Regression: Predicting House Prices

Emily Fox & Carlos Guestrin
Machine Learning Specialization
University of Washington
Predicting house prices
How much is my house worth?

I want to list my house for sale
How much is my house worth?
Look at recent sales in my neighborhood

• How much did they sell for?
**Plot recent house sales (Past 2 years)**

Terminology:
- $x$ – feature, covariate, or predictor
- $y$ – observation or response

- Past 2 years
- One prev. house sale in my neighborhood

price ($) vs. square feet (sq.ft.)
Predict your house by similar houses

No house sold recently had *exactly* the same sq.ft.
Predict your house by similar houses

- Look at average price in range
- Still only 2 houses!
- Throwing out info from all other sales
Linear regression
Use a **linear** regression model

Fit a line through the data

\[ f(x) = w_0 + w_1 x \]

parameters of model

square feet (sq.ft.)

price ($)

©2015 Emily Fox & Carlos Guestrin
Use a **linear** regression model

Fit a line through the data

\[
y = f_w(x) = w_0 + w_1 x
\]

function parameterized by \( w = (w_0, w_1) \)
Which line?

\[ f_w(x) = w_0 + w_1 x \]

different parameters \( w \)
“Cost” of using a given line

Residual sum of squares (RSS)

\[
\text{RSS}(w_0, w_1) = \\
(\$_{\text{house } 1} - [w_0 + w_1 \text{ sq.ft.}_{\text{house } 1}])^2 \\
+ (\$_{\text{house } 2} - [w_0 + w_1 \text{ sq.ft.}_{\text{house } 2}])^2 \\
+ (\$_{\text{house } 3} - [w_0 + w_1 \text{ sq.ft.}_{\text{house } 3}])^2 \\
+ \ldots \text{ [include all houses]}
\]
Find “best” line

Minimize cost over all possible $w_0, w_1$

$y$

price ($)

square feet (sq.ft.)

RSS($w_0, w_1$) =

$(\text{price}_{\text{house 1}} - (w_0 + w_1 \text{sq.ft.}_{\text{house 1}}))^2$

$+ (\text{price}_{\text{house 2}} - (w_0 + w_1 \text{sq.ft.}_{\text{house 2}}))^2$

$+ (\text{price}_{\text{house 3}} - (w_0 + w_1 \text{sq.ft.}_{\text{house 3}}))^2$

$+ ...$ [include all houses]

$\hat{w} = (\hat{w}_0, \hat{w}_1)$
Predicting your house price

$$f_{\hat{w}}(x) = \hat{w}_0 + \hat{w}_1 x$$

Best guess of your house price:

$$\hat{y} = \hat{w}_0 + \hat{w}_1 \text{ sq.ft. of your house}$$
Adding higher order effects
Fit data with a line or ... ?

You show your friend your analysis.
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!
What about a quadratic function?

\[ f_w(x) = w_0 + w_1 x + w_2 x^2 \]
Even higher order polynomial

I can minimize your RSS
Do you believe this fit?

My house isn’t worth so little
Evaluating overfitting via training/test split
Do you believe this fit?

Minimizes RSS, but bad predictions

price ($) vs. square feet (sq.ft.)
What about a quadratic function?

\[ f_w(x) = w_0 + w_1 x + w_2 x^2 \]
How to choose model order/complexity

• Want good predictions, but can’t observe future

• **Simulate predictions**
  1. Remove some houses
  2. Fit model on remaining
  3. Predict heldout houses
Training/test split

Terminology:
- training set
- test set
Training error

Training error ($w$) =

\[(\$_{\text{train } 1} - f_{w}(\text{sq.ft.}_{\text{train } 1}))^2 + (\$_{\text{train } 2} - f_{w}(\text{sq.ft.}_{\text{train } 2}))^2 + (\$_{\text{train } 3} - f_{w}(\text{sq.ft.}_{\text{train } 3}))^2 + \ldots \text{ [include all training houses]}\]

Minimize to find $\hat{w}$
Test error

\[
\text{Test error} (\hat{\omega}) = (\$_{\text{test } 1} - f_{\hat{\omega}}(\text{sq.ft.}_{\text{test } 1}))^2 + (\$_{\text{test } 2} - f_{\hat{\omega}}(\text{sq.ft.}_{\text{test } 2}))^2 + (\$_{\text{test } 3} - f_{\hat{\omega}}(\text{sq.ft.}_{\text{test } 3}))^2 + \ldots \text{ [include all test houses]}
\]
Training/Test Curves

Error

Model complexity

linear

polynomial

15th order
Adding other features
Predictions just based on house size

- Only 1 bathroom!
- Not same as my 3 bathrooms

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

©2015 Emily Fox & Carlos Guestrin
Add more features

\[
f_w(x) = w_0 + w_1 \text{sq.ft.} + w_2 \# \text{bath}
\]
How many features to use?

• Possible choices:
  – Square feet
  – # bathrooms
  – # bedrooms
  – Lot size
  – Year built
  – ...

• See Regression Course!
Other regression examples
Salary after ML specialization

• How much will your salary be? \( y = \$$ \)
• Depends on \( x = \) performance in courses, quality of capstone project, # of forum responses, ...
Salary after ML specialization

\[ \hat{y} = \hat{w}_0 + \hat{w}_1 \text{ performance} + \hat{w}_2 \text{ capstone} + \hat{w}_3 \text{ forum} \]

informed by other students who completed specialization
Stock prediction

• Predict the price of a stock
• Depends on
  – Recent history of stock price
  – News events
  – Related commodities
Tweet popularity

• How many people will retweet your tweet?
• Depends on # followers,
  # of followers of followers,
  features of text tweeted,
  popularity of hashtag,
  # of past retweets,...
Smart houses

• Smart houses have many distributed sensors
• What’s the temperature at your desk? (no sensor)
  – Learn spatial function to predict temp
• Also depends on
  – Thermostat setting
  – Blinds open/closed or window tint
  – Vents
  – Temperature outside
  – Time of day
Summary for regression
What you can do now...

- Describe the input (features) and output (real-valued predictions) of a regression model
- Calculate a goodness-of-fit metric (e.g., RSS)
- Estimate model parameters by minimizing RSS (algorithms to come...)
- Exploit the estimated model to form predictions
- Perform a training/test split of the data
- Analyze performance of various regression models in terms of test error
- Use test error to avoid overfitting when selecting amongst candidate models
- Describe a regression model using multiple features
- Describe other applications where regression is useful