Analysis of Microarray Data with Methods from Machine Learning and Network Theory

Summer Lecture 2015

Prof. Dr. A. B. Cremers
Dr. Jörg Zimmermann

Bayesian Statistics

Bayesian statistics is an alternative approach to estimation and decision problems.

The main difference to classical statistics is that model parameters like θ are now treated as random variables, too.

A probability distribution on model parameters is interpreted as a knowledge state of an observer about the true model parameters, thus quantifying the uncertainty about the true model parameters.

In Bayesian statistics it is possible to make probabilistic statements about the unknown (and unobservable) model parameters (in contrast to classical statistics).

The knowledge state of an observer can change if new data is observed. Probability theory implies that there is exactly one way to do this: Bayes' rule.

Bayes' Rule (or Bayes Theorem)

posterior probability of θ

prior probability of θ

$$p(\theta|x) = \frac{p(x|\theta)}{p(x)} \cdot p(\theta)$$

update factor

Bayes Rule states how to learn from data. It tells you the revised probability of a model θ after seeing data x.

Remember: p(x|θ) is the likelihood function, i.e. the family of stochastic models one has to choose to define the inference problem (e.g. a Bernoulli likelihood or a Gaussian likelihood)..

Bayes' Rule: prior probability

The prior probabilities for θ (the prior distribution, or just prior) are interpreted as the knowledge state of an observer before seeing data x. How to choose a prior for a specific inference problem is in general a difficult problem and is treated extensively in the literature (e.g. R. Yang, J. Berger: A Catalog of Noninformative Priors, where you can look up priors for many standard inference problems).

A main goal of prior theory is the definition of “noninformative priors”, which represent maximal ignorance of model parameters with regard to a well-defined criterion (see A. R. Syversveen: Noninformative Bayesian Priors). Bayesian inference based on such noninformative priors can be regarded as “objective” in a well-defined manner, containing no hidden assumptions or subjective knowledge.

Bayes' Rule: update factor

The update factor is the ratio of the likelihood of the observed data x given a specific model parameter θ and the probability of data x (before you have seen the data).

If the likelihood of x given θ is greater than the (unconditioned) probability of data x, than this can be regarded as positive evidence for model parameter θ, resulting in an update factor greater than 1. Correspondingly, if the likelihood of x given θ is less than the probability of data x, than is negative evidence and the update factor will be less than 1.

The unconditional probability of data x can be reduced to known quantities. It holds:

conditionalization (3. axiom)

$$p(x) = \int_{\theta \in \Theta} p(x, \theta) d\theta = \int_{\theta \in \Theta} p(x|\theta) \cdot p(\theta) d\theta$$

marginalization (see next slide)

Bayes' Rule: Marginalization

The process of reducing a joint probability distribution p(x,θ) to only one variable by ignoring (or averaging) the other variable leads to the “marginal distribution”:

$$p(x) = \int_{\theta \in \Theta} p(x, \theta) d\theta$$

Continuous case

$$p(x) = \sum_{i \in I} p(x, \theta_i)$$

Discrete case

Marginalization can be justified via the 2. axiom (sum of probabilities).

Derivation of Bayes' Rule

The core idea is to split the joint distribution p(x,θ) in two different ways, according to the third axiom of probability theory:

$$p(\theta|x) \cdot p(x) = p(\theta, x) = p(x, \theta) = p(x|\theta) \cdot p(\theta)$$

A simple transformation now leads to:

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$$

So the derivation of Bayes' Rule is mathematically trivial, but its philosophical and practical implications are deep.

Bayes-Estimator for Bernoulli-Experiment

Assume a probability distribution on parameter θ (formerly known as p):

This is called the apriori-distribution and can be viewed as representing the current (uncertain) knowledge that we have about θ.

For simplicity, let's use the uniform distribution on θ:

$$p(\theta) = 1 \qquad \theta \in [0, 1]$$

as a kind of representation that we have no knowledge of the true value (more on that later).

Bernoulli-Experiment

Uniform Prior (Laplace-Prior): $p(\theta) = 1 \qquad \theta \in [0, 1]$

Likelihood: $p(x|\theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$

(Unconditional) Probability p(x): $\int_0^1 p(x|\theta) \cdot p(\theta) d\theta = \frac{1}{n+1}$

The Beta Distribution

In order to define the posterior distribution of the Bernoulli experiment, we need the Beta Distribution. It is another important univariate continuous distribution which often occurs in applications. It depends on two parameters, α and β:

$$\alpha > 0, \beta > 0 \qquad 0 < x < 1$$

$$Be(x|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \cdot x^{\alpha-1}(1-x)^{\beta-1}$$

$$E(x) = \frac{\alpha}{\alpha+\beta} \qquad V(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Bernoulli-Experiment

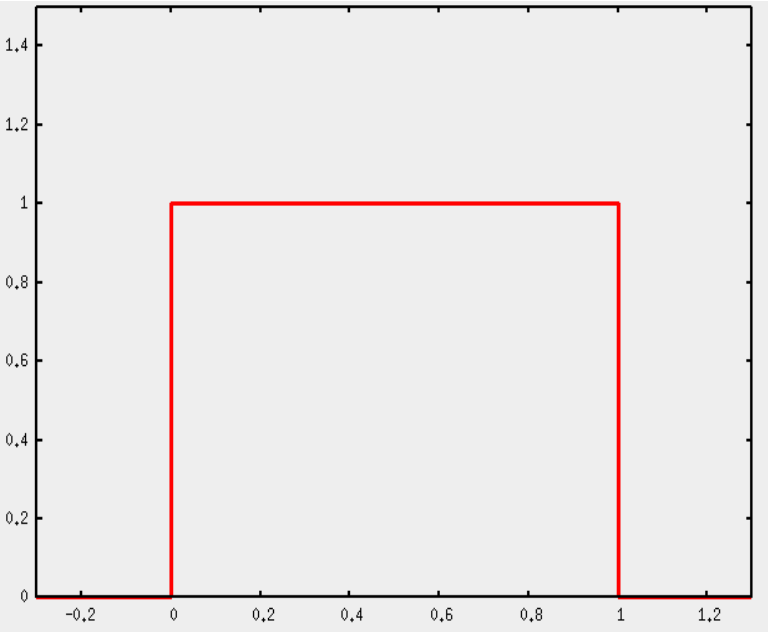
This leads to the Posterior distribution:

$$Be(\theta|x+1, n-x+1) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} \cdot \theta^x (1-\theta)^{n-x}$$

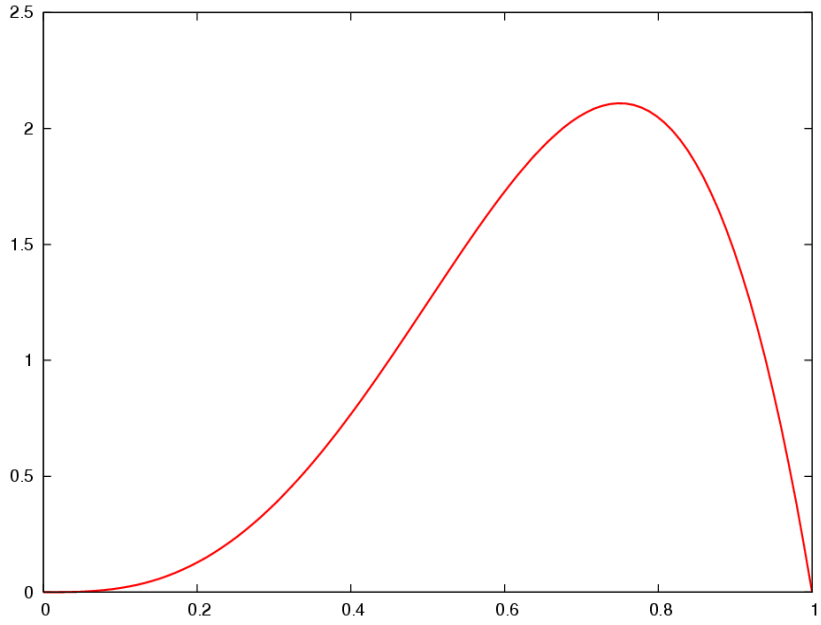
Γ(x) is the Gamma function. It is an extension of the factorial function to real numbers. If *n* is a positive integer, then: *Γ(n) = (n-1)!*

Bernoulli-Experiment

Laplace-Prior



Posterior (n=4, x=3)



Bernoulli-Experiment

How can we get an estimator from a posterior distribution?

By introducing Loss-functions. A loss-function specifies the loss incurred by assuming θ as true, while the true value θ*.

Most common loss function is Quadratic loss (but it could be any other problem-specific function):

$$Loss_{quadratic}(\theta, \theta^*) = (\theta - \theta^*)^2$$

Bernoulli-Experiment

Now we can assign every value of θ an expected loss based on the posterior-distribution:

$$E(L(\theta)) = \int_0^1 L_{quadratic}(\theta, \theta^*) \cdot p(\theta^*|x) d\theta^*$$

Now we get an estimator by the Minimum Expected Loss-Principle:

Take that θ as estimator (after having seen data x) which minimizes the expected loss (wrt. chosen loss function).

It holds that the minimizer of quadratic loss is just the expectation value of the posterior distribution.

In the our case (Bernoulli-Experiment with Laplace-Prior and quadratic loss):

$$\hat{\theta}(x) = \frac{x+1}{n+2}$$

Prior Problem

- The water/wine problem:

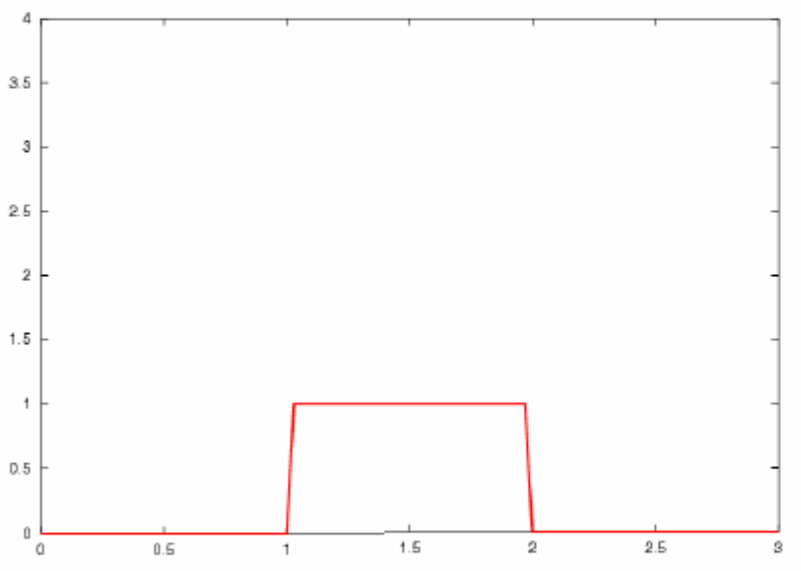
  Assume we have a glass with a mixture of water and wine, and we know:

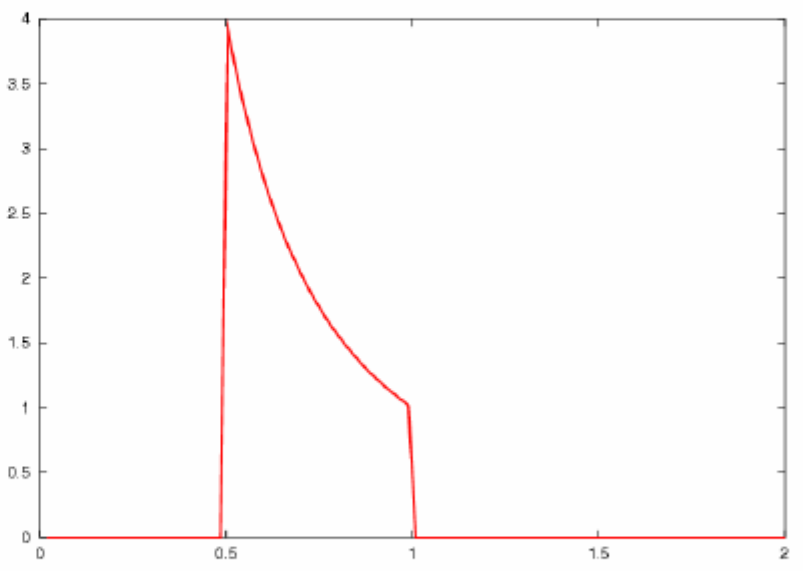
  (1) $1 < \text{water/wine} < 2$.

  Now let us assume a uniform prior on a = water/wine. But another person might choose to build the ratio b = wine/water with $\frac{1}{2} < b < 1$. How does a uniform prior for a translate into a prior for b?

Prior Problem

Uniform prior for a





translates to this prior for b!

Information Geometry

- Likelihoodfunction f(h|x) parameterizes a set of probability models.

- If there would be a notion of distance or geometry on the space of probability models, one could find a „natural parameterization", i.e., equal changes in the parameter should lead to equal changes in the respective probability models.

