INTRO TO MACHINE LEARNING

Slides: Andrew NG

(→ free course on Coursera)

In a problem of prediction, the outcome is what is to be predicted.

The outcome can be a continuous variable

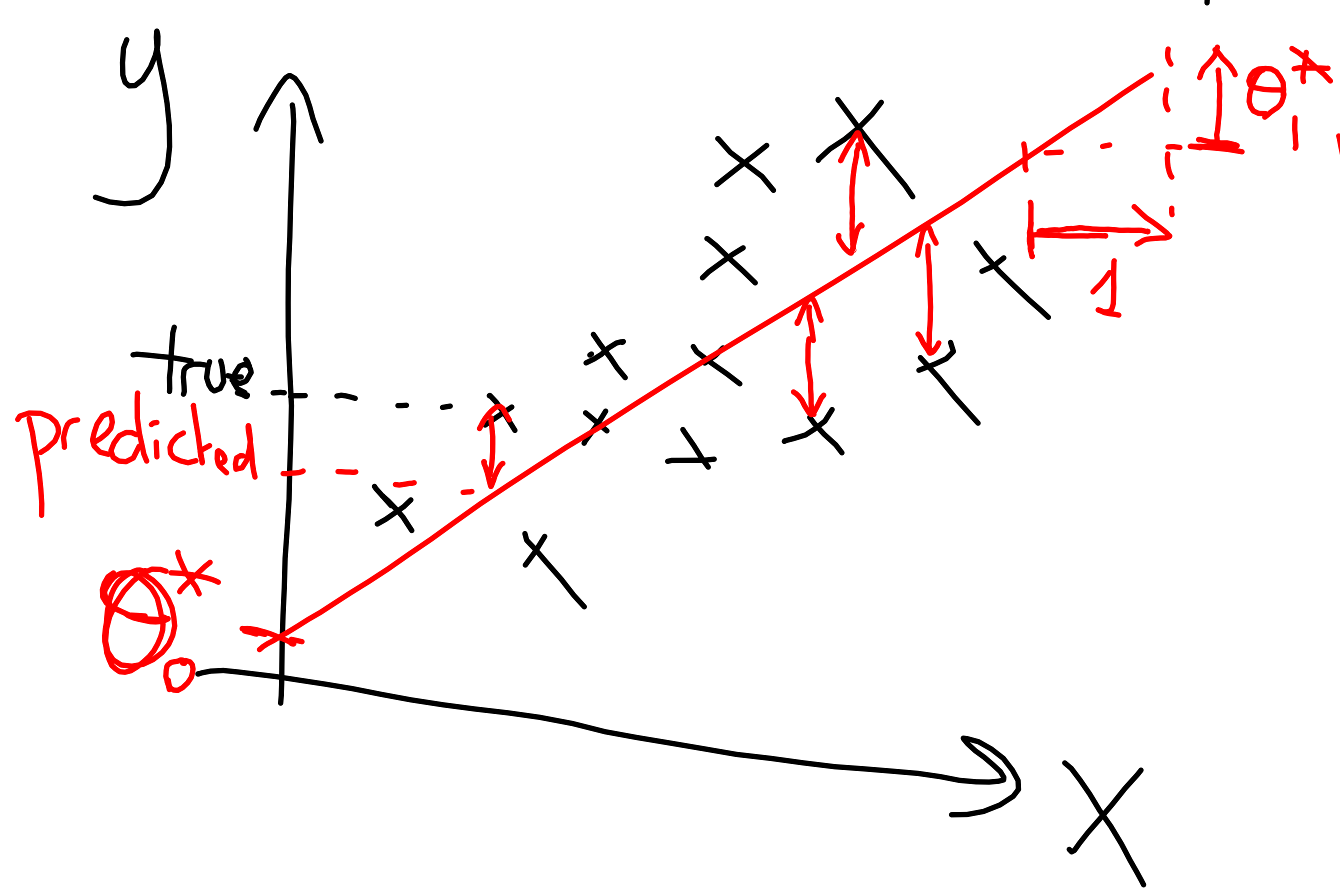
→ regression analysis

or a discrete/categorical variable

→ classification/clustering problem

LINEAR REGRESSION (UNIVARIATE)

