Obraz zawierający tekst

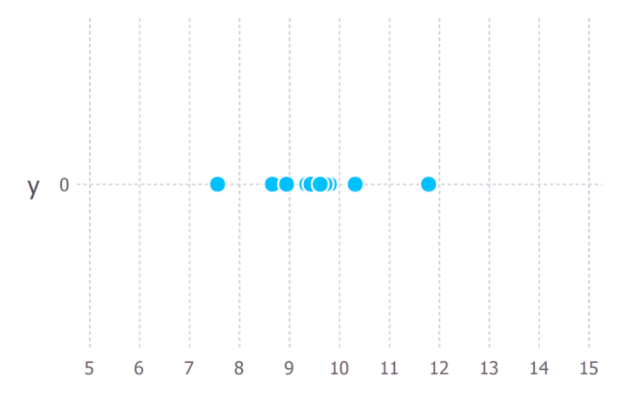
Opis wygenerowany automatycznie

Exercise 25

Overall we can say that MLE is a maximum likelihood of a function. More precisly that this is a metod of estimation that determines the values of parameters and it works in such way that the parameters values are found with intention to maximise the likelihood.

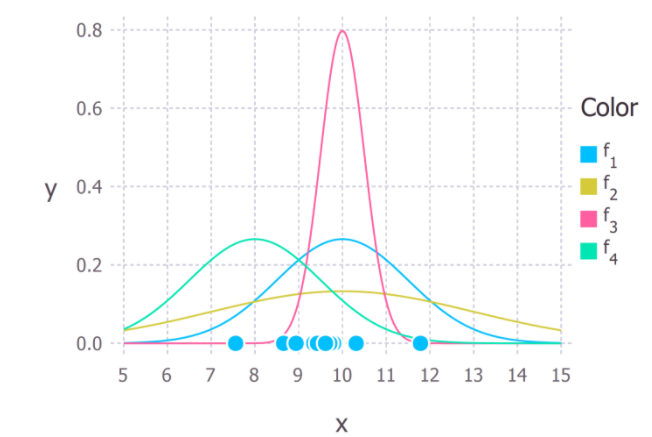
The likelihood function written as is defined for probability mass function of discrete random variable and also for continous random variable. We can also note that this is not probablity mass function or probability density function for a single random variable, but for entire joint distribution of a random variables.

A bit of explanation, let’s suppose that weh ave 10 observations of data points, for example for cat weight measurments. Each data represent the given weight of individual cat and is considered a data point.



We need to decide which model will best describe the process of generating the data. We will choose then the best distribution. For given example we will assume that the data generation process can be adequatly described by a Gaussian Normal Distribution (we can see that the Gaussian will suit good in this example becouse most oft he 10 data points are clustered in the middle with only few scattered on the left and right).

Recall that the Gaussian distribution has 2 parameters. The mean, , and the standard deviation, . Different values of these parameters result in different curves (just like with the straight lines above). Maximum likelihood estimation is a method that will find the values of and that result in the curve that best fits the data.



The 10 data points and possible Gaussian distributions from which the data were drawn. is normally distributed with mean 10 and variance 2.25 (variance is equal to the square of the standard deviation), this is also denoted and The goal of maximum likelihood is to find the parameter values that give the distribution that maximise the probability of observing the data. The true distribution from which the data were generated was , which is the blue curve in the figure above.

The Consitency of the MLE is that the MLE specific value of as well as function of can give us tiny interval estimate of function with extremly high (close to 100%) confidence level, when the value is sufficently large, for example:

The Asymptotic Efficency oft he MLE is that the MLE specific value of as well as function of beside being unbiased, has variance that converges tot he lower bound of smallest possible variances (of inbiased estimator). Which is in fact Cramer-Rao Lower Bound divided by , for example:

Exercise 26

We will use the equation for normal distribution, in this case:

We take logarithm of the likelihood function, it kmakes taking the derivative easier:

We can also state here that the likelihood function and the log of it both peak at the same values for and

Fast calc :

Then:

Combining from the first term with the remaining occurances and also all from the :

Whether we cant to calculate (our case) or , we treat one of them as constant and find where the slope of its likelihood function is 0. Then (we look what terms contains

After using chain rule and some simplifications:

Then we need to find when :

Exercise 27

**Entropy** is a therm from physics introduced into information theory, where in physics it usually refers into measure of chaos in a given system and a higher entropy means lower chaos value, in information theory entropy can be defined as the expected number of bits of information contained in an event. For example, when we are tossing a coin, the entropy of a coin is 1. It is becouse of the probability of having head or tail is equal to 0.5. The overall amount of information required to identify if it’s head or tail comes from a simple question „is it head or tail?“. If the entropy is higher than 1, it means that more infromation is needed to represent an event. In conclusion we can say that entropy increases with increase in uncertainty.

