9.6 Maximum Likelihood Estimation

What is the essence of the so called *Maximum Likelihood Estimation*?

Imagine the following situation. You are a bold spaceman who unfortunately crashed on an alien planet. Soon after your crash you see the alien sun setting and decide to call that side of the planet *west*. Minutes after the sunset you meet two bug-eyed but otherwise very friendly alien monsters - one green and one purple. The green one, keen to make small-talk, tells you that one interesting fact about this planet is that the sun sets in the west only with a 50% chance. Other 50% of the time it sets in the south. The purple one laughs and tells you that his friend is pulling your leg. "On this planet the sun sets always in the west" he says.

Which explanation should you accept as the "right" one?

The intuitive way to answer that question[4] would be: *settle for the explanation which is backed best by the observed facts*.

This simple and intuitive principle is at the core of the *Maximum Likelihood Estimation*[5]. The Maximum Likelihood Estimation is the estimation procedure which enables you to find the explanation, which is supported by the most data.

Three weeks and 21 western sunsets later on the alien planet you start to wander about the strange sense of humor of its inhabitants.

In order to bridge the gap to Computer Vision we will denote things in a slightly different manner. First we will talk more specifically about measured *data* instead about general observed facts. Second we will talk about a *model* with its *parameters* instead of an explanation. We assume that we know the correct model beforehand. Then it should be clear that a different sets of parameters for that model will provide different explanations for the data we are measuring. In the following we will also use the more formal term *hypothesis* for a general explanation. In the our context, hypothesis will mean a set of parameters for a given model. And finally, we will use *probability* in order to express how strongly the data support a hypothesis, i.e. how likely this hypothesis is.

To sum up with our new notation: The Maximum-Likelihood Estimation is the one estimation which is maximally likely, i.e. most probable, given the data.

In the following we will introduce the general form of the *Maximum Likelihood Estimation (MLE)* which is a special case of the so called *Maximum a Posteriori (MAP)* estimation. MAP in turn is an approximation of the *Bayesian* prediction. So we will first briefly explain Bayesian prediction, go on with MAP and finally end with MLE. In the end we will give examples for MLE in Computer Vision .

line fitting example should be cool

The derivation of MLE follows Russell and Norvig's AI Book [6] (Chap. 20, pp. 712-715).

In all of the following we will denote the observed data by d, and the hypotheses from the respective hypothesis space by h_i.

[4] and also the answer adopted in science

[5] Actually it is at the core of the so called *Maximum a Posteriori (MAP)* estimation, of which the MLE is but a special case.

9.6.1 Bayesian Prediction

Bayesian prediction is a more general concept than Maximum Likelihood Estimation. It would not be used directly for estimation purposes in general because of its complexity. The standard application of the Bayesian prediction is the prediction of a single variable X, given the observed data.

insert balls+urns or alien sunset example!

In order to compute the probability $P(h_i|d)$ of the hypothesis h_i given the data d, we will use the well known *Bayes' Rule*. This is useful in case that the probability of the quantity A given B $P(A|B)$ is either not given or hard to estimate but the probability $P(A|B)$ is easily accessible. In its general form, the Bayes' Rule reads like this.

Definition 9.30 (Bayes' Rule)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} = \alpha\,P(B|A)\,P(A). \tag{9.17}$$

In our case we are interested in $P(h_i|d)$, which we unfortunately can not measure directly. But on the other hand, we are in most cases able to compute $P(d|h_i)$ if we assume a certain probability density function (PDF), e.g. a Gaussian PDF, which describes the measurement errors of the data d.

$$P(h_i|d) = \frac{P(d|h_i)\,P(h_i)}{P(d)} = \alpha\,P(d|h_i)\,P(h_i) \tag{9.18}$$

The key quantities of the above equation are the so called *hypothesis prior* $P(h_i)$ and the data *likelihood* under a certain hypothesis $P(d|h_i)$.

The hypothesis prior describes the probability of a certain hypothesis which is known a priori - that is, before we actually observe any data. If we have a reason to consider a certain hypothesis more probable than the others we just assign it a higher a priori probability.

The likelihood of the data $P(d|h_i)$ describes how probable the observed data is if we assume that the hypothesis h_i is valid.

elaborate!

The idea behind the Bayesian prediction is to make a mixture of all predictions, weighted by the probability of the used hypothesis. This way, we can expect that the true hypothesis will dominate the Bayesian prediction if enough data is given. This is so because with a sufficiently large set of data, it is very improbable that the data will support the wrong hypotheses.

Definition 9.31 (Bayesian Prediction)

$$P(X|d) = \sum_i P(X|h_i)\,P(h_i|d) \tag{9.19}$$

The most important property of the Bayesian prediction is that it is optimal, i.e. that for any data set, it will be correct more often than any other prediction.

In many real world cases however, the hypothesis space is too large or even infinite and so the cost of computing the Bayesian prediction gets too large. Thus the optimal Bayesian prediction is often approximated. A very common approximation is the so called Maximum a Posteriori, which we will now introduce.

9.6.2 Maximum a Posteriori

The *Maximum a Posteriori (MAP)* prediction is based only on one hypothesis, not on the sum of all hypotheses as in Bayesian learning. The hypothesis h_{MAP} chosen for prediction is the most probable one, given the data.

Definition 9.32 (MAP Hypothesis)

$$\begin{aligned} h_{MAP} &= \arg\max_i P(h_i|d) && (9.20)\\ &= \arg\max_i \alpha\, P(d|h_i)\, P(h_i) && (9.21)\\ &= \arg\max_i P(d|h_i)\, P(h_i) && (9.22) \end{aligned}$$

This way the original task of a large (infinite) summation for the Bayesian Learning is replaced by the task of optimization for MAP. MAP is already a reasonable method for a estimation task. It estimates the hypothesis h_{MAP}. But we will see that in most practical cases we can even make a further simplification of the estimation process. This will lead us to the Maximum Likelihood Estimation (MLE).

9.6.3 Maximum Likelihood

A further simplification of the MAP can be made by the assumption that we have the same prior probability $P(h_i)$ for all hypotheses. In applications where a hypothesis is represented by a set of parameters for a given model, we do not have a reason to prefer one single set of parameters. Thus the upper assumption is valid. The formula for the computation of the most probable hypothesis now becomes even more simple.

Definition 9.33 (ML Hypothesis)

$$\begin{aligned} h_{ML} &= \arg\max_i P(h_i|d) && (9.23)\\ &= \arg\max_i \alpha P(d|h_i)\, P(h_i) && (9.24)\\ &= \arg\max_i P(d|h_i)\, P(h_i) && (9.25)\\ &= \arg\max_i P(d|h_i) && (9.26) \end{aligned}$$

This formula tells us that we can compute the optimal hypothesis h_{ML} which is the most probable one given the data d by maximizing just $P(d|h_i)$. The fact that the maximized quantity $P(d|h_i)$ is the likelihood of the data should explain the term *Maximum Likelihood*.

9.6.4 Examples