* predictors : one continuous variable (x)
* outcome : a continuous variable (y)
* training dataset : n points $\left(x^{(i)}, y^{(i)}\right)$ $i \in [1, n]$



we want to get to a model whereby y is a function of X : $y = f(X)$

Linear regression means that our model must be of the form : $y = \underbrace{\theta_0}_{\text{intercept}} + \underbrace{\theta_1}_{\text{slope}} x \; \cancel{+ \theta_2 x^2}$ (quadratic term)

$\Rightarrow$ the learning will consist in learning the best values for the 2 parameters θ_0 and θ_1

"the best"?? $\Rightarrow$ we will define a cost function $J(\theta_0, \theta_1)$ to be minimized. Least-squares approach is to have

$$J(\theta_0, \theta_1) = \sum_i \left(\text{predicted}^{(i)} - \text{trueoutcome}^{(i)} \right)^2$$

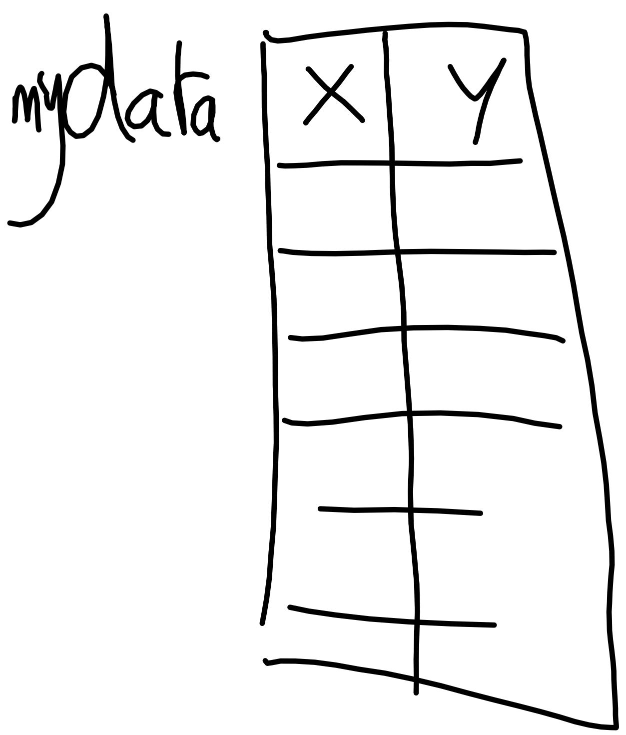
$$J(\theta_0, \theta_1) = \sum \left(\text{predicted}^{(i)} - \text{trueoutcome}^{(i)} \right)^2$$

Means: $J(\theta_0, \theta_1) = \sum_{i=1}^{n} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)} \right)^2$

There is an algebraic formula to calculate (θ_0^*, θ_1^*) that minimizes $J(\theta_0, \theta_1)$. Therefore we have our <u>regression model</u>:

$$y = \theta_0^* + \theta_1^* x$$

In R, linear regression is performed with the function lm() (think "linear model")



$\Rightarrow$ call: lm(y ~ x, data=mydata)

