## **Maximization of the Likelihood**

In connection with the logistics task, we will consider how to calculate the likelihood for a curved coin.

So, let’s have a coin, and if you flip it, the probability of landing on heads is p: p (H) = p; this probability is unknown to us. Naturally, the probability of landing on tails is *1-p*: *p(T) = 1 – p*.

Let`s make an experiment to determine the value of p. Suppose we tossed a coin 10 times (N = 10), and it landed on heads seven times, and on tails three times.

How can we calculate the total likelihood, which is the probability of obtaining the result that we have? It is equal to the probability of falling heads out, multiplied by the probability of falling tails out:

L = p^7 (1 - p)^3. 

We can write it down since each flip is independent so that we can multiply the probabilities of each coin toss.

We need to find the maximum of L to p. More precisely, we need to find such a p that L is maximal. For this, of course, we have to use the differential calculus. We take the logarithm of the likelihood function to avoid difficulties. There is a reason to do it since the logarithm function is monotonically increasing. This means that the point at which the likelihood function reaches its maximum is the point at which the log-likelihood function also reaches its maximum at the same time.

Let’s do this. We denote the logarithm of the likelihood function by l:

l = log [p^7 (1 - p)^3] = log p^7 + log (1 - p)^3 = 7 log;p + 3 log (1 - p).

Equate the derivative to zero:

frac {partial l} {partial p} = frac {7} {p} + frac {3} {(1 - p)} * (-1) = 0.

And find the solution to p:

frac {7} {p} = frac {3} {1 - p}; ;hereof ;p = frac {7} {10} = p (H).

So we have p = 7/10, which is equal to the probability of heads falling out, that is exactly what we have expected.

Now we use a similar idea for logistic regression.

We have a probability *P(y=1*|*x)* . We can consider it as the probability of dropping heads out when we flip a coin. It is equal to the sigmoid of the product of the transposed w by x. We denote it as y:

P (y = 1|x) = sigma (w^T x).

The likelihood function for N cases is equal to

L = prod_{n=1}^{N} y_n;^{t_n} (1 - y_n)^{1 - t_n}.

It follows that if the target variable t is equal to one, then the likelihood function is yn, and if the target variable is zero, then the likelihood is 1 – yn.

It is curious that if we take the logarithm of the likelihood function, we get the equation of the cross-entropy error function, which we discussed earlier:

l = sum_{n} t_n ;log ;y_n + (1 - t_n) ;log (1 - y_n). 

The only difference is the absence of a minus sign before the formula. Thus, the maximum of the log-likelihood function is the same as the minimum of the cross-entropy error function.