2. \tag{1}$$

- Choose the model with the smallest generalization error

estimated

as $k \to \infty$ converges to $\mathbb{E}_{\vec{x},Y}\left[(Y - \hat{f}(\vec{x}))^2 \mid \mathcal{X}, \mathcal{T} \right]$ (training data)

Validation

Why does this work?

Testing

Importantly, the validation error of the best model in the validation step is typically an *underestimate* estimate of the true generalization error. Why?

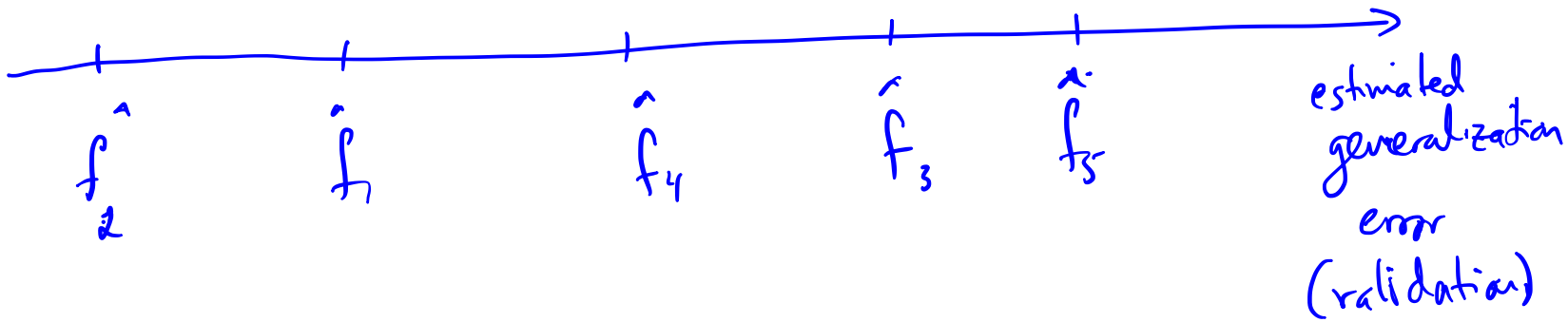
Consider this example:

- Suppose two i.i.d. random variables Z_1, Z_2.
- We choose the minimum.
- Is $\mathbb{E}[\min\{Z_1, Z_2\}]$ the same as $\mathbb{E}[Z_1]$ or $\mathbb{E}[Z_2]$? No: it is less than both.

Key point: Expected value of the minimum is smaller than the minimum of the expected value.

Testing

Importantly, the validation error of the best model in the validation step is typically an *underestimate* of the true generalization error. Why?



Testing

To obtain an accurate (i.e., *unbiased*) estimate of the generalization error of the selected model, we use another holdout set, called the *test* set.

Suppose that samples $(\tilde{\mathbf{X}}_{k+1}, \tilde{Y}_{k+1}), \ldots, (\tilde{\mathbf{X}}_{\ell}, \tilde{Y}_{\ell})$ are in the test set.

Let $\hat{f}^*$ be selected model. Then an unbiased estimate of generalization error is:

$$\frac{1}{\ell - k} \sum_{i=k+1}^{\ell} (\tilde{Y}_i - \hat{f}^*(\tilde{\mathbf{X}}_i))^2.$$

Note that in some instances, an estimate of generalization error is not needed, so there is no test set; in that case the terms "validation set" and "test set" are sometimes used interchangeably.

Train, validate, test: Linear regression

Suppose we are given a large dataset with p covariates per observed outcome.

We can build a predictive linear regression model as follows:

1. Separate data into three groups: training, validation, and test.
2. Use the training data to build a collection of linear regression models, using different sets of covariates, higher order terms, interactions, transformed variables, regularization, etc.
3. Use validation data to estimate generalization error of the different models, and pick the best one.
4. Use test data to assess performance of the chosen model.

Examples

Example: Model selection, validation, and testing

For this example, we generate 300 X_1, X_2 as i.i.d. $N(0,1)$ random variables.

We then generate 300 Y random variables as:

$$Y_i = 1 + 2X_{i1} + 3X_{i2} + \varepsilon_i,$$

where ε_i are i.i.d. $N(0,5)$ random variables.
The training, validation, and test separation is 100/100/100 samples, respectively.