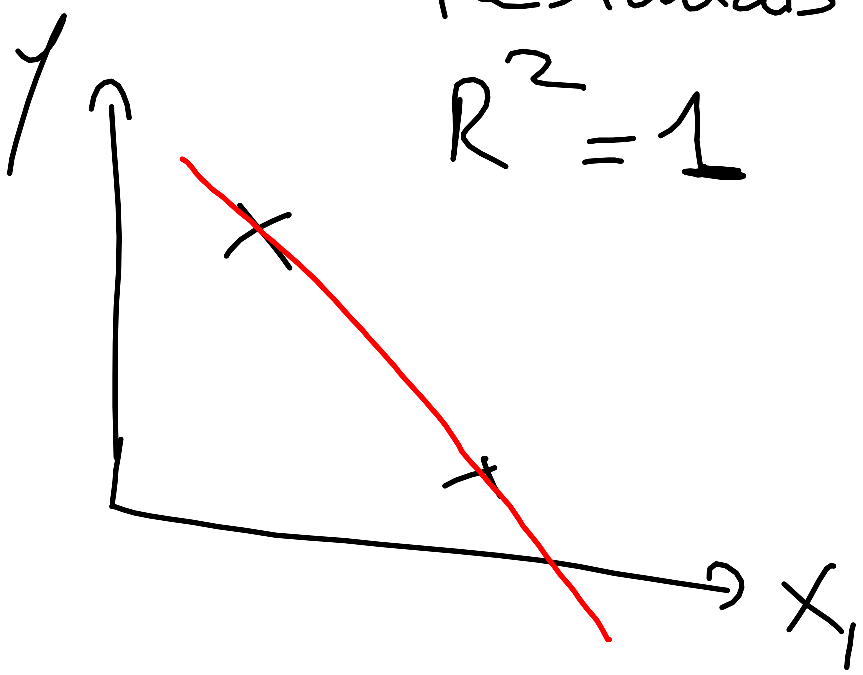
MULTIVARIATE (MULTIPLE) LINEAR REGRESSION

- still have a continuous outcome variable y
- $m > 1$ (continuous or not) predictor variables $x_1, x_2, x_3, \ldots, x_m$
- one training datapoint #i : $(x_1^{(i)}, x_2^{(i)}, \ldots, x_m^{(i)}, y^{(i)})$
- model : $y = \theta_0 + \sum_{j=1}^{m} \theta_j x_j$ ($m+1$ parameters)
- same Least squares approach to minimize the sum of the squared prediction errors

