logistic regression: three views

\[
\log \frac{P(1|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)}{1 - P(1|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m)} = w_0 + w_1 x_2 + w_2 x_2 + \ldots + w_m x_m
\]

Linear classifier

\[
P(1|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m) = \frac{1}{1 + e^{-(w_0 + w_1 x_2 + w_2 x_2 + \ldots + w_m x_m)}}
\]

Conditional model

logistic

\[
\arg\min_w \sum \log(1 + e^{-(w_0 + w_1 x_2 + w_2 x_2 + \ldots + w_m x_m)})
\]

Linear model

minimizing logistic loss

Logistic regression

Why is it called logistic regression?
A digression: regression vs. classification

Raw data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

extract features

features: $f_1, f_2, f_3, \ldots, f_n$

Label:

classification: discrete (some finite set of labels)

regression: real value

linear regression

Given some points, find the line that best fits/explains the data.

Our model is a line, i.e. we're assuming a linear relationship between the feature and the label value.

$$h(y) = w_1x_1 + b$$

Linear regression

Learn a line $h$ that minimizes some loss/error function.

$$error(h) = ?$$

Sum of the individual errors:

$$error(h) = \sum_{i=1}^{n} |y_i - h(f_i)|$$

0/1 loss!

Error minimization

How do we find the minimum of an equation?

$$error(h) = \sum_{i=1}^{n} |y_i - h(f_i)|$$

Take the derivative, set to 0 and solve (going to be a min or a max)

Any problems here?

Ideas?
Linear regression

We’d like to minimize the error
Find $w_1$ and $w_0$ such that the error is minimized

$\text{error}(h) = \sum_{i=1}^{n} (y_i - (w_1 f_i + w_0))^2$

We can solve this in closed form

Multiple linear regression

If we have $m$ features, then we have a line in $m$ dimensions

$h(\hat{f}) = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m$

weights
Multiple linear regression

We can still calculate the squared error like before

\[ h(\vec{f}) = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m \]

\[ \text{error}(h) = \sum_{i=1}^{n} (y_i - (w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m))^2 \]

Still can solve this exactly!

Logistic function

\[ \text{logistic}(x) = \frac{1}{1 + e^{-x}} \]

Logistic regression

Find the best fit of the data based on a logistic

Big Data

What is “big data”?  
What are some sources of big data?  
What are the challenges of dealing with big data?  
What are some of the tools you’ve heard of?  

For more info:  
http://www.youtube.com/watch?v=eEpxNQHtRkI
Hadoop: guest speaker

Patricia Florissi

CTO of EMC
PhD from Columbia University

Hadoop

http://www.youtube.com/watch?v=XtLXPLb6EXs