Example: Model selection, validation, and testing

We trained the following five models, then ran them through the validation and test set.

For each we computed the square root of the mean squared prediction error (RMSE).[3]

Model	Training	Validation	Test
`Y ~ 1 + X1`	5.37	5.58	6.64
`Y ~ 1 + X2`	4.87	4.95	6.06
`Y ~ 1 + X1 + X2`	4.44	4.67	5.76
`Y ~ 1 + X1 + X2 + I(X1^2) + I(X2^2)`	4.39	4.64	5.80
`Y ~ 1 + X1 + X2 + I(X1^2) + I(X2^2) + . . . I(X1^5) + I(X2^5)`	4.29	4.75	5.91

[3] RMSE = "root mean squared error"

The models

```
> display(fm1)
lm(formula = Y ~ 1 + X1, ...)
            coef.est coef.se
(Intercept) 1.10     0.54
X1          1.98     0.52
---
n = 100, k = 2
residual sd = 5.43, R-Squared = 0.13
```

The models

```
> display(fm2)
lm(formula = Y ~ 1 + X2, data = ...)
            coef.est coef.se
(Intercept) 1.22     0.49
X2          2.81     0.45
---
n = 100, k = 2
residual sd = 4.92, R-Squared = 0.29
```

The models

```
> display(fm3)
lm(formula = Y ~ 1 + X1 + X2, ...)
            coef.est coef.se
(Intercept) 1.09     0.45
X1          1.91     0.43
X2          2.76     0.41
---
n = 100, k = 3
residual sd = 4.51, R-Squared = 0.41
```

The models

```
> display(fm4)
lm(formula = Y ~ 1 + X1 + X2 + I(X1^2) + I(X2^2), ...)
            coef.est coef.se
(Intercept) 0.55     0.63
X1          1.89     0.43
X2          2.73     0.41
I(X1^2)     0.49     0.35
I(X2^2)     0.01     0.22
---
n = 100, k = 5
residual sd = 4.51, R-Squared = 0.42
```

The models

```
lm(formula = Y ~ 1 + X1 + X2 + I(X1^2) + I(X2^2) +
   + I(X1^3) + I(X2^3) + I(X1^4) + I(X2^4)
   + I(X1^5) + I(X2^5), ...)
             coef.est coef.se
(Intercept)  0.39     0.81
X1           0.41     1.50
X2           3.87     0.96
I(X1^2)      1.33     1.13
I(X2^2)     -0.34     0.62
I(X1^3)      1.19     1.27
I(X2^3)     -0.42     0.43
I(X1^4)     -0.22     0.27
I(X2^4)      0.04     0.07
I(X1^5)     -0.19     0.22
I(X2^5)      0.02     0.04
---
n = 100, k = 11
residual sd = 4.55, R-Squared = 0.44
```

An example with regularization: Baseball hitters

Data taken from *An Introduction to Statistical Learning*.

Consists of statistics and salaries for 263 Major League Baseball players.

We use this dataset to:

- Develop the train-test method
- Apply lasso and ridge regression
- Compare and interpret the results

We'll use the `glmnet` package for this example.

Loading the data

glmnet uses matrices rather than data frames for model building:

```
> library(ISLR)
> library(glmnet)

> data(Hitters)
> hitters.df = subset(na.omit(Hitters))

> X = model.matrix(Salary ~ 0 + ., hitters.df)
> Y = hitters.df$Salary
```

NOTE: the data should be standardized before running lasso or ridge! See updated notes on course site.

