Information Theory & Statistical Physics

Lecture: Johannes Berg
Exercises: Stephan Kleinbölting

Sheet 5

Due date: 06.07.17 12:00 *To be discussed on: 12.07.17*

Website: http://www.thp.uni-koeln.de/~skleinbo/teaching/Information2017/

16 A bent coin

30+10 pts.

Tossing a possibly unfair coin coin F times, we find that it shows F_a times head and F_b times tails. Given a sequence of tosses $D = (s_1, \ldots, s_M)$ we want to predict the outcome of the next toss using Bayesian reasoning. Call the parameter of the model $p \in [0, 1]$. It describes the probability of the coin coming up as heads.

(a) *15pt* - Write down the likelihood of the data $P(D|p)$ as a function of the parameter p. Assume a flat prior $P(p) = 1$ and use Bayes' theorem to derive the posterior $P(p|D)$.

Hint: You will find the integral representation of the beta function

$$B(n, m) = \int_0^1 x^{m-1}(1-x)^{n-1}dx = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

useful.

(b) *Bonus 10pt* - Sketch the posterior distribution under different data, e.g $(F_a, F_b) \in \{(3, 7), (30, 70), (300, 700)\}$.

Hint: It is advisable to do this with a computer. Due to the factorials you probably cannot implement $P(p|D)$ directly. Implement and possibly simplify $\log P(p|D)$ instead and plot its exponential.

(c) *5pt* - Calculate the probability $P(s|D)$ of the next toss s.

Note how the Bayesian approach effortlessly incorporates the limited knowledge due to the finite sample size into the prediction!

On the other hand, it is often impossible to evaluate the necessary integrals, e.g $P(D) = \int P(p)P(D|p)dp$, explicitly. In that case we need to find ways to approximate the posterior. One such method is know as *MAP* (maximum a posteriori). In this approximation one replaces the posterior distribution by the value of its maximum.

$$p^* = \arg\max_p P(p|D) = \arg\max_p P(D|p)P(p).$$

Note that for a flat prior, the MAP estimator is the maximum likelihood estimator.

(d) *10pt* - Find the MAP estimator of p for a flat prior. Compare to the result of (b), in particular if $F = 1$ or $F = 0$.

17 Maximum a priori estimators

30 pts.

Bayesian inference requires specifying prior distributions. One might feel a little uneasy due to this freedom, and also because there is usually no canonical choice of a prior. But actually, one should rather embrace it as a blessing, because it enables us to make all our assumptions explicit. If we want to convey that nothing about the model parameters was known beforehand, we may chose an uninformative (flat) prior.

At the same time one should not overestimate the importance of the prior. Intuitively one expects that the prior becomes less important as more data becomes available. Whatever we assumed to know about the model is either phased out or confirmed by real data. We would like to examine this behavior by an example.

Let $(x_1, \dots x_M)$ be a sequence of samples from a normal distribution with mean μ and variance σ^2. Let $p(x|\mu, \sigma^2)$ denote the density.

The goal is to estimate the parameters using MAP and ML (maximum likelihood) estimators.

(a)*10pt* - Maximize the *log*-likelihood function $L(\{x_j\}|\mu, \sigma^2) = \sum_j \ln p(x_j|\mu, \sigma^2)$ and show that the ML estimates of μ and σ^2 are given by

$$\hat{\mu}_{ML} = \frac{1}{M}\sum_{j=1}^{M} x_j \qquad \hat{\sigma}^2_{ML} = \frac{1}{M}\sum_{j=1}^{M}(x_j - \hat{\mu}_{ML})^2. \tag{1}$$

Let us focus on the mean μ only. We would like to find another estimator under a more general prior. A very opportune choice is itself a Gaussian

$$p(\mu|\mu_m, \sigma_m^2) = \frac{1}{\sqrt{2\pi\sigma_m^2}} \exp\left(-\frac{1}{2}\left(\frac{\mu - \mu_m}{\sigma_m}\right)^2\right) \tag{2}$$

The reason is that now the posterior is a Gaussian itself. Priors with this property are called *conjugate priors*[1]. The parameters μ_m and σ_m of the prior are called hyper-parameters. The results of our inference will depend on them, but hopefully in a non-crucial manner.

(b)*20pt* - In analogy to (a) maximize the posterior (better: its logarithm)

$$\hat{\mu}_{MAP} = \arg\max_{\mu} \left\{ \prod_j p(x_j|\mu, \sigma^2) \cdot p(\mu|\mu_m, \sigma_m^2) \right\} \tag{3}$$

and show that

$$\hat{\mu}_{MAP} = \frac{M\sigma_m^2}{M\sigma_m^2 + \sigma^2}\left(\frac{1}{M}\sum_j x_j\right) + \frac{\sigma^2}{M\sigma_m^2 + \sigma^2}\mu_m. \tag{4}$$

Discuss the result, in particular the cases $M \to \infty$ and $\sigma_m \to \{0, \infty\}$.

Note: Only for the mean is the conjugate prior a Gaussian. If we wanted to construct an analogous estimator for the variance, we would use a prior for σ that is gamma-distributed.

[1] see Wikipedia for a table of conjugate priors for a range of different models.