ML_faultyexample

November 6, 2018

1 Introduction

This document uses an example to point out an interesting dilemma in the maximum likelihood (ML) "principle". The general approach in machine learning and signal processing is to use ML as a golden compass to formulate solutions. However, the reality is that ML is associated with a number of short-comings, including problems with contamination and the famous problem of estimating variances in mixtures of Gaussians when one of the data samples falls on the mean. The former has been an interesting area of research for me on generalized divergences in the past and the latter is a famous example and a discussion for it can be found in Bishop's Machine learning and Pattern Recognition book [1]. Compared to the two examples mentioned above, the example used in the present document is much more trivial. I found this in one of Lucien Le Cam's papers [2] and couldn't resist writing it down, as it points out such an outrageous dilemma when using ML.

1.1 Example

Consider two sequences of independent random variables X_j and Y_j indexed by $j = 1, ..., n$. Let's assume that $X_j \sim N(\mu_j, \sigma^2)$ and $Y_j \sim N(\mu_j, \sigma^2)$. In other words, for each j the pair X_j and Y_j are identically distributed with the same mean and variance and the mean is a function of j.

We consider the poblem of estimating σ^2 using two separate approaches.

1.1.1 First Approach for estimating σ^2

We can define a third random variable $Z_j = X_j - Y_j$, which using the independence property of X_j and Y_j is $Z_j \sim N(0, 2\sigma^2)$. So the ML estimate of σ^2 becomes straight-forward and is equal to:

$$\hat{\sigma}^2 = \frac{1}{2n} \sum_{j=1}^{n} (X_j - Y_j)^2$$

1.1.2 Second Approach for estimation σ^2

Let's say we're not lucky enough to think of defining the auxiliary random variable Z_j and decide to routinly calculate the ML esitmate of all μ_js along with σ^2. To do so, we have the joint probability

$$f(X_j, Y_j) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(X_j - \mu_j)^2}{2\sigma^2}) \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(Y_j - \mu_j)^2}{2\sigma^2})$$

To simplify the maximization, we take its log:

$$\log f(X_j, Y_j) = -\log(2\pi\sigma^2) - \frac{(X_j - \mu_j)^2}{2\sigma^2} - \frac{(Y_j - \mu_j)^2}{2\sigma^2}$$

We have n of these terms:

$$\log(f(\{X_j\}_{j=1}^n, \{Y_j\}_{j=1}^n)) = -n\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(X_j - \mu_j)^2 + (Y_j - \mu_j)^2$$

We start with estimating each μ_j:

$$\frac{\partial}{\partial \mu_j}\log(f(\{X_j\}_{j=1}^n, \{Y_j\}_{j=1}^n)) = 0$$

$$-2(X_j - \mu_j) - 2(Y_j - \mu_j) = 0$$

$$\mu_j = \frac{X_j + Y_j}{2}$$

Now, with respect to σ^2

$$\frac{\partial}{\partial \sigma^2}\log(f(\{X_j\}_{j=1}^n, \{Y_j\}_{j=1}^n)) = 0$$

$$-n\frac{2\pi}{2\pi\sigma^2} - \frac{2}{4\sigma^4}\sum_{j=1}^{n}(X_j - \mu_j)^2 + (Y_j - \mu_j)^2 = 0$$

$$\sigma^2 = \frac{1}{2n}\sum_{j=1}^{n}(X_j - \mu_j)^2 + (Y_j - \mu_j)^2$$

Replacing $\mu_j = \frac{X_j + Y_j}{2}$

$$\sigma^2 = \frac{1}{2n}\sum_{j=1}^{n}(\frac{X_j - Y_j}{2})^2 + (\frac{X_j - Y_j}{2})^2$$

$$\hat{\sigma}^2 = \frac{1}{2n}\sum_{j=1}^{n}\frac{Z_j^2}{2}$$

So the Second approach gives an estimate that is half the size of the first estimate.

1.1.3 Discussion

This is an interesting observation, and points out one of the puzzling issues of using ML. The problem was initially raised by Neyman and Scott [3] and has been attributed to the additional degrees-of-freedom of the second approach. Therefore, the ML solutions in the second approach should be corrected to give the "best ML estimate", which is quite ironic. The question now becomes: why should there be a different solution if part the data is omitted?. More importantly, how should we know when to omit some parts of the data to obtain the best estimate. This casts a big shadow on the principle of maximum likelihood. I think a good conclusion to this example is the following self-referencial quote from Le Cam:

The main principle is that one should not believe in principles but study each problem for its own sake. Lucien Le Cam (1990)

1.1.4 References

[1] Bishop, C. (2006). Pattern Recognition and Machine Learning. Springer.

[2] Le Cam, Lucien. "Maximum likelihood: an introduction." International Statistical Review/Revue Internationale de Statistique (1990): 153-171.

[3] Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. Econometrica: Journal of the Econometric Society, 1-32.