Linear Regression

Niclas Ståhl
niclas.stahl@his.se

Linear Regression

Regression in general

- Regression is a predictive analysis task concerned with predicting the real values y based on a set of attributes of an instance x.
- So the task is to find a model h so that $h(x) = y$.
- This is often intractable so we aim to find a model h so that $h(x) + \epsilon = y$ where ϵ is a small error.
- One example of regression analysis is to predict the price of a house given its size, number of rooms, ...

Linear regression

- Predict Y with $h(X|\theta)$ where h is a linear function.
- Recall the equation for a straight line:

$$y = mx + b$$

- m denotes the gradient or slope b denotes the intercept with the y axis.
- In machine learning the notations $y = \theta_1 * x + \theta_0$ or $y = w * x + b$ are often used instead.

The problem

So the problem is: **How to find the best values for w and b??**

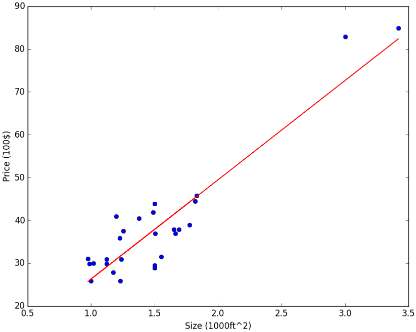


Figure: How to find these values of w and b.

The cost function

- First we need to define what *best* means.
- This is done by creating a cost function. That is a function that defines what we pay for being wrong.
- In some literature this is called the energy function.
- There are an infinite number of possible cost functions. But for simplicity it should be differentiable.
- (It is also good if the cost function is convex).

Cost function for linear regression

For linear regression the mean squared error (MSE) is often used as the cost function:

The Mean Squared Error

$$MSE(Y, \hat{Y}) = \frac{1}{2m} \sum_{i=1}^{m} (y_i - \hat{y}_i)^2$$

A graphical view of the MSE

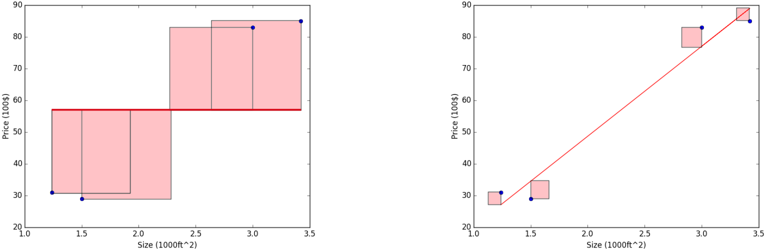


Figure: Illustration of the mean squared error

Optimization

- How can we find the parameters that give the mineralize the cost?

- In the same way as we can find the minimum of x^2: Gradient decsent.

- Example on whiteboard.

Gradient decsent - Example

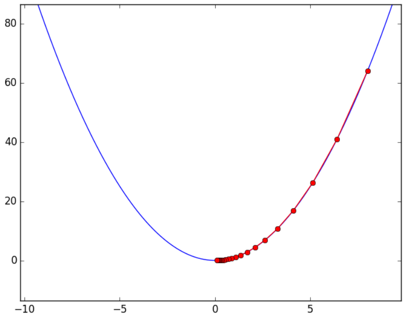


Figure: Example of gradient decsent.

Gradient decsent

- How does the cost function change when w and b changes?

Gradient decsent

- How does the cost function change when w and b changes?
- Take the gradient of the cost with respect to w and b:

$$\frac{\partial cost}{\partial w} = \frac{\partial}{\partial w} MSE(Y, \hat{Y})$$
$$\frac{\partial cost}{\partial b} = \frac{\partial}{\partial b} MSE(Y, \hat{Y})$$

Gradient decsent

- How does the cost function change when w and b changes?
- Take the gradient of the cost with respect to w and b:

$$\frac{\partial cost}{\partial w} = \frac{\partial}{\partial w}\frac{1}{2m}\sum_{i=1}^{m}(y_i - (w * x_i + b))^2$$

$$\frac{\partial cost}{\partial b} = \frac{\partial}{\partial b}\frac{1}{2m}\sum_{i=1}^{m}(y_i - (w * x_i + b))^2$$

