

Learning About a Coin (cont.)

Lecture Notes: Machine Learning, Wednesday, Nov 2, 2011.

Lecturer: Prof. Dr. Patrick van der Smagt

Held by: Dipl.-Inf. Christian Osendorfer

Julian Zafiris (with minor lecturer corrections)

Introduction

Recall from last lecture an important aspect of what Machine Learning is about: Given a couple of observations, by means of statistical inference reveal something about the structure of the underlying generating process. In the example of the coin, the process is given by the repeated draws in succession, whereas its structure can be described by a set of parameters $\{\theta_i\}$, representing the probability that w.l.o.g. “head” is drawn on flip i , that is

$$\theta_i \equiv P(F_i(H)) ,$$

with $F_i : \{H, T\} \rightarrow \{0, 1\}$ being a random variable, such that

$$F_i(H) = 1 \text{ and } F_i(T) = 0 ,$$

if the i -th flip results in head (H) and tail (T), respectively. According to the definition of the script ([0130-refresher-pt]), the probability density function of F_i is then referred to as $p(f_i)$. When no distinction between different flips is made, the according random variable will be referred to as F .

The Model

- In a first idea of a general model of the coin-flipping process could be to assume that the outcome of the first flip is governed by some parameter θ_1 , the outcome of the second flip by some parameter θ_2 , aso.. In this model, θ_i might somehow, by non-trivial circumstances, depend on θ_j , with $j < i$.
- As an already simplified version a graphical model incorporating the dependencies between two subsequent flips is depicted below

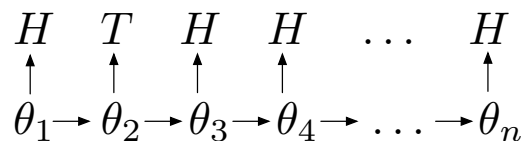


Fig. 1: Generalized model of coin-flipping process.

- However, from an intuition-based point of view and in order to tackle the problem mathematically, the depicted model can be simplified further by making the assumption that both, the coin and the circumstances under which the coin is drawn, do not change during the process, which - put into a more formal manner - result in the following two assumptions
 1. θ_i and θ_j with $i \neq j$ are *independent* of each other.
 2. $\theta_1 \equiv \theta_2 \equiv \dots \equiv \theta_n \equiv \theta$, i.e. the individual parameters θ_i are *identically distributed*.
- The 1st and 2nd assumption put together then result in the famous property of *identically and independent distributed (i.i.d.)* samples, which in this case correspond to the individual coin flips.

Given this setup, the task of this lecture was to recaptulate and introduce some of the fundamental ways to obtain an *educated guess* about how the structure of the coin-flipping process and therefore the outcome of the next coin flip could look like. As to this, the approach of *Maximum Likelihood Estimation (MLE)* was repeated and this *Maximum A Posteriori (MAP)* Estimation covered, which allow to obtain estimators for the parameter θ as underlying structural parameter of the coin-flipping process. The third approach aims to estimate the probability of the next coin flip being head directly.

Maximum Likelihood Estimation (MLE) - Repeated

- As a first approach of estimating θ , the estimator could be chosen in a way, such that the observation under the respective choice becomes “quite likely”.
- In a concrete mathematical sense, “quite” is replaced by “maximum” at this point, resulting in the name of the method, *Maximum Likelihood Estimation (MLE)*.
- The respective estimator for θ is then called Maximum Likelihood Estimator (denoted: $\hat{\theta}_{MLE}$) (Note: both the method as well as the estimator are abbreviated MLE).
- The question now is, how to determine $\hat{\theta}_{MLE}$ analytically.

- As to this, the probability for obtaining the observed sequence $HTHH \dots H$ out of all possible sequences of the same length is written down in terms of the probability $P(f_i | f_{i-1}f_{i-2} \dots f_1, \theta)$ that a specific outcome is obtained on flip i , that is

$$\begin{aligned} P(HTHH \dots H | \theta) &= P(H | THH \dots H) P(THH \dots H | \theta) \\ &= P(H | THH \dots H) P(T | HH \dots H) P(HH \dots H | \theta) \\ &= \dots \\ &= \prod_{i=1}^n P(f_i | f_{i-1}f_{i-2} \dots f_1, \theta) . \end{aligned}$$

(Note the capital P , whenever talking of probability rather than density, according to the definition in the script.)

- From the assumption of conditional independence between any two flips, one receives

$$P(f_i | f_{i-1}f_{i-2} \dots f_1, \theta) = P(f_i | \theta)$$

and therefore

$$P(HTHH \dots H | \theta) = \prod_{i=1}^n P(f_i | \theta) .$$

- This probability is called *likelihood* wrt. θ . The difference between probability and likelihood in this case is determined by the respective viewpoint or objective: by using the term likelihood, one refers to the parameters conditioned on (in this case θ). That is, functions of the form $\theta \mapsto p(\mathcal{D} | \theta)$, where p is some probability density function, are generally referred to as likelihood.