Calculating how much information is in a random variable, is same as calculating the information for the events drawn from probability distribution. In that sense, entropy is considered as average bits of information required to represent an event drawn from the probability distribution. Entropy for a random variable can be computed using the below equation:

**Cross-Entropy** is the average number of bits required to represent an event from one distrubution, when compared to another distribution. It can be explained as the average bits of information needed to identify event drawn from the estimated probability distribution , rather than true distribution . It can be view as measure of difference between two distributions.

Cross-entropy is widely used in Deep Learning as a loss function to enable the learning. In that, the true probability distribution is the label and predicted distribution is the value from the current model. The lower the number of bits computed higher the chance of approximating the predicted probability distribution with a true probability distribution. The cross-entropy loss of two distribution and can be defined as below:

In this formula we treat as a true distribution oft he event and as the estimated probability distribution of the event .

**Kullback–Leibler Divergence,** also KL divergence or relative entropy. It can be described as the difference between cross-entropy and entropy.

So in the case when we consider two distributions and , KL divergence between those two distributions can be calculated as:

KL divergence of those two distributions is then the difference between cross-entropy of those distributions, and entropy oft he distribution . In a case then we KL divergence value is 0, those two distributions are identical. The full formula of KL divergence:

Exercise 28

The negative log-likelihood is also known as the multiclass cross-entropy as they are in fact two different interpretations of the same formula.

Mathematically, it is easier to minimise the negative log-likelihood function than maximising the direct likelihood. We have then negative log-likelihood:

For multiclass classification problem, we can use cross-entropy as a loss function. As it was mentioned before, cross-entropy is the average bits of information needed to identify event drawn from the estimated probability distribution , rather than true distribution . The lower the bit higher the approximation of . When we use the cross-entropy loss function, the idea of reducing the bit using the optimisation process.

If we plug, according to previous informations:

as probability of ground truth

as probability of predicted classes

We can then identify that cross-entropy is equal to negative log-likelihood:

Exercise 29

For decision tree, Gini Index is calculated using the following formula:

Where is probability of class . This is called overall Gini Impurity and it measures the frequency at which any element of dataset will be mislabelled when it is randomly labeled.  
The minimum value oft he Gini Index is 0. This happens when the node is pure, this means that all the contained elements in the node are of one unique class. Therefore, this node will not be split again. Thus, the optimum split is chosen by the features with less Gini Index. Moreover, it gets the maximum value when the probability of the two classes are the same:

.5

As it was allready mentioned, we calculate Entropy with following formula:

Where also stay for probability of class . Entropy however is a measure of information that indicated the disorded oft he features with the target. It’s similiar tot he Gini Index, as the optimum split is chosen by the feature with less entropy (less entropy is better). It gets its maximum value when the probability of the two classes is the same and a node is pure when the entropy has its minimum value, which is 0:

The Gini Index and the Entropy have two main differences:

* Gini Index has values inside the interval whereas the interval of the Entropy is .
* Computationally, entropy is more complex since it makes use of logarithms and consequently, the calculation of the Gini Index will be faster.

Exercise 30

Let’s defiane the Gini Index fort he measurment across the K-classes:

Where represents the proportion of training observations in the -th region that are from the -th class.

The Gini index for case 1:

The Gini index for case 2:

The smaller Gini index value for case 2 indicates that this split would be favored (intuitively the split is preferable since of the nodes is pure, it only contains one class).

Same goes for cross-entropy:

Since , it follows that − . It is not hard to show that the cross-entropy will take on a value near zero if the values are all close to zero or one. Thus, like Gini Index, the cross-entropy will take on a small value if the mth node is pure. Minimizing cross-entropy is equivalent to maximizing information gain (between the input variables and class label).

Exercise 32

With the majority vote approach we will have:

0.1, 0.15, 0.2, 0.2 - GREEN

0.55, 0.6, 0.6, 0.65, 0.7, and 0.75 - RED

We classify then X as RED.

With the average probability approach, we classify X as GREEN, becouse the average of all probabilities is 0.45.