Linear Regression

https://xkcd.com/1007/

Slides from: Andrew Ng

CSC411: Machine Learning and Data Mining, Winter 2017

Michael Guerzhoy
<table>
<thead>
<tr>
<th>Training set of housing prices (Portland, OR)</th>
<th>Size in feet² ($x$)</th>
<th>Price ($) in 1000's ($y$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2104</td>
<td>460</td>
</tr>
<tr>
<td></td>
<td>1416</td>
<td>232</td>
</tr>
<tr>
<td></td>
<td>1534</td>
<td>315</td>
</tr>
<tr>
<td></td>
<td>852</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Notation:
- $m$ = Number of training examples
- $x$'s = “input” variable / features
- $y$'s = “output” variable / “target” variable
Housing Prices (Portland, OR)

Price (in 1000s of dollars)

Supervised Learning
Given the “right answer” for each example in the data.

Regression Problem
Predict real-valued output
How do we represent \( h \) ?

- We represent hypotheses about the data using the parameters \( \theta = (\theta_0, \theta_1) \).
- If the data is correctly predicted according to hypothesis \( h_\theta \), then \( y \approx h_\theta(x) = \theta_0 + \theta_1 x \).
- The learning algorithm finds the best hypothesis \( h_\theta \) for the training set.
- We can then estimate the values of \( y \) for the test set using that \( h_\theta \).
- If \( h_\theta(x) \) is a linear function of a real number \( x \), this procedure is called linear regression.
### Training Set

<table>
<thead>
<tr>
<th>Size in feet$^2$ ($x$)</th>
<th>Price ($$) in 1000's ($$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2104</td>
<td>460</td>
</tr>
<tr>
<td>1416</td>
<td>232</td>
</tr>
<tr>
<td>1534</td>
<td>315</td>
</tr>
<tr>
<td>852</td>
<td>178</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

**Hypothesis:** \( h_\theta(x) = \theta_0 + \theta_1 x \)

\( \theta_i \)'s: Parameters

How to choose \( \theta_i \)'s?
\[ h_\theta(x) = \theta_0 + \theta_1 x \]

\[
\begin{align*}
\theta_0 &= 1.5 \\
\theta_1 &= 0
\end{align*}
\]

\[
\begin{align*}
\theta_0 &= 0 \\
\theta_1 &= 0.5
\end{align*}
\]

\[
\begin{align*}
\theta_0 &= 1 \\
\theta_1 &= 0.5
\end{align*}
\]
Idea: Choose $\theta_0, \theta_1$ so that $h_\theta(x)$ is close to $y$ for our training examples $(x, y)$
Quadratic cost function – on the board
Hypothesis:  \( h_\theta(x) = \theta_0 + \theta_1 x \)

Parameters:  \( \theta_0, \theta_1 \)

Cost Function:  \( J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \)

Goal:  minimize \( J(\theta_0, \theta_1) \)
Cost Function Surface Plot
Contour Plots

• For a function $F(x, y)$ of two variables, assigned different colours to different values of $F$
• Pick some values to plot
• The result will be *contours* – curves in the graph along which the values of $F(x, y)$ are constant
\[ h_{\theta}(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))
Cost Function Contour Plot

\[ h_\theta(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))
$$h_\theta(x)$$
(for fixed $\theta_0, \theta_1$, this is a function of $x$)

$$J(\theta_0, \theta_1)$$
(function of the parameters $\theta_0, \theta_1$)
Have some function $J(\theta_0, \theta_1)$

Want $\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$

Outline:

- Start with some $\theta_0, \theta_1$
- Keep changing $\theta_0, \theta_1$ to reduce $J(\theta_0, \theta_1)$
  until we hopefully end up at a minimum
Gradient Descent on the board
$J(\theta_0, \theta_1)$
For Linear Regression, J is bowl-shaped ("convex")
Gradient Descent Example
\[ h_\theta(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))
\[ h_\theta(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))
$h_\theta(x)$
(for fixed $\theta_0$, $\theta_1$, this is a function of $x$)

$J(\theta_0, \theta_1)$
(function of the parameters $\theta_0$, $\theta_1$)
\[ h_\theta(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))

- **Training data**
- **Current hypothesis**
\( h_\theta(x) \)
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\( J(\theta_0, \theta_1) \)
(function of the parameters \( \theta_0, \theta_1 \))
\[ h_\theta(x) \]

(for fixed $\theta_0, \theta_1$, this is a function of $x$)

\[ J(\theta_0, \theta_1) \]

(function of the parameters $\theta_0, \theta_1$)
\[ h_\theta(x) \]
(for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\[ J(\theta_0, \theta_1) \]
(function of the parameters \( \theta_0, \theta_1 \))
$h_\theta(x)$
(for fixed $\theta_0$, $\theta_1$, this is a function of $x$)

$J(\theta_0, \theta_1)$
(function of the parameters $\theta_0$, $\theta_1$)
\( h_\theta(x) \) (for fixed \( \theta_0, \theta_1 \), this is a function of \( x \))

\( J(\theta_0, \theta_1) \) (function of the parameters \( \theta_0, \theta_1 \))
Linear Regression vs. k-Nearest Neighbours

Orange: $y = 1$
Blue: $y = 0$
Linear Regression vs. k-Nearest Neighbours

- Linear Regression: the boundary can only be linear
- Nearest Neighbours: the boundary can more complex

Which is better?
- Depends on what the *actual boundary* looks like
- Depends on whether we have enough data to figure out the *correct* complex boundary