- A natural measure of divergence between probability distributions is the Kullback-Leibler Divergence:

$$D_{KL}(P||Q) = \int_{x \in X} p(x) \cdot \log \left(\frac{p(x)}{q(x)} \right)$$

Jeffreys' Prior

- This idea leads to the Jeffreys' prior, which uses the likelihood function as the key:

$$p(x) \sim \sqrt{det(I(x))}$$

where I(x) is the Fisher-Information Matrix

Binomial Model

- Let $B_{n,p}(x)$ be a binomial model. The Jeffreys' prior for this model is:

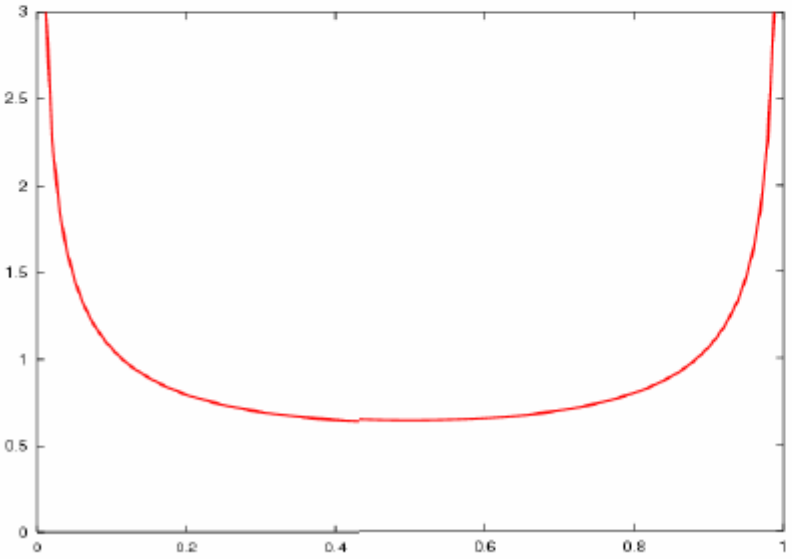
$$\frac{1}{\pi} * \frac{1}{\sqrt{p * (1-p)}}$$

After having seen n results in total and x positive results, the Posterior distribution is:

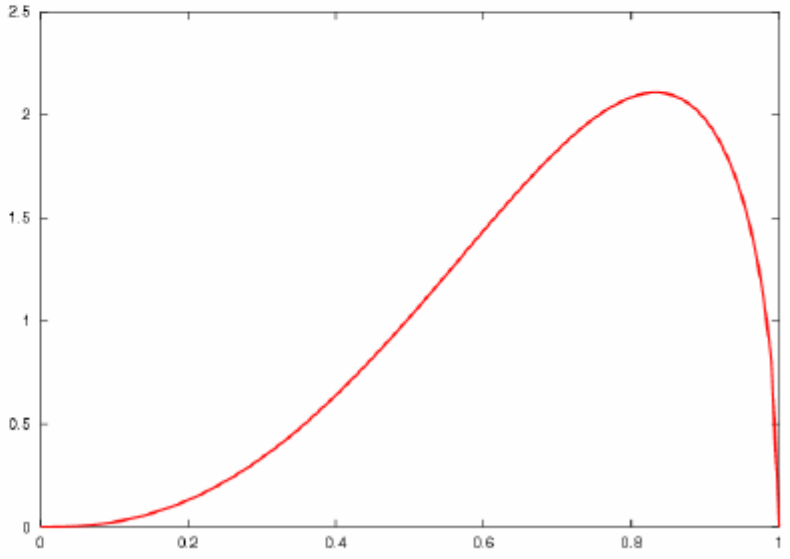
$$\frac{p^{x-1/2}(1-p)^{n-x-1/2}}{B(x+1/2, n-x+1/2)}$$

Binomial Model

Prior



Posterior (n = 4, x = 3)



Hypotheses Tests

Setting: You have a working hypothesis and you want to know if your data support your hypothesis or not.

Example: tea tasting Lady

A Lady claims she can distinguish between two cases

1. tea with milk, but pouring milk into the cup first

2. pouring first tea and then milk into the cup

Hypotheses Tests

We want statements of the kind:

$$P(X \geq t) \leq \alpha$$

where α is a significance niveau, and if X ≥ t, we can reject the "null hypothesis" on this significance niveau.