Training vs. test set

Here is a simple way to construct training and test sets from the single dataset:

```
train.ind = sample(nrow(X), round(nrow(X)/2))
X.train = X[train.ind,]
X.test = X[-train.ind,]
Y.train = Y[train.ind]
Y.test = Y[-train.ind]
```

Ridge and lasso

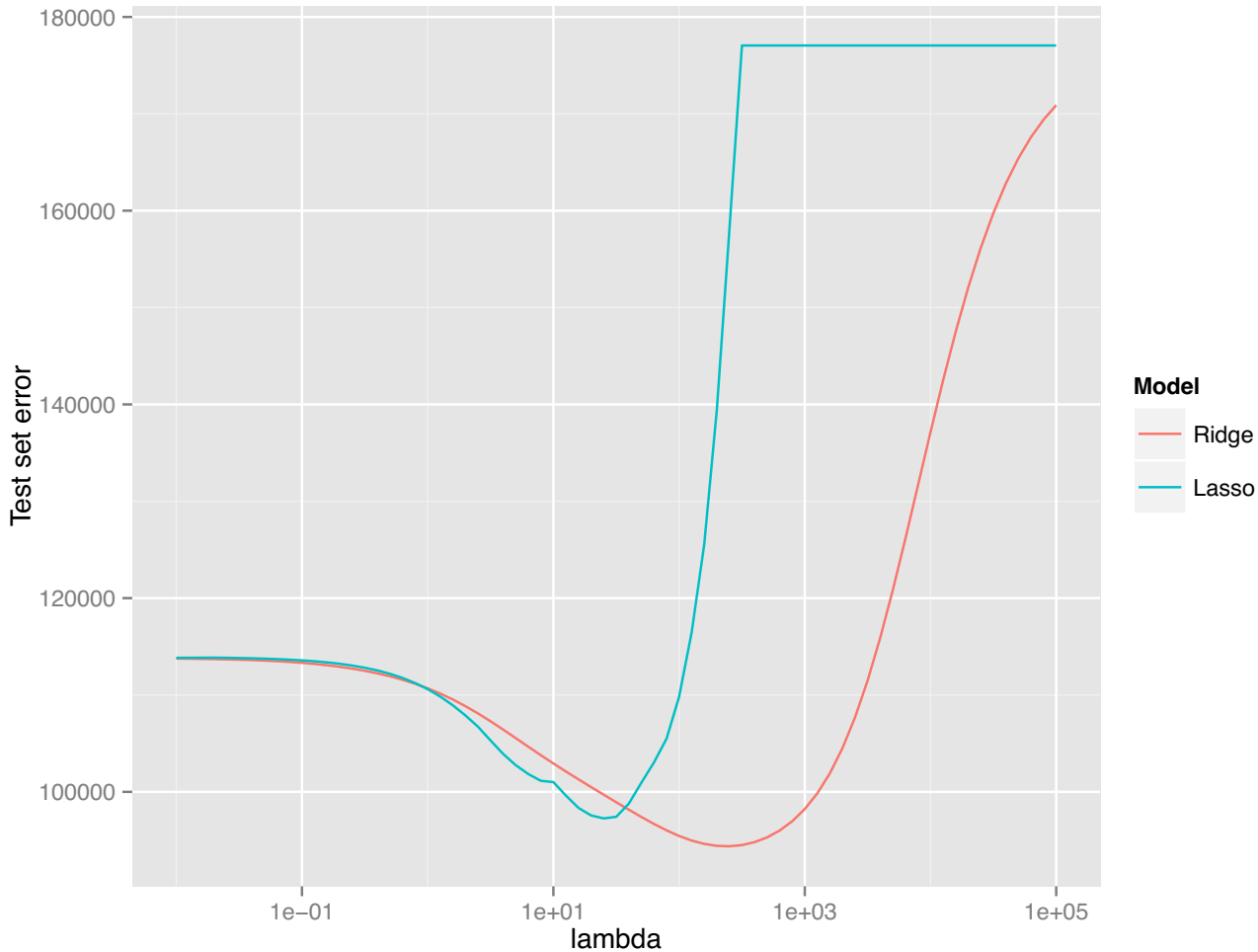
Building a lasso model:

```
> lambdas = 10^seq(-2,3.4,0.1)
> fm.lasso = glmnet(X.train,
      Y.train, alpha = 1,
      lambda = lambdas, thresh = 1e-12)
```

Setting `alpha = 0` gives ridge regression.
Make predictions as follows at $\lambda =$ `lam`:

```
> mean( (Y.test -
    predict(fm.lasso, s = lam, newx = X.test))^2 )
```

Results



What is happening to lasso?

Lasso coefficients

Using `plot(fm.lasso, xvar="lambda")`: