Homework 1: Maximum Likelihood Estimation

Giacomo Sanna

October 27, 2019

Introduction

While discussing noisy targets, the learning of p(y|x), and the prediction of the value y of when Y is noisy, we saw that one of the approaches we can follow consists in first learning p(y|x), and then estimating y in a second, separate, step. Like in the noiseless case, when we deal with the first step, the "learning problem", we have a probabilistic Model H = $\{q_1(y|x), q_2(y|x), ...\}$ and our aim is to select from this model a hypothesis that approximates p well. Here p(y|x) denotes the true distribution and $q_i(y|x)$ is the probability distribution described by each hypothesis.

In class we focused more on using the KL divergence as the loss function that minimizes the distance between two distributions but we acknowledged that minimizing the KL divergence is equivalent to maximizing the likelihood function. My aim with this brief report is to expand the lecture material and talk about the alternative of using the maximum likelihood estimation approach, briefly touched during our class discussion. But first of all, what is the likelihood function?

The likelihood function

When our goal is the estimation of an unknown parameter, we collect data and we can describe the joint pdf of our sample as $f_{\underline{X}}(\underline{x};\theta)$. Here all we are saying is that the joint pdf is a function of the sample and of an unknown parameter. But if we frame our questions in a different way, then we can look at the joint pdf as the likelihood function $L(\underline{\theta},\underline{x})$. The subtle difference is that the likelihood function is specific to the data we have drawn and answer the question: *"What it the probability of observing this particular data, given all the possible values the unknown parameters can take?"*

Maximum likelihood estimation

Maximixing the likelihood function means finding the most likely value of the unknown parameter, given the data we observed. If we assume that each observation is *i.i.d.*, the joint pdf and the likelihood function can be written as $f_{\underline{X}}(\underline{x};\theta) = L(\underline{\theta},\underline{x}) = \prod_{i=1}^{n} f_X(x_i;\theta)$, which can be a fairly tractable form with optimization in mind.

It is common to come across pratictioners that maximize the loglikelihood $logL(\underline{\theta},\underline{x}) = l(\underline{\theta},\underline{x})$ instead of $L(\underline{\theta},\underline{x})$. Why is that? One reason for this is that the $argmax\ l(\underline{\theta},\underline{x}) = \hat{\theta}_{MLE}$ that maximizes the likelihood function also maximizes the log likelihood fuction (as the log transformation is monotonic increasing), but the additional advantage of the log transformation is that it allows us to simplify optimization calculations because it converts the product of marginal probabilities in a summation. In that case, under some regularity conditions, we can maximize the function by setting $\frac{\partial}{\partial\theta}L(\underline{\theta},\underline{x}) = 0$ or $\frac{\partial}{\partial\theta}logL(\underline{\theta},\underline{x}) = 0$ and check that the second derivative $\frac{\partial^2}{\partial\theta^2}L(\underline{\theta},\underline{x}) < 0$.

To relate these concepts to the introduction, it is easy to see that we can develop an estimate for the conditional likelihood of p(y|x) using the maximum likelihood method.
In this case, $L(\underline{\theta},\underline{y}|\underline{x}) = \prod_{i=1}^{n} f_{Y|X}(y_i|x_i;\theta)$.

Historical perspective

The concept of maximum likelihood has been around for longer than many may think. Even though it was formalized by Fisher in 1922, in a paper pointing out the limitations of finding estimators using the methods of moments, prior references and contributions in basic forms appear as early as the 1820s. As noted by Anders Hald (1999), we can trace them back to Gauss (1816), Hagen (1837) and Edgeworth (1909) and published in textbooks ranging from astronomy, geodesy and civil engineering, often calling it not "maximum likelihood" but "most probable value for the unknown".

A numerical computation - The Newton Raphson algorithm

Ofentimes, the likelihood function cannot be solved analytically using the optimization method described earlier or calculations are too complex to be carried out, and a numerical maximization approach has to be followed instead. One such approach is the Newton - Raphson algorithm. For simplicity I focus on the case when the parameter is monodimensional but multidimensional extensions to this method exist. Given a differentiable function, the Newton - Raphson method allows us to find the roots the function by iteration. It is obvious that our application is to find where the derivative of the likelihood function is equal to zero, or $L'(\theta) = 0$. If we don't know the shape of the distribution, it is useful to perform the algorithm multiple times, using different starting values, to make sure that the point of convergence is the only global maximum.

Starting from a first guess k=1, θ^k, we know that the θ_{MLE} can be approximated using the Taylor expansion $L'(\hat{\theta})L'(\theta^k) + L''(\theta^k)(\hat{\theta} - \theta^k)$, which we set equal to 0. Rearranging the equation, and solving for $\hat{\theta}$, we obtain $\hat{\theta} = \theta^k - \frac{L'(\theta^k)}{L''(\theta^k)}$. The value we obtained is used as our new guess and the process is repeated, until we reach convergence to 0, or the absolute value $|\hat{\theta}^{k+1} - \hat{\theta}^k|$ is small "enough".

Conclusions

Although there are functions in R that will perform for us all the calculations in the background (e.g."optimize"), it is important to understand the theoretical frameworks behind these formulas. An historical point of view is also useful to put in perspective the problems that we are facing today, given where we are coming from. I hope the reader will be encouraged to learn more about the various steps that many scientists have discovered and that have led us where we are now.

References

Anders Hald (1999). On the History of Maximum Likelihood in Relation to Inverse Probability and Least Squares. Statistical Science. Vol. 14, No. 2, 214-222.

Casella, G & Berger, R. (2001). Statistical Inference, Second edition. Duxbury Press

Edgeworth, F.Y. (1909). Addendum on Probable errors of frequency constants. J. Roy. Statist. Soc. 72 81-90.

Fisher, R. A. (1921). On the probable error of a coefficient of correlation deduced from a small sample. Metron 1 3-32

Hurlin, C. (2013). Maximum Likelihood Estimation - Lecture notes. University of Orleans. https://www.univ-orleans.fr/deg/masters/ESA/CH/Chapter2_MLE.pdf

Nitis Mukhopadhyay (2000). Probability And Statistical Inference. CRC Press