- $\hat{\theta}_{MLE}$ is then defined as

$$\hat{\theta}_{MLE} := \arg \max_{\theta} P(HTHH \dots H | \theta) .$$

- Note: as pointed out in the lecture of 2011-10-31 that for the calculation of $\hat{\theta}_{MLE}$, it might be useful to make use of the fact that

$$\begin{aligned} \arg \max_{\theta} P(HTHH \dots H | \theta) &= \arg \max_{\theta} \log P(HTHH \dots H | \theta) \\ &\equiv \arg \max_{\theta} \sum_{i=1}^n \log P(f_i | \theta) , \end{aligned}$$

which works out since log is convex and therefore does not change the point of the extrema of the corresponding argument.

- In the case of the coin the number of heads is *Binomial* distributed. For a Binomial distributed random variable, here $|H|$, the MLE is (in general) given by

$$\hat{\theta}_{MLE} = \frac{|H|}{|H| + |T|} .$$

Drawbacks of MLE

- Consider the following sequence:

HH .

- The application of the above formula yields $\hat{\theta}_{MLE} = 1$, which seems to be quite unlikely, in particular due to the fact that only a few flips of the coin were observed.
- This phenomenon (in this case as property of MLE) is called *overfitting* and expresses the general fact that the estimation of parameters of a model (in this case θ) might too strong related to the training data \mathcal{D} (in this case the sequence HH of coin flips) and lack a *generalization*. Overfitting might occur in several methods of statistical learning, when the number of observations is the same order of magnitude as the number of free parameters of the model.

Maximum APosteriori (MAP) Estimator

- To overcome the problem of overfitting in connection with MLE, a new approach has been made, in which θ is considered to be an random variable itself and in particular being non-constant. That is, the value of θ is not fixed, but depending on the inherently random sequence at hand and therefore a random quantity itself.
- Thus, one is interested in modelling the the density $p(\theta | \mathcal{D})$, which - from Bayes' Theorem - can be expressed as follows

$$p(\theta | \mathcal{D}) = \frac{p(\mathcal{D} | \theta)p(\theta)}{p(\mathcal{D})} ,$$

with

$$p(\mathcal{D}) = \int_{-\infty}^{\infty} p(\mathcal{D} | \theta') p(\theta') d\theta' .$$

(Note, that p represents the respective *density function* here. Thus, $p(\theta | \mathcal{D})$ is referred to as *conditional desity* in terms of θ .)

- Now, let $x \in [0, 1]$. The probability $P(\theta = x | \mathcal{D})$ can then be approximated by

$$P(\theta = x | \mathcal{D}) \approx \int_{x-\delta x}^{x+\delta x} p(\theta = x | \mathcal{D}) dx$$

for small positive δx .

- In the above formula $p(\theta)$ is the so-called *prior*, whereas $p(\theta | \mathcal{D})$ is referred to as *posterior*. The names result in the fact that $p(\theta)$ provides prior information about the distribution of θ , that is, before any observations are drawn.

- The idea of the *Maximum A posteriori (MAP) estimator* is now to choose the estimator for θ as the *mode* of the posterior distribution, that is

$$\begin{aligned}\hat{\theta}_{MAP} &:= \arg \max_{\theta} p(\theta | \mathcal{D}) \\ &\equiv \arg \max_{\theta} p(\mathcal{D} | \theta) p(\theta) .\end{aligned}$$

Note, that the second step is valid since $p(\mathcal{D})$ is independent of θ and thus does not play any role for the maximization.

- Taking a closer look at the above definition of $\hat{\theta}_{MAP}$ reveals that it in fact is similar to $\hat{\theta}_{MLE}$. The only but important difference is given by the introduction of the prior $p(\theta)$, which acts as *regularization* term and therefore avoids overfitting.

Choosing the prior

- The remaining question is, how to choose the prior in the approach of MAP.
- From the equations above it turns out that there are hardly any restrictions for the choice of $p(\theta)$. In fact, it does not even need to be a probability density function (i.e. integrate to 1), since

$$\begin{aligned}\hat{\theta}_{MAP} &\equiv \arg \max_{\theta} p(\mathcal{D} | \theta) p(\theta) \\ &\equiv c \cdot \arg \max_{\theta} p(\mathcal{D} | \theta) p(\theta) ,\end{aligned}$$

for any constant c and $p(\theta | \mathcal{D})$ still integrates to 1.

