Regression: Predicting House Prices

STAT/CSE 416: Intro to Machine Learning
Hunter Schafer (slides by Emily Fox)
University of Washington
April 5, 2018
Generic linear regression model

Model:
\[ y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + \ldots + w_D h_D(x_i) + \varepsilon_i \]
\[ = \sum_{j=0}^{D} w_j h_j(x_i) + \varepsilon_i \]

feature 1 = \( h_0(x) \) … e.g., 1
feature 2 = \( h_1(x) \) … e.g., \( x[1] = \) sq. ft.
feature 3 = \( h_2(x) \) … e.g., \( x[2] = \) #bath

…

or, \( \log(x[7]) \times x[2] = \log(\text{#bed}) \times \text{#bath} \)

feature \( D+1 = h_D(x) \) … some other function of \( x[1], \ldots, x[d] \)
Training Data \rightarrow \text{Feature extraction} \rightarrow \text{ML model} \rightarrow \hat{y} \rightarrow \text{Quality metric} \rightarrow \text{ML algorithm} \rightarrow \hat{w} \rightarrow \text{Feature extraction} \rightarrow \text{ML model} \rightarrow \hat{y} \rightarrow \text{Quality metric} \rightarrow \text{ML algorithm} \rightarrow \hat{w}
Measuring loss

**Loss function:**

\[ L(y, f_\hat{w}(x)) \]

Cost of using \( \hat{w} \) at \( x \) when \( y \) is true

Actual value \( \hat{f}(x) = \text{predicted value} \hat{y} \)

**Examples:** (assuming loss for underpredicting = overpredicting)

**Absolute error:** \( L(y, f_\hat{w}(x)) = |y - f_\hat{w}(x)| \)

**Squared error:** \( L(y, f_\hat{w}(x)) = (y - f_\hat{w}(x))^2 \)
Fit data with a line or … ?

Dude, it’s not a linear relationship!
What about a quadratic function?

Dude, it’s not a linear relationship!
Even higher order polynomial

I can minimize your RSS

price ($)

square feet (sq.ft.)

©2018 Emily Fox
Assessing the loss
Part 1: Training error
Define training data

\[ y \]

price ($)

\[ x \]

square feet (sq.ft.)

\[ X \]
Define training data

(price ($) vs. square feet (sq.ft.))
Example:

Fit quadratic to minimize RSS

\[ \hat{w} \text{ minimizes RSS of training data} \]
Compute training error

1. Define a loss function $L(y, f_\hat{w}(x))$
   - E.g., squared error, absolute error,…

2. Training error
   = avg. loss on houses in training set
   $= \frac{1}{N} \sum_{i=1}^{N} L(y_i, f_\hat{w}(x_i))$
Example:
Use squared error loss \((y-f_\hat{w}(x))^2\)

Training error \((\hat{w}) = \frac{1}{N} * \)
\[\left( (\text{Price}_{\text{train} \ 1} - f_\hat{w}(\text{sq.ft}_{\text{train} \ 1}))^2 \right. \]
\[+ \left( (\text{Price}_{\text{train} \ 2} - f_\hat{w}(\text{sq.ft}_{\text{train} \ 2}))^2 \right. \]
\[+ \left( (\text{Price}_{\text{train} \ 3} - f_\hat{w}(\text{sq.ft}_{\text{train} \ 3}))^2 \right. \]
\[+ \ldots \text{include all} \]
\[\text{training houses}] \]
Training error vs. model complexity

Error vs. Model complexity graph.
Training error vs. model complexity

Model complexity

Error

price ($) vs. square feet (sq.ft.)
Training error vs. model complexity
Training error vs. model complexity

Error vs. Model complexity

price ($) vs. square feet (sq.ft.)
Training error vs. model complexity
Assessing the loss
Part 2: Generalization (true) error
Generalization error

Really want estimate of loss over all possible (house, $) pairs
Distribution over houses

In our neighborhood, houses of what # sq.ft. (_house_ ) are we likely to see?
Distribution over sales prices

For houses with a given # sq.ft. (ハウス), what house prices $ are we likely to see?
Generalization error definition

Really want estimate of loss over all possible (house, $) pairs

Formally:

generalization error = \mathbb{E}_{x,y}[L(y,f_\hat{w}(x))]
Generalization error vs. model complexity

Model complexity vs. Error

square feet (sq.ft.) vs. price ($)

$y = f(\hat{\mathbf{w}})$
Generalization error vs. model complexity

- Error
- Model complexity
- $f_{\hat{w}}$
- Price ($\$$)
- Square feet (sq.ft.)