Gradient decsent

- How does the cost function change when w and b changes?
- Take the gradient of the cost with respect to w and b:

$$\frac{\partial cost}{\partial w} = \frac{1}{m}\sum_{i=1}^{m} -x_i * (y_i - (w * x_i + b))$$

$$\frac{\partial cost}{\partial b} = \frac{1}{m}\sum_{i=1}^{m} -(y_i - (w * x_i + b))$$

Gradient decsent - Pseudo-code

Pseudo-code for steepest gradient decsent:

1. Select starting parameters.
2. Calculate the gradient for the cost function with respect to the parameters.
3. Update the parameters by taking a "step" in the opposite direction of the gradient.
4. Repeat step 2-3 until convergence.

Example of gradient decsent

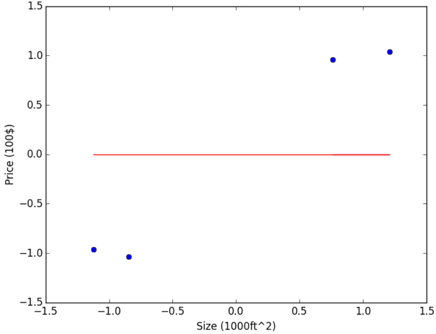


Figure: The line given the initial parameters.

Example of gradient decsent

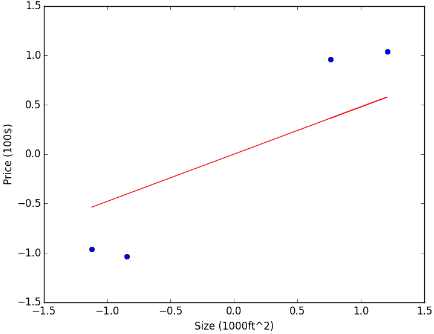


Figure: The line after 5 iterations.

Example of gradient decsent

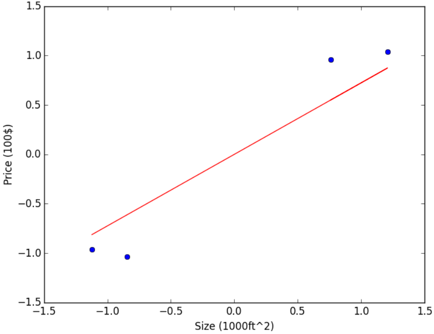


Figure: The line after 10 iterations.

Example of gradient decsent

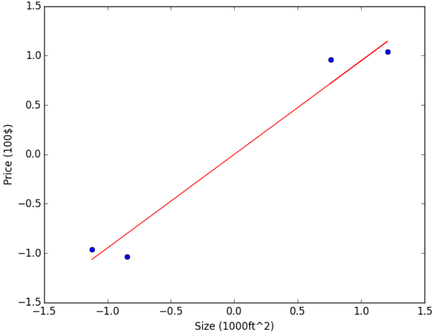


Figure: The line after 25 iterations.

Example of gradient decsent

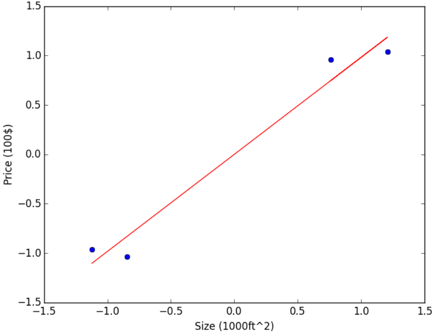


Figure: The line after 100 iterations.

Example of gradient decsent

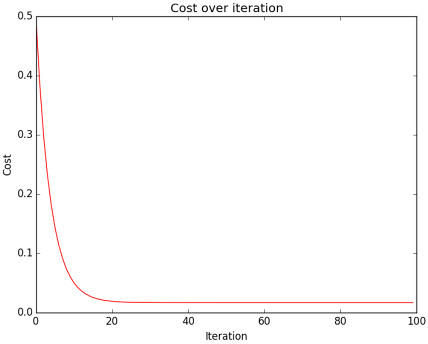


Figure: The cost over the number of iterations.