- Nevertheless, taking a look at the likelihood $p(\mathcal{D} | \theta)$, corresponding to a Binomial distributed random variable (the number of heads and tails respectively), allows one to make an appropriate choice of prior for the likelihood in this case, which is the density of the so-called *Beta distribution*. Assuming θ to be Beta distributed leads to a mathematical nice property of the corresponding prior, namely to *conjugate*. Informally, conjugate means that

$$\text{Posterior} \propto \text{Likelihood} \times \text{Prior}$$

will result in a density that corresponds to a Beta distribution again and therefore might be used as prior in the next estimation of a process of continuous data acquisition (see below).

- The definition of the density of the Beta distribution is based upon the of the *Beta function* (see [Wikipedia]):

$$B(a, b) \equiv \int_0^1 x^{a-1} (1-x)^{b-1} dt .$$

- For the case that a, b are positive integers, it holds that

$$B(a, b) = \frac{\Gamma(a) \Gamma(b)}{\Gamma(a+b)},$$

where Γ denotes the *gamma function*, being defined as

$$\Gamma(z) \equiv (z-1)!, \text{ for } z \in \mathbb{C}.$$

(The proof for the above connection between beta and gamma function is quite elaborate and can be found on the net.)

- The density function of the *beta distribution* is now given by

$$\begin{aligned} p(x | a, b) &\equiv \frac{1}{B(a, b)} \cdot x^{a-1} (1-x)^{b-1} \\ &\equiv \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \cdot x^{a-1} (1-x)^{b-1}. \end{aligned}$$

(From the above connection between beta and gamma function it can easily be checked that $p(x | a, b)$ integrates to unity, i.e. is indeed a probability density function.)

- The approach for choosing the prior $p(\theta)$ is now, to assume that $\theta \sim \text{Beta}(a, b)$, i.e.

$$p(\theta) = p(\theta | a, b) = \frac{\Gamma(a+b)}{\Gamma(a) \Gamma(b)} \cdot \theta^{a-1} (1-\theta)^{b-1}.$$

- a, b are called *hyperparameters* to $p(\theta | a, b)$, since they have to be determined from outside the model boundaries (see below).
- A rationale behind the choice of the density of the Beta distribution as prior can be given as follows: assume the coin has been flipped $a+b-2$ times resulting a sequence of $a-1$ heads and $b-1$ tails. Now, let $\hat{\theta}$ be some estimator for θ . The probability that the sequence at hand was generated under the assumption that θ is equal to $\hat{\theta}$ is given

$$P(a, b | \theta = \hat{\theta}) = \hat{\theta}^{a-1} (1-\hat{\theta})^{b-1}.$$

The probability that the sequence was generated by any choice of parameter is

$$P(a, b) = \int_0^1 \theta^{a-1} (1-\theta)^{b-1} d\theta.$$

Thus, the proportion

$$P(a, b | \theta = \hat{\theta}) / P(a, b)$$

(which corresponds to the density function of the Beta distribution at point $\hat{\theta}$) is *proportional* to the probability that the sequence at hand was generated using $\hat{\theta}$, i.e. $P(\theta | a, b)$.

- Hence, the problem of computing the MAP results in substituting $P(\theta | \mathcal{D})$ by $P(\theta | a, b)$, since \mathcal{D} is fully determined by a, b and the final definition of MAP with the density of a Beta distribution as prior is given

$$\hat{\theta}_{MAP} = \arg \max_{\theta} p(\mathcal{D} | \theta) p(\theta | a, b) ,$$

where \mathcal{D}, a, b and therefore $\hat{\theta}_{MAP}$ itself are changing over time (see below).

- With

$$p(\mathcal{D} | \theta) = \theta^{|H|} (1 - \theta)^{|T|} ,$$

and $|H|$ and $|T|$ denoting the number of heads and tails in \mathcal{D} , such as $p(\theta | a, b)$ as stated above, one finally obtains

$$\hat{\theta}_{MAP} = \frac{|T| + a - 1}{|H| + |T| + a + b - 2} .$$

Learning

- For the case that no data has been generated yet ($t = 0$), a meaningful choice for the hyperparameters could be to set $a = b = 1$. However, one might already include some of his experience about coins by setting for example $a = b = 2$. The meaning of this choice would be that one *expects* the same number of heads and tails (in this case $a - 1 = 1$ and $b - 1 = 1$) to be drawn out of $a + b - 2 = 2$ samples. Therefore, the respective setting reflects the assumption of the coin to be *fair* in a certain sense. Accordingly, a choice of $a = b = 10$ would still represent the same assumption of fairness, but in a much stronger sense, since the assumption of 9 heads and tails drawn out of 18 samples is statistically stronger than in case of $a = b = 2$.
- For the case that some data has already been generated ($t > 0$), the most meaningful choice would be to set the a, b according to the number of heads and tails in the data, since any personal believe is clearly dominated by an observation.
- If the process of data acquisition is described by a sequence $\mathcal{D}_1, \mathcal{D}_2, \dots$ of growing data sets and therefore increasing information, then from any new data set \mathcal{D}_i a new pair of hyperparameters a_i, b_i can be inferred, corresponding to the actual number of heads and tails in the dataset. Then the training process of MLE and MAP can be described by the following two tables, where the increasing number of lines corresponds to an increasing accuracy of the respective estimator:

Data	Likelihood	MLE
\mathcal{D}_1	$p(\mathcal{D}_1 \theta)$	$\hat{\theta}_{MLE,1}$
\mathcal{D}_2	$p(\mathcal{D}_1 \theta)$	$\hat{\theta}_{MLE,2}$
\mathcal{D}_3	$p(\mathcal{D}_1 \theta)$	$\hat{\theta}_{MLE,3}$
\vdots	\vdots	\vdots

(a)

Data	Hyperparameters	Prior	Posterior	MAP
\mathcal{D}_1	(a_1, b_1)	$p(\theta)$	$p(\theta \mathcal{D}_1)$	$\hat{\theta}_{MAP,1}$
\mathcal{D}_2	(a_2, b_2)	$p(\theta \mathcal{D}_1)$	$p(\theta \mathcal{D}_2)$	$\hat{\theta}_{MAP,2}$
\mathcal{D}_3	(a_3, b_3)	$p(\theta \mathcal{D}_3)$	$p(\theta \mathcal{D}_4)$	$\hat{\theta}_{MAP,3}$
\vdots	\vdots	\vdots	\vdots	\vdots

(b)

Tab. 1: Training process of MLE and MAP.

(Note, that in the case of MLE the likelihood and in the case of MAP the posterior is subject to maximization.)

Fully Bayesian Approach

- Recall the overall target of the estimating the parameter θ : one is interested in determining the outcome of the next coin flip as precise as possible. The estimation of θ by MLE/MAP is one way to cope with this, since the probability $P(f)$ that the next is head is fully determined by this parameter.
- Yet, another approach would be, to calculate $P(f)$ directly, by means of *marginalization* (see [0130-refresher-pt]).
- As to this, θ is considered to be a random variable, just like in the approach of MAP. The probability density $p(f)$ can be re-written as integral of the density $p(f, \theta)$ over all possible values of θ . Since the estimation of $p(f)$ is subject to the modelling approach, it also depends on the data at hand as well as the hyperparameters a, b , if the probability density according to a Beta-distribution is again used as a prior, that is $p(f) = p(f | \mathcal{D}, a, b)$.

- The marginalization is then given by

$$\begin{aligned}
 p(f | \mathcal{D}, a, b) &= \int_0^1 p(f, \theta | \mathcal{D}, a, b) d\theta \\
 &= \int_0^1 p(f | \theta, \mathcal{D}, a, b) p(\theta | \mathcal{D}, a, b) d\theta \\
 &= \int_0^1 p(f | \theta) p(\theta | \mathcal{D}, a, b) d\theta ,
 \end{aligned}$$

which is derived by application of sum and product rule (first two lines) as well as using the fact that

$$p(f | \theta) = p(f | \theta, \mathcal{D}, a, b) .$$

This holds, since the probability of any flip resulting in head is independent of the hyperparameters a, b rather than on θ , whose distribution is determined by the former.

- Plugging in the appropriate expressions for the probability density functions

$$\begin{aligned}
 p(f | \theta) &= \theta^f (1 - \theta)^{1-f} \\
 p(\theta | \mathcal{D}, a, b) &= \frac{p(\mathcal{D} | \theta) p(\theta | a, b)}{p(\mathcal{D})} \\
 p(\mathcal{D}) &= \int_0^1 p(\mathcal{D} | \theta) p(\theta | a, b) d\theta \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(|T|+a)\Gamma(|H|+b)}{\Gamma(|H|+|T|+a+b)} ,
 \end{aligned}$$

yields

$$\begin{aligned}
 p(f | \mathcal{D}, a, b) &= \frac{\Gamma(|H|+|T|+a+b)}{\Gamma(|T|+a)\Gamma(|H|+b)} \int_0^1 \theta^{|T|+a+f-1} (1-\theta)^{|H|+b-f} d\theta \\
 &= \frac{\Gamma(|H|+|T|+a+b)}{\Gamma(|T|+a)\Gamma(|H|+b)} \cdot \frac{\Gamma(f+|T|+a)\Gamma(|H|+b-f+1)}{\Gamma(|T|+a+|H|+b+1)} ,
 \end{aligned}$$

where the last step could be carried out by the general fact about the connection of beta and gamma function stated above.

Summary

In the lecture of 2011-11-02 three generic approaches have been discussed, which allow to make an educated guess of the hidden parameters of a statistical process. Maximum Likelihood and Aposteriori Estimation (MLE/MAP) aim to estimate the latent parameter θ (or a set of latent parameters), whereas a fully Bayesian approach tries to estimate the probability that a certain event occurs.