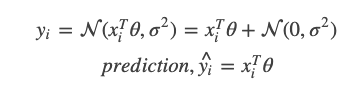
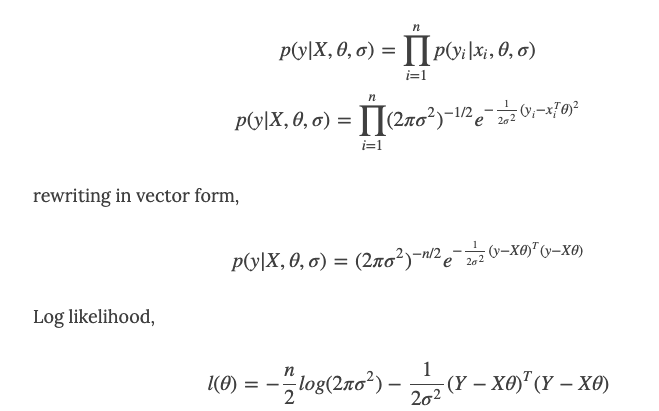
3. Write out the Maximum likelihood Estimation for linear regression. How is this related to the MSE loss for linear regression derived in the last point? Derive the relation between them.

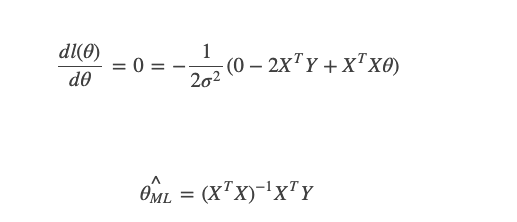


Assuiming that yi, is normal distributed with mean, xiT θ and variance, σ2. Where xi is a vector of form (xi1 = 1, x2 1)

If yi is normal distributed, the process of learning becomes the process of maximizing the product of the individual probabilities, which is equivalent to maximizing the log likelihood.



The first term is a constant and the second term is a parabola, the peak (maxima) of which can be found by equating the derivative of l(θ)l(θ) to zero. Equating first derivative to zero.



Finally, we get the linear regression equation.

Maximum likelihood estimation is a method of [estimating](https://en.wikipedia.org/wiki/Estimation_theory) the [parameters](https://en.wikipedia.org/wiki/Statistical_parameter) of a [distribution](https://en.wikipedia.org/wiki/Probability_Distribution) by [maximizing](https://en.wikipedia.org/wiki/Mathematical_optimization) a [likelihood function](https://en.wikipedia.org/wiki/Likelihood_function), so that under the assumed [statistical model](https://en.wikipedia.org/wiki/Statistical_model) the [observed data](https://en.wikipedia.org/wiki/Realization_(probability)) is most probable. Mean squared error is a measure of how good the estimator of a parameter is. The estimator can be the maximum likelihood estimator.

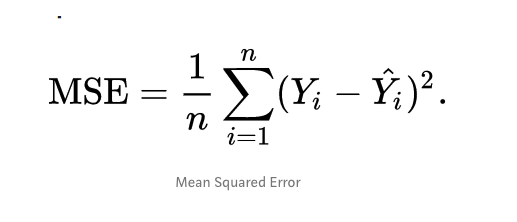
2: How is linear regression related to Pytorch and gradient descent?

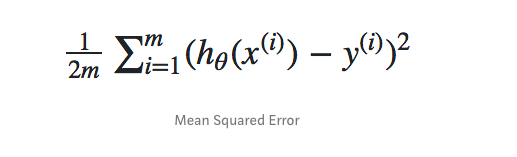
Write out MSE loss for linear regression. Could we also use this loss for classification?

Linear regression finds the best fit line to a set of points and thus minimises the cost functions. The idea is to find the paramaters for the values (x, y). In order to find these values, we need the gradient descent. For example, if we

are to calculate the price for cars, we will start with a random choice of x and y values, then we calculate the predictions (car price). Once all the predicted values are obtained, we will compare them with what was actually

observed. The difference between the observation and the prediction is called error, which is a measure of the cost function. Usually the mean squred error is calculated.





1. 1:What is Maximum Likelihood estimation (MLE) ? Can you give an example?

Maximum likelihood estimation (MLE) is a method of estimating the parameters of a distribution by maximizing a

likelihood function, so that under the assumed statistical model the observed data is most probable.

The point in the parameter space that maximizes the likelihood function is called the maximum likelihood estimate.

Example:

Suppose we have a package of seeds, each of which has a constant probability p of success of germination.

We plant n of these and count the number of those that sprout. Assume that each seed sprouts independently of

the others. How do we determine the maximum likelihood estimator of the parameter p?

We begin by noting that each seed is modeled by a Bernoulli distribution with a success of p. We let X be

either 0 or 1, and the probability mass function for a single seed is f( x ; p ) = px (1 - p)1 - x. Our sample consists of n different Xi, each of with has a Bernoulli distribution. The seeds that sprout

have Xi = 1 and the seeds that fail to sprout have Xi = 0.

The likelihood function is given by:

L ( p ) = Π pxi (1 - p)1 - xi

We see that it is possible to rewrite the likelihood function by using the laws of exponents.

L ( p ) = pΣ xi (1 - p)n - Σ xi

Next we differentiate this function with respect to p. We assume that the values for all of the Xi are known,

and hence are constant. To differentiate the likelihood function we need to use the product rule along with

the power rule:

L' ( p ) = Σ xip-1 +Σ xi (1 - p)n - Σ xi - (n - Σ xi )pΣ xi (1 - p)n-1 - Σ xi

We rewrite some of the negative exponents and have:

L' ( p ) = (1/p) Σ xipΣ xi (1 - p)n - Σ xi - 1/(1 - p) (n - Σ xi )pΣ xi (1 - p)n - Σ xi

= [(1/p) Σ xi - 1/(1 - p) (n - Σ xi)]ipΣ xi (1 - p)n - Σ xi

Now, in order to continue the process of maximization, we set this derivative equal to zero and solve for p:

0 = [(1/p) Σ xi - 1/(1 - p) (n - Σ xi)]ipΣ xi (1 - p)n - Σ xi

Since p and (1- p) are nonzero we have that

0 = (1/p) Σ xi - 1/(1 - p) (n - Σ xi).

Multiplying both sides of the equation by p(1- p) gives us:

0 = (1 - p) Σ xi - p (n - Σ xi).

We expand the right hand side and see:

0 = Σ xi - p Σ xi - p n + pΣ xi = Σ xi - p n.

Thus Σ xi = p n and (1/n)Σ xi = p. This means that the maximum likelihood estimator of p is a sample mean. More specifically this is the sample proportion of the seeds that germinated. This is perfectly in line with what intuition would tell us. In order to determine the proportion of seeds that will germinate, first consider a sample from the population of interest.

0: What is a likelihood function? Also add a formula and explain what it means.

It shows how likely particular values of statistical parameters are for a given set of observations.

Likelihood is used to summarize the data’s evidence about unknown parameters.

For discreteprobability dsitribution, the likelyhood formula is

𝑁

𝐿(𝜃;𝑋)=∏𝑓𝑖(𝑥𝑖;𝜃)

𝑖=1

=𝑓1(𝑥1;𝜃)𝑓2(𝑥2;𝜃)⋯𝑓𝑁(𝑥𝑁;𝜃)

Here 𝑓𝑁(𝑥𝑁;𝜃) is the PDF of the underlying distribution.