Graph showing the relationship between error and model complexity.
Generalization error vs. model complexity
Generalization error vs. model complexity

Error vs. Model complexity graph

Price ($) vs. Square feet (sq.ft.) graph
Generalization error vs. model complexity

Error

Model complexity

price ($)

square feet (sq.ft.)

\( f_\hat{w} \)
Generalization error vs. model complexity
Generalization error vs. model complexity

Can’t compute!
Assessing the loss
Part 3: Test error
Approximating generalization error

Wanted estimate of loss over all possible (house, $) pairs

Approximate by looking at houses not in training set
Forming a test set

Hold out some \((\text{},\text{)}\) that are \textit{not} used for fitting the model.
Forming a test set

Hold out some \((\_\_\_\_, \$)\) that are *not* used for fitting the model.

Proxy for “everything you might see”

Test set
Compute test error

Test error

\[ \text{Test error} = \text{avg. loss on houses in test set} \]

\[ = \frac{1}{N_{\text{test}}} \sum_{i \text{ in test set}} L(y_i, f_\hat{w}(x_i)) \]

# test points

fit using training data

has never seen test data!
Example:
As before, fit quadratic to training data

\[ y = \hat{w} \text{ minimizes } \text{RSS of training data} \]
Example:
As before, use **squared error loss** \((y - f_{\hat{w}}(x))^2\)
Training, true, & test error vs. model complexity

Overfitting if:
Training/test split
Training/test splits

Training set

Test set

how many? vs. how many?
Training/test splits

Too few $\rightarrow$ $\hat{w}$ poorly estimated
Training/test splits

Too few $\to$ test error bad approximation of true error
Training/test splits

Typically, just enough test points to form a reasonable estimate of true error

If this leaves too few for training, other methods like cross validation (will see later…)}
3 sources of error +
the bias-variance tradeoff
3 sources of error

In forming predictions, there are 3 sources of error:

1. Noise
2. Bias
3. Variance
Data inherently noisy

\[ y_i = f_{w(true)}(x_i) + \varepsilon_i \]

Irreducible error
Bias contribution

Assume we fit a constant function
Bias contribution

Over all possible size N training sets, what do I expect my fit to be?
Bias contribution

\[ \text{Bias}(x) = f_{w(\text{true})}(x) - f_{\bar{w}}(x) \]

Is our approach flexible enough to capture \( f_{w(\text{true})} \)? If not, error in predictions.
Variance contribution

How much do specific fits vary from the expected fit?
Variance contribution

How much do specific fits vary from the expected fit?
Variance contribution

How much do specific fits vary from the expected fit?

Can specific fits vary widely?
If so, erratic predictions

\[ \text{low complexity} \rightarrow \text{low variance} \]
**Variance** of high-complexity models

Assume we fit a high-order polynomial
Variance of high-complexity models

Assume we fit a high-order polynomial
Variance of high-complexity models

high complexity $\rightarrow$ high variance
Bias of high-complexity models

\[ f_w(\text{true}) \]

\[ f_{\bar{w}} \]

High complexity \( \rightarrow \) low bias

Price ($) vs. square feet (sq.ft.)
Bias-variance tradeoff
Error vs. amount of data

Error

# data points in training set

©2018 Emily Fox
Summary of assessing performance
What you can do now…

• Describe what a loss function is and give examples
• Contrast training and test error
• Compute training and test error given a loss function
• Discuss issue of assessing performance on training set
• Describe tradeoffs in forming training/test splits
• List and interpret the 3 sources of avg. prediction error
  - Irreducible error, bias, and variance