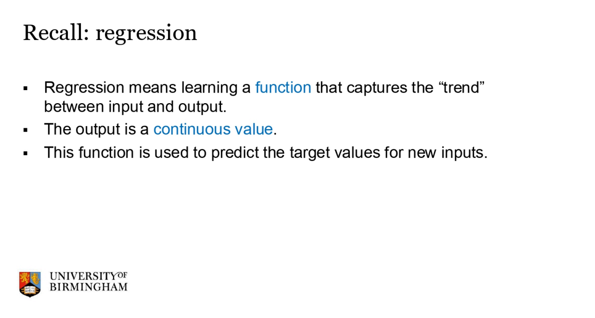
W2.2: Linear Regression

Linear Regression:

A ML {machine learning} algorithms for regression problems

Gradient descent:

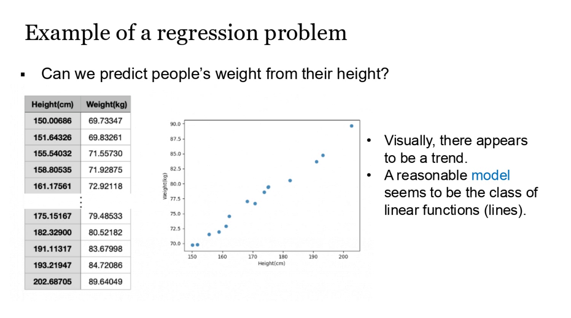
An optimisation technique used in ML {machine learning} algorithms



Recall: regression

- Regression means learning a function that captures the “trend” between input and output.
- The output is a continuous value.
- This function is used to predict the target values for new inputs.

UNIVERSITY OF BIRMINGHAM



Example of a regression problem

- Can we predict people’s weight from their height?

Height(cm)	Weight(kg)
150.00686	69.73347
151.64326	69.83261
155.54032	71.55730
158.80535	71.92875
161.17561	72.92118
⋮	
175.15167	79.48533
182.32900	80.52182
191.11317	83.67998
193.21947	84.72086
202.68705	89.64049

- Visually, there appears to be a trend.
- A reasonable model seems to be the class of linear functions (lines).

Univariate linear regression

- We are making our assumption on the function here.
- We have one input attribute (height) – hence the name **univariate**.

$$y = f(x; \underbrace{w_0, w_1}_{\text{free parameters}}) = w_1 x + w_0$$

dependent variable　　free parameters　　independent variable

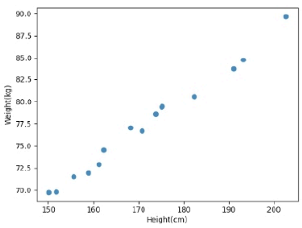
- Any line is described by this equation by specifying values for w_1 and w_0.

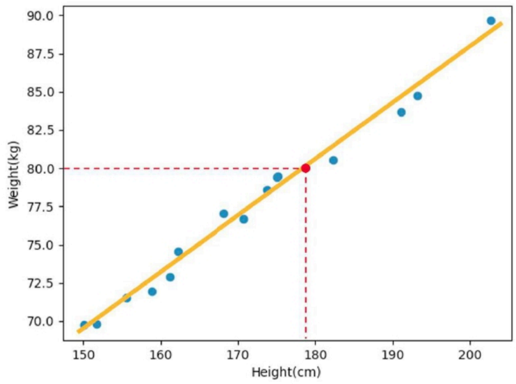
UNIVERSITY OF BIRMINGHAM

- ***The "free parameters" are not input attributes but rather values that the model learns to help it learn the trend based on the input values***

Check your understanding

Height(cm)	Weight(kg)
150.00686	69.73347
151.64326	69.83261
155.54032	71.55730
158.80535	71.92875
161.17561	72.92118
⋮	
175.15167	79.48533
182.32900	80.52182
191.11317	83.67998
193.21947	84.72086
202.68705	89.64049

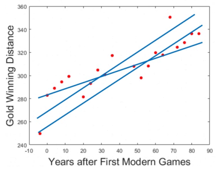




Suppose that from historical data someone calculated the parameters of our linear model are w_0 =1.68, w_1 =0.44. A new person (James) has height x=178cm. What is James weight?

UNIVERSITY OF BIRMINGHAM

Our goal: find the “best” line



- Which is the “best” line? That captures the trend in the data.
- Determine the “best” values for w_0 and w_1.

UNIVERSITY OF BIRMINGHAM

Loss/cost functions

- We need a criterion that tells us how good/bad that line is.
- Such criterion is called a loss function.

Terminology

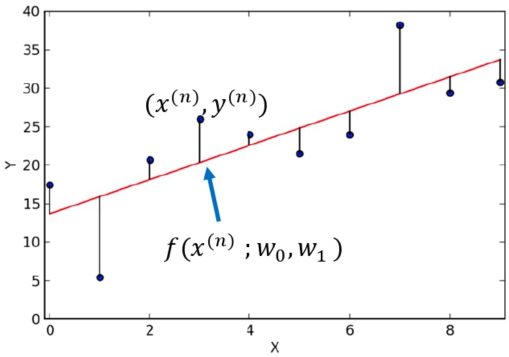
- Loss function = cost function = loss = cost = error function

UNIVERSITY OF BIRMINGHAM

We average the losses on all training examples

- For each training example (point) n = 1,…, N,

  The loss on the n-th point is the <u>mismatch/distance</u> between the output of the model for this point $f(x^{(n)}; w_0, w_1)$ and the observed target $y^{(n)}$.
- Average these losses.