OVERFITTING 2

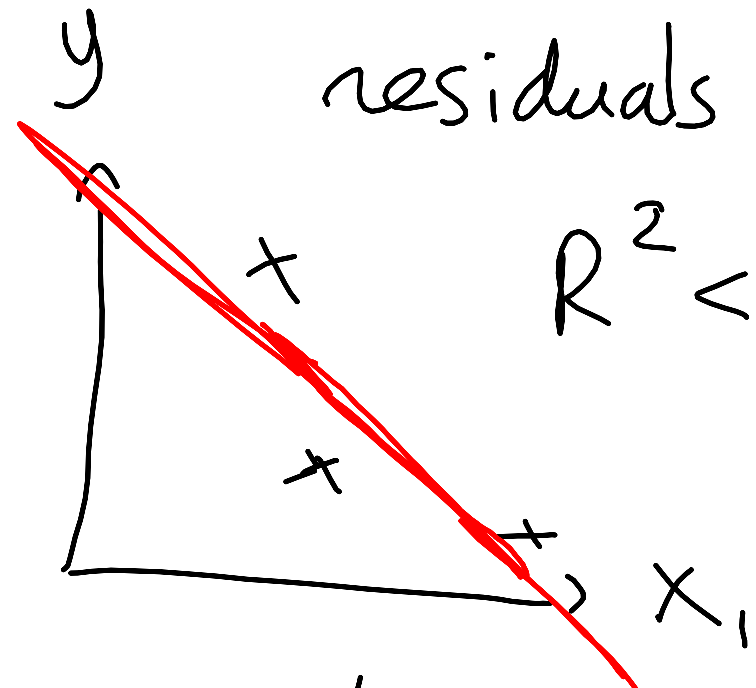
residuals $= 0$

$R^2 = 1$



residuals not zero

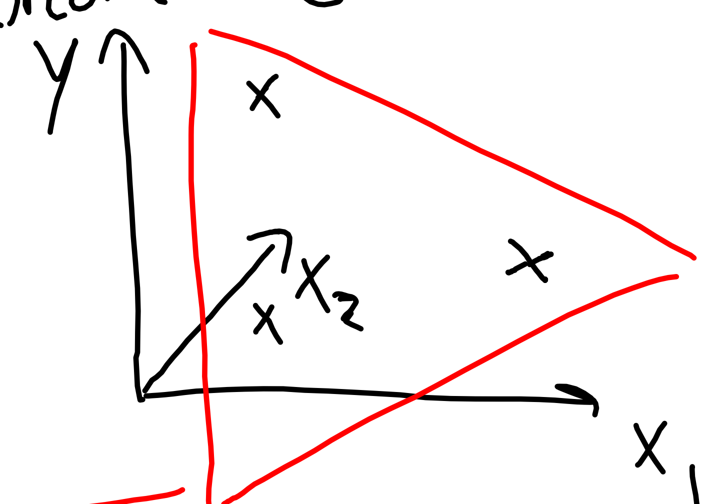
$R^2 < 1$



let us add one predictor

residuals $= 0$

$R^2 = 1$



model: $y = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

defines a plane

$\text{plane}(\theta_0^*, \theta_1^*, \theta_2^*)$

contains all three training datapoints

With a given fixed training set of n datapoints, increasing the number of predictors automatically improves the R squared measure (better fit).

$\Rightarrow$ with $(n-1)$ predictors, we reach $R^2 = 1$ even if some individual predictors don't make sense.

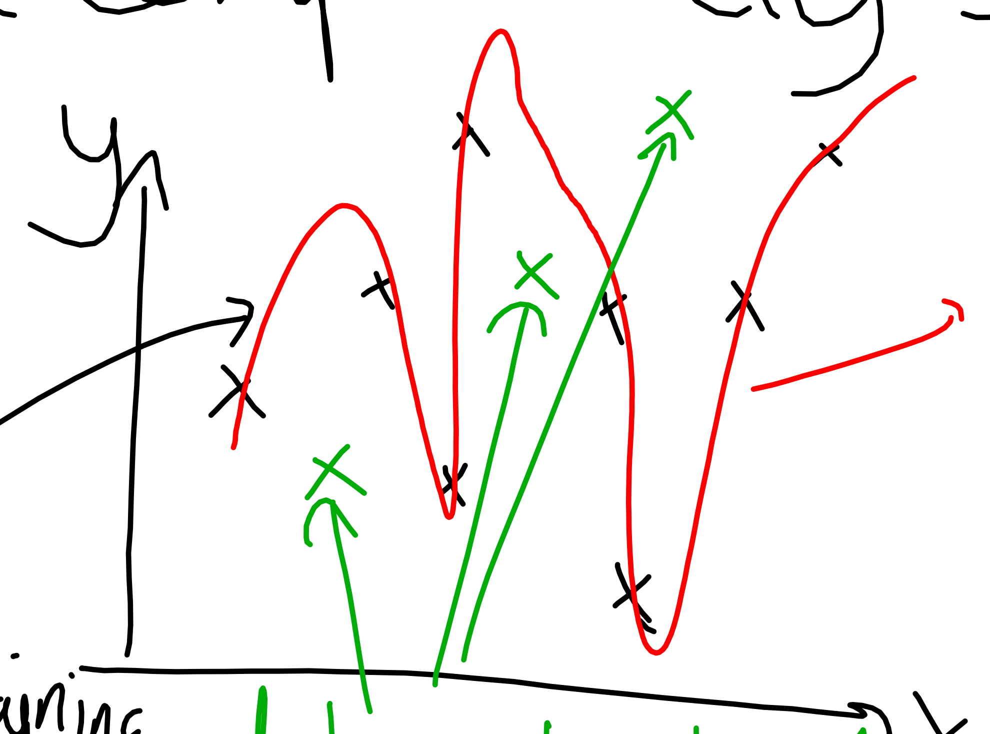
The adjusted R^2 includes a penalty in proportion with the number of predictors.

$\Rightarrow$ fairer to use the adjusted R^2.

Overfitting occurs when one designs a model of high complexity based on a comparatively small dataset.

example:

$R^2 = 1$ (on the training set)



many parameters, some with v. large absolute values

$$\left(y = \sum_{i=0}^{m} \theta_i \, x^i \right)$$

datapoints not in the original training set come with very poor predictions to avoid overfitting, don't try and go for too complex a representation on "simple" data:

- check the adjusted R squared
- add a predictor only if it <u>individually</u> helps decrease significantly the value of the cost function

→ automatic procedures (based on the AIC/BIC values) to determine the best model, at the same time accurate and generic enough.