Introduction to Bayesian Statistics

Alexandra Posekany

Dec 14 2021

Probabilistic Learning

Machine Learning is based on data with one or more of the following properties:

- **Stochastic** and/or generated by a complex non-deterministic or not fully understood process
- Noisily observed
- Partially observed

Probability theory is a wide field of research focussed on expressing, modelling such uncertainties and finding appropriate data generating processes. Among those are quite prominently Probabilistic models. **Probabilistic** is directly connected with the term "probability".

Some fields of unsupervised learning where no Outcome data is availble for backpropagation and Training of an algorithm, introduced probabilities rather than deterministic decisions. Among those is naive Bayesian filtering which is applied e.g. for filtering spam emails.

Probabilistic Modelling

Probabilistic models do not simply transform information in the data into single number outputs, but allow us to include previous knowledge and provide an information about how probable any one possible value is based on the information contained in the data.

Therefore instead of fitting a deterministic model to the data, as is done by regression where the only stochastic part is left in the residuals, we will fit models where each part, the data, the parameters/model coefficients are inherently random. As random variables, we describe them through the means probabilities and their distributions. Based on these, algorithmic "learning" happens through "updating" information based on our observed data.

Introduction to Bayesian Inference

Bayes' Theorem

$$\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_\Theta \pi(\theta)f(x|\theta)d\theta}.$$

The denominator contains the **marginal distribution**, $m(x) = \int_\Theta \pi(\theta)f(x|\theta)d\theta$ which has the normalisation purpose of assuring that $\pi(\theta|x)$ is a probability distribution.
The quintessential parts of the Bayesian model are:

- the **prior distribution** $\pi(\theta)$ which expresses the uncertainty about a model parameter θ from parameter space Θ;
- the **Likelihoodfunction** $f(x|\theta)$ which transforms the information contained in the data to the model structure and evaluates its 'fittingness',
- and the resulting **posterior distribution** $\pi(\theta|x)$.

Why bother working with distributions?

Prior distributions introduce information available before the statistical analysis from any external sources into the model. This distribution basically assigns a probability to every possible value of the parameter for being a "proper" choice for the current model

The **Likelihood function** as in all statistical models weighs how well the observed data fit into the chosen model based on for a certain choice of parameter. The dependence on the choice of parameter is relevant, as it is exactly this parameter we wish to learn about in our Bayesian model.

In all our previous methods for regression with and without regularisation, cross validation or Monte Carlo simulation where we aimed for estimating a single specific value of parameters/a set of model coefficients/an area under a curve etc. Contrary to all of them, probabilistic modelling aims for obtaining a **probability** of being "appropriate" for **every single value** considered. The probability distribution of every single parameter value after combining previous information (prior) and data information (likelihood) is the **posterior distribution**.

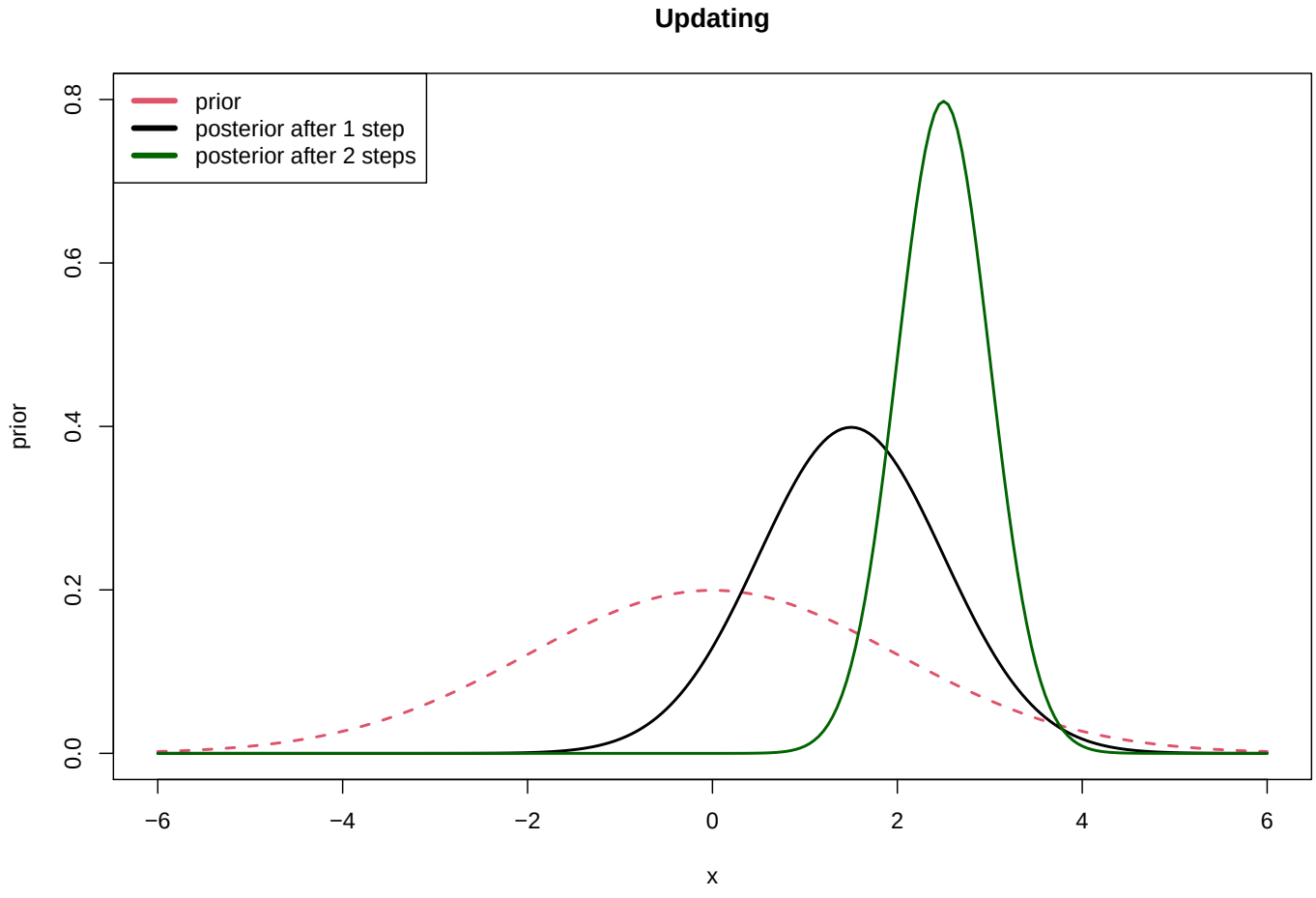
Updating Information

Maximum Likelihood estimation and Frequentist estimation can only use the whole sample at once and not include additional information.

Bayesian estimation allows for updating: taking the posterior after including the first observation as prior for including the second information results in the same posterior as including both observations at once. "**Updating**''

$$\begin{aligned} p(\theta|x_1, x_2) &\propto f(x_1, x_2|\theta)p(\theta) \\ &\propto f(x_2|\theta)f(x_1|\theta)p(\theta) \\ &\propto f(x_2|\theta)p(\theta|x_1) \end{aligned}$$

Updating


Updating
prior
posterior after 1 step
posterior after 2 steps
prior
0.0
0.2
0.4
0.6
0.8
−6
−4
−2
0
2
4
6
x

Why Bayesian methods? - Advantages

- All uncertainty is modelled as probability, probability laws obeyed.
  Complex hierarchical models are possible.

- Allows the incorporation of (prior) scientific information.

- Appropriateness of methods does not depend on having large sample sizes (asymptotics). Relevant for medical research!

- Can easily obtain inferences for any quantity of interest in an intuitively interpretable manner.
  - Direct probability interpretations.

  - Prediction is straightforward.

Why Bayesian methods? - Potential disadvantages

- Inferences depend on choice of prior and likelihood function, which may be incorrectly specified.
- Real prior distributions can be difficult to obtain in complicated models.
- In fact, bad prior information may be worse than no prior information.
- Sensitivity analysis is necessary.

The battle between Frequentist and Bayesian statistics

- **Frequentist**
  - procedure that quantifies uncertainty (e.g., p-value, confidence interval, etc.) in terms of repeating the process that generated the data many times
  - parameters are a single fixed value and unknown
  - unbiased estimation
  - makes probability statements only about the data
  - unconditional probabilities
- **Bayesian**
  - procedure that represents the uncertainty about parameters with probability distributions
  - parameters are random variables and unknown
  - biased estimation
  - makes probability statements about model parameters and the data
  - conditional probabilities

Probability

- The main difference between classical and Bayesian statistics is the concept of probability
  - from the classical point of view, probability is an "objective" concept
  - from the Bayesian point of view, probability is an "subjective" concept, as probabilities are conditional
- Both approaches have pros and cons. However, when they are both applicable, they are unlikely to give different results.

Likelihood function

Likelihood function

The **likelihood function** of a sample $x_1, \ldots, x_n$ of a random variable X which is either distributed according to the discrete distribution $\mathbb{P}[X = x|\theta]$ or according to the continuous distribution $f(x|\theta)$ is defined as

$$\begin{aligned} \mathcal{L}(\theta|(x_1, \ldots, x_n)) &= \prod_{i=1}^{n} \mathbb{P}[X = x_i|\theta] \qquad \text{or} \\ &= \prod_{i=1}^{n} f(x_i|\theta) \qquad \text{respectively.} \end{aligned}$$

The **log-likelihood function** $\ell(\theta|(x_1, \ldots, x_n))$ is the logarithmised likelihood function.

Revision of Distributions and their use as likelihood or prior

Distributions will by the building stones of Bayesian inference. We will therefore revise them - we already used them for random number generation. Now we also add the consideration for which type of scenarios which distributions are reasonable as likelihood functions and priors.

Discrete distributions will be in use for selected scenarios:

- Actual categories: univariately we start with a uniform distribution and will end up with different weights for categories a-posteriori.
- Indicators of groups: These are drawn from a Bernoulli/Binomial distribution in the dichotomous case or a multinomial distribution for more than 2 groups.
- Counts: typically we model counts without a known maximum number of counts and apply Poisson or negative binomial distributions.

Applying continuous distributions

For continuous distributions again several scenarios exist:

- **Modelling continuous observations**: The most frequently used distribution for modelling anything is the normal distribution due to the mean and standard deviation having a direct interpretation and simple to calculate estimators.

  Depending on the domain of the data and the typical shape of their values other distributions become appropriate:

  - For skewed data with positive values log-normal distributions or Gamma distributions are applied.
  - For bounded data Beta distributions may be applied.
  - For overdispersed data with heavy tails student's t distributions can be utilized.

- **Modelling residuals**: Assuming a normal distribution of residuals is common for estimating confidence bounds of linear, generalised linear and regularised regression settings, as well as other algorithms based on least squares estimation such as LDA. This distribution assumption directly transfers the models into the Bayesian setting with the option of choosing different residual distributions.

Types of prior Distributions - Informative priors

- **Natural conjugate prior distribution**
  The prior has the same "structure" as the likelihood.
  Therefore we can obtain a solution in closed form.
  Because prior, likelihood and posterior have a common "structure" we are also able to interpret the prior, data and resulting posterior information in terms of this structure and the model's parameters or distributional shape and properties.

Theorem (Pitman-Koopman Lemma)

If for large enough sample size there exists a sufficient statistic of constant dimension for a family of distributions $f(.|\theta)$*, then this family*

- **Elicitated prior**
  A prior is built 'manually' in such a way that specific 'weights' are put on specific values of the parameter based on concrete information.

Non-informative Priors

- **'noninformative prior'**
  - **Jeffrey's prior**
    Idea: invariant under diffeomorph parameter transformations
    It is determined based on the Fisher information of the likelihood function.
    $\pi_J = [det(\mathcal{I})]^{-\frac{1}{2}}$
  - **reference priors**
    These are specifically chosen priors which are linked to asymptotic properties of the posterior.
  - **Maximum Entropy prior**
    maximizes the entropy, i. e. prior uncertainty about the parameter with side conditions

Important Conjugate prior combinations for discrete likelihoods