UNIVERSITY OF BIRMINGHAM

Loss function

- The loss expresses an error, so it must be always non-negative.
- Absolute value loss (L1 loss):

$$L1 = |f(x) - y|$$

- Mean squared error loss (L2 loss):

$$L2 = (f(x) - y)^2$$

$$\boxed{g(w_0, w_1) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{(f(x^{(n)}; w_0, w_1) - y^{(n)})^2}_{\text{Loss for the n-th training example}}} \quad \textit{Empirical loss used by LR}$$

- 0/1 loss:

$$L_{0/1} = 0 \; if \; f(x) = y, else \; 1$$

UNIVERSITY OF BIRMINGHAM

The $\frac{1}{N}$ is important. Gives you the loss for the n-th training example not the sum of losses for the n-th training example which is what it would do with just the $\sum_{n=1}^{N}(f(x^{(n)}; w_0, w_1) - y^{(n)})^2$, without the $\frac{1}{N}$.

I'd assume the "0/1 loss" method is very general and imprecise

Check your understanding

- Suppose a linear function with parameters w_0=*0.5*, w_1 =*0.5*
- Computer the MSE value at the training example (1,3).

UNIVERSITY OF BIRMINGHAM

{Question strangely written but...: }

$$y = f(x; w_0, w_1) = w_0 x + w_1$$

$\Rightarrow$

$f(x) = (0.5 * 1) + 0.5 = 1$

Actual $y = 3$

Therefore:

Absolute value loss:

$|1 - 3| = 2$

Mean squared loss:

$(1 - 3)^2 = 4$

or

$\frac{1}{1} \sum_{n=1}^{1} (f(1^{(3)}; 0.5, 0.5) - 3^{(1)})^2$
$\Rightarrow (1 - 3)^2 = 4$

0/1 loss:

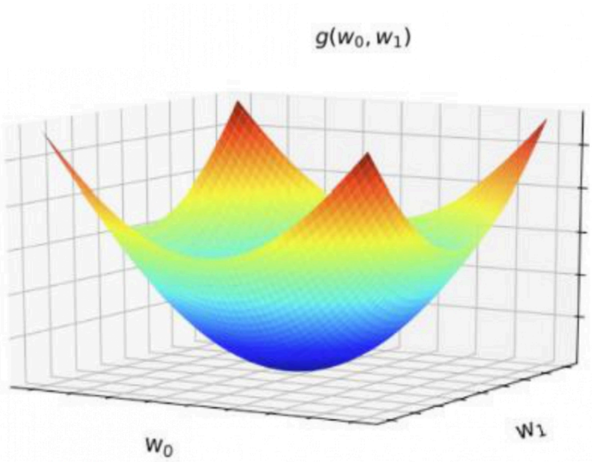
1

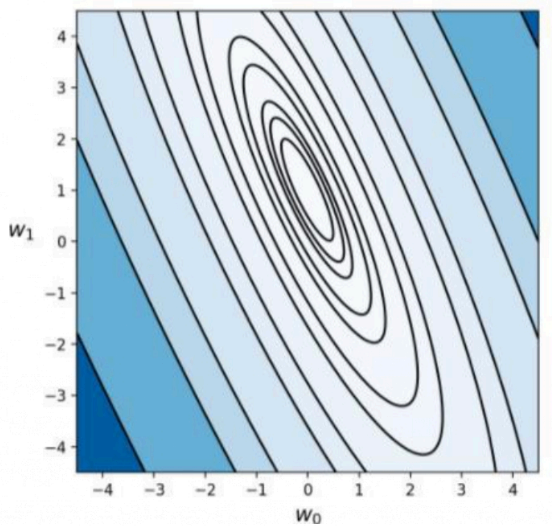
Univariate linear regression

- Given training data
$$(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots \ (x^{(N)}, y^{(N)})$$
- Fit the model
$$y = f(x; w_0, w_1) = w_1 x + w_0$$
- By minimizing the cost function
$$g(w_0, w_1) = \frac{1}{N} \sum_{n=1}^{N} (f(x^{(n)}; w_0, w_1) - y^{(n)})^2$$

UNIVERSITY OF BIRMINGHAM

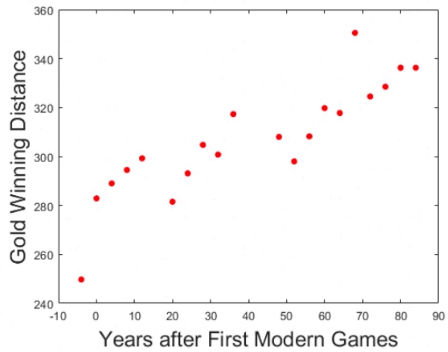
Cost function depends on the free parameter





Univariate linear regression

- Every combination of (w_0, w_1) has an associated cost.
- Key training task: find the 'best' values of (w_0, w_1) such that the cost is minimum.



UNIVERSITY OF
BIRMINGHAM