HADOOP

David Kauchak
CS451 – Fall 2013

Admin

Assignment 7

logistic regression: three views

$$\log\frac{P(1 \mid x_1,x_2,\ldots,x_m)}{1-P(1 \mid x_1,x_2,\ldots,x_m)} = w_0 + w_1x_2 + w_2x_2 + \ldots + w_mx_m$$

linear classifier

$$P(1 \mid x_1,x_2,\ldots,x_m) = \frac{1}{1+e^{-(w_0+w_1x_2+w_2x_2+\ldots+w_mx_m)}}$$

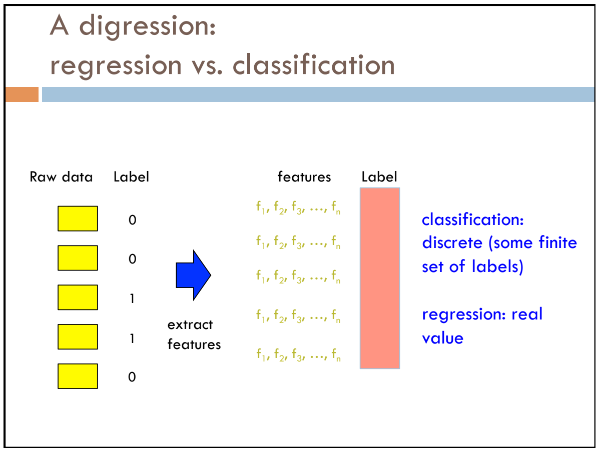
conditional model
logistic

$$\text{argmin}_{w,b}\sum_{i=1}^{n}\log(1+e^{-y_i(w_1x_2+w_2x_2+\ldots+w_mx_m+b)})$$

linear model
minimizing logistic loss

Logistic regression

Why is it called logistic regression?

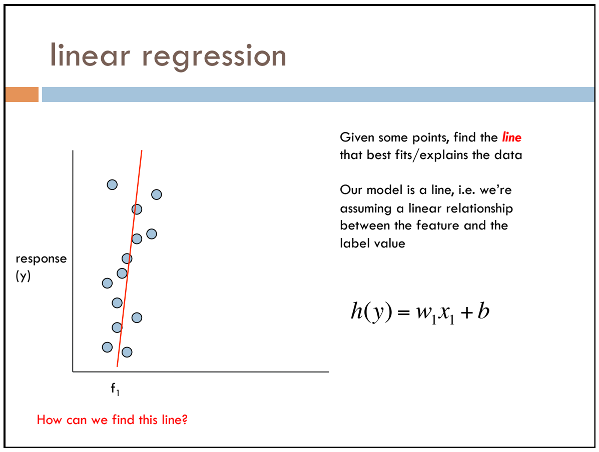


A digression: regression vs. classification

Raw data
Label
0
0
1
1
0
extract features
features
Label
$f_1, f_2, f_3, \ldots, f_n$
$f_1, f_2, f_3, \ldots, f_n$
$f_1, f_2, f_3, \ldots, f_n$
$f_1, f_2, f_3, \ldots, f_n$
$f_1, f_2, f_3, \ldots, f_n$

classification: discrete (some finite set of labels)

regression: real value



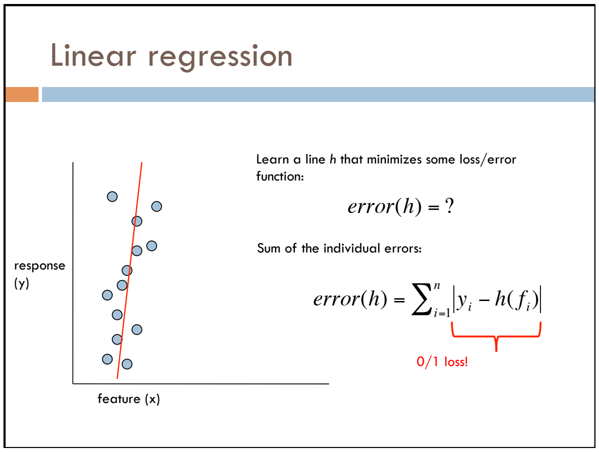
linear regression

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

$$h(y) = w_1 x_1 + b$$

How can we find this line?



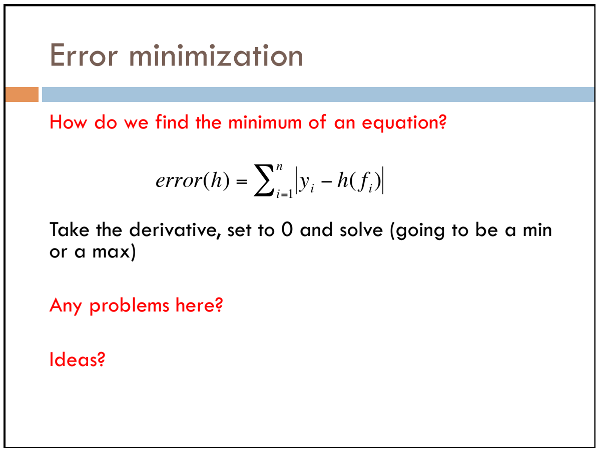
Linear regression

Learn a line *h* that minimizes some loss/error function:

$$error(h) = ?$$

Sum of the individual errors:

$$error(h) = \sum_{i=1}^{n} \underbrace{|y_i - h(f_i)|}_{\text{0/1 loss!}}$$



Error minimization

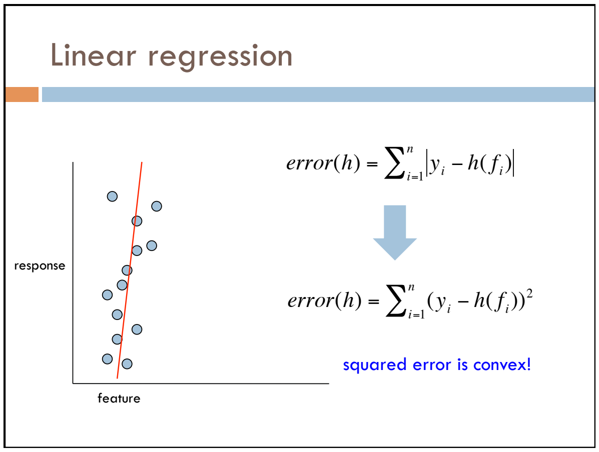
How do we find the minimum of an equation?

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

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

Any problems here?

Ideas?

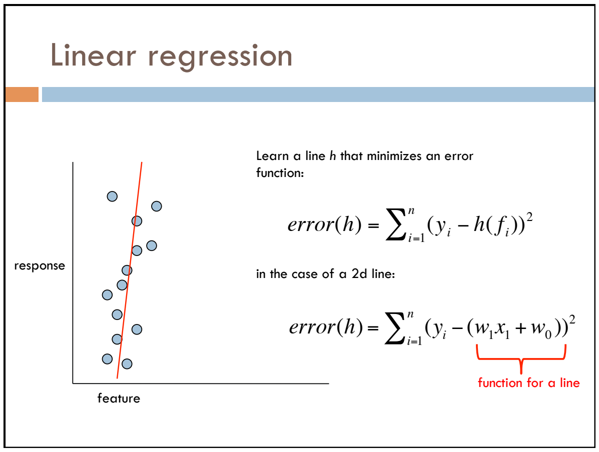


Linear regression

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

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

squared error is convex!



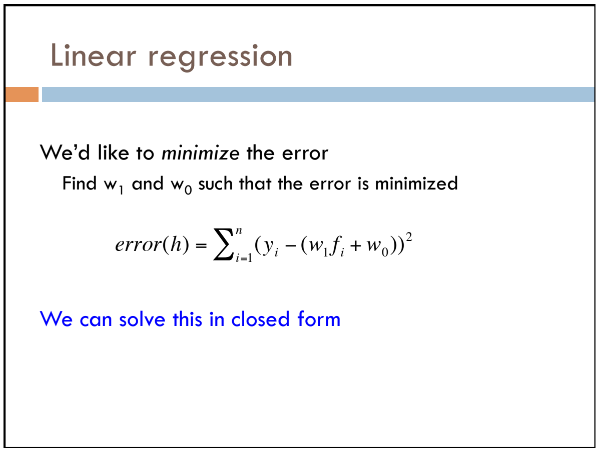
Linear regression

Learn a line *h* that minimizes an error function:

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

in the case of a 2d line:

$$error(h) = \sum_{i=1}^{n} (y_i - \underbrace{(w_1 x_1 + w_0)}_{\text{function for a line}})^2$$



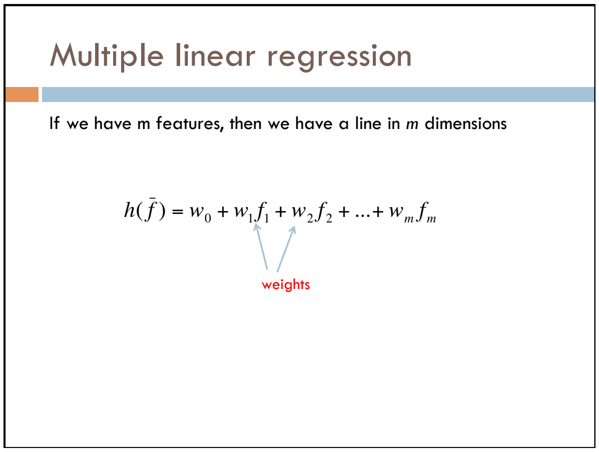
Linear regression

We'd like to *minimize* the error

Find w_1 and w_0 such that the error is minimized

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

We can solve this in closed form



Multiple linear regression

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

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

weights

Multiple linear regression

We can still calculate the squared error like before

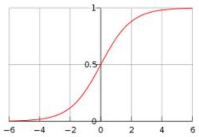
$$h(\bar{f}) = w_0 + w_1 f_1 + w_2 f_2 + \ldots + w_m f_m$$

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

Still can solve this exactly!

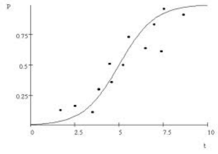
Logistic function

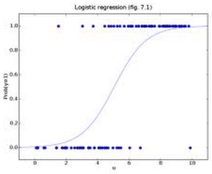
$$\text{logistic} = \frac{1}{1+e^{-x}}$$



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=eEpxN0htRKI

Hadoop: guest speaker

Patricia Florissi

http://www.mobileworldmag.com/emc-leads-it-transformation-at-the-emc-forum/

CTO of EMC

PhD from Columbia University

Hadoop

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