Gradient decsent - Step size

The only tricky part with gradient decsent is to set the set the "step size", (this is often called the learning rate).

- If the learning rate is to small it will take a long time to reach the optimum.
- If the learning rate is to large we will step over the minimum.

Energy landscape

For functions with few parameters and a cost function that is easy to compute, we can plot the cost over a reasonable large set of parameters.

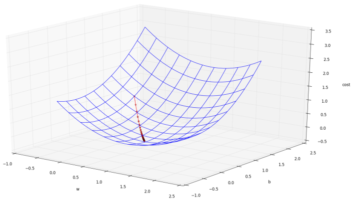


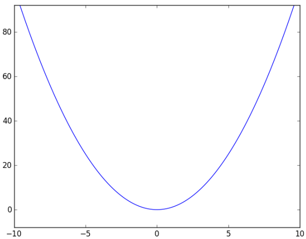
Figure: The cost function for w and b.

Convex functions

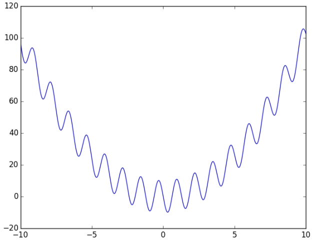
A *convex* function is a function where a line between two points on the function graph only intersects the function graph at these two points. If a function is (downward) *convex* it means that:

- The function has one single minimum.
- The gradient is always decreasing when this minimum is approach.

Convex functions and none convex functions



(a) A convex function



(b) A non convex function

Figure: An example of a convex and a non convex function

Linear regression with multiple variables

- We can also do regression with multiple input variables or multiple input variables.
- In the case of multiple input variables the equation for the prediction will be:

$$\hat{y}_i = w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_m x_{m,1} + b$$

- If we have multiple output variables the model will be a hyperplane instead of a line.

Vectorization

The presented equations for linear regression can be vectorized.
For a single input row:

$$\hat{y}_i = \begin{bmatrix} x_{i,1} & x_{i,2} & \dots & x_{i,m} \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix} + b =$$

$$= \sum_{j=0}^{m} \cdot w_j x_{i,j} + b$$

$$= x_i \cdot w + b$$

Vectorization

The presented equations for linear regression can be vectorized. For the whole dataset:

$$
\begin{aligned}
\hat{Y} &= \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix}_{n,m} \cdot \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_m \end{bmatrix}_{m,1} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}_{n,1} = \\
&= \begin{bmatrix} x_1 \cdot w + b \\ x_2 \cdot w + b \\ \vdots \\ x_n \cdot w + b \end{bmatrix}_{n,1} \\
&= X * W + B
\end{aligned}
$$

Polynomial regression

- In polynomial regression we try to predict the outcome Y with a polynomial expression.
- A polynomial expression involves only addition, subtraction, multiplication and non-negative integer exponents of variables.
- Example: $y = w_1 * x + w_2 * x^2 + w_3 * x^3 + \cdots + w_n * x^n$

Polynomial regression

- Is polynomial regression any more difficult than linear regression?

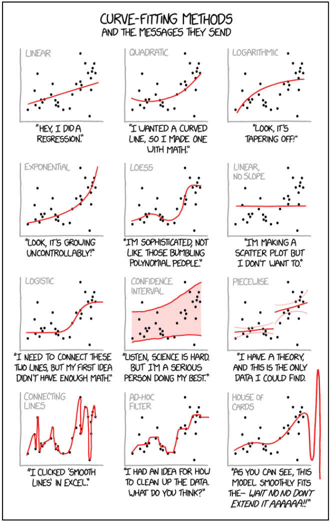
Polynomial regression

- Is polynomial regression any more difficult than linear regression?
- **No!** The same methods can be used to solve polynomial regression.

Polynomial regression

- Is polynomial regression any more difficult than linear regression?
- **No!** The same methods can be used to solve polynomial regression.
- Polynomial regression can also be transformed into a linear regression problem with multiple variables through variable transformation.
- For Example: $w_1 * x + w_2 * x^2 = w_1 * x + w_2 * z$

There are many ways to fit a line



But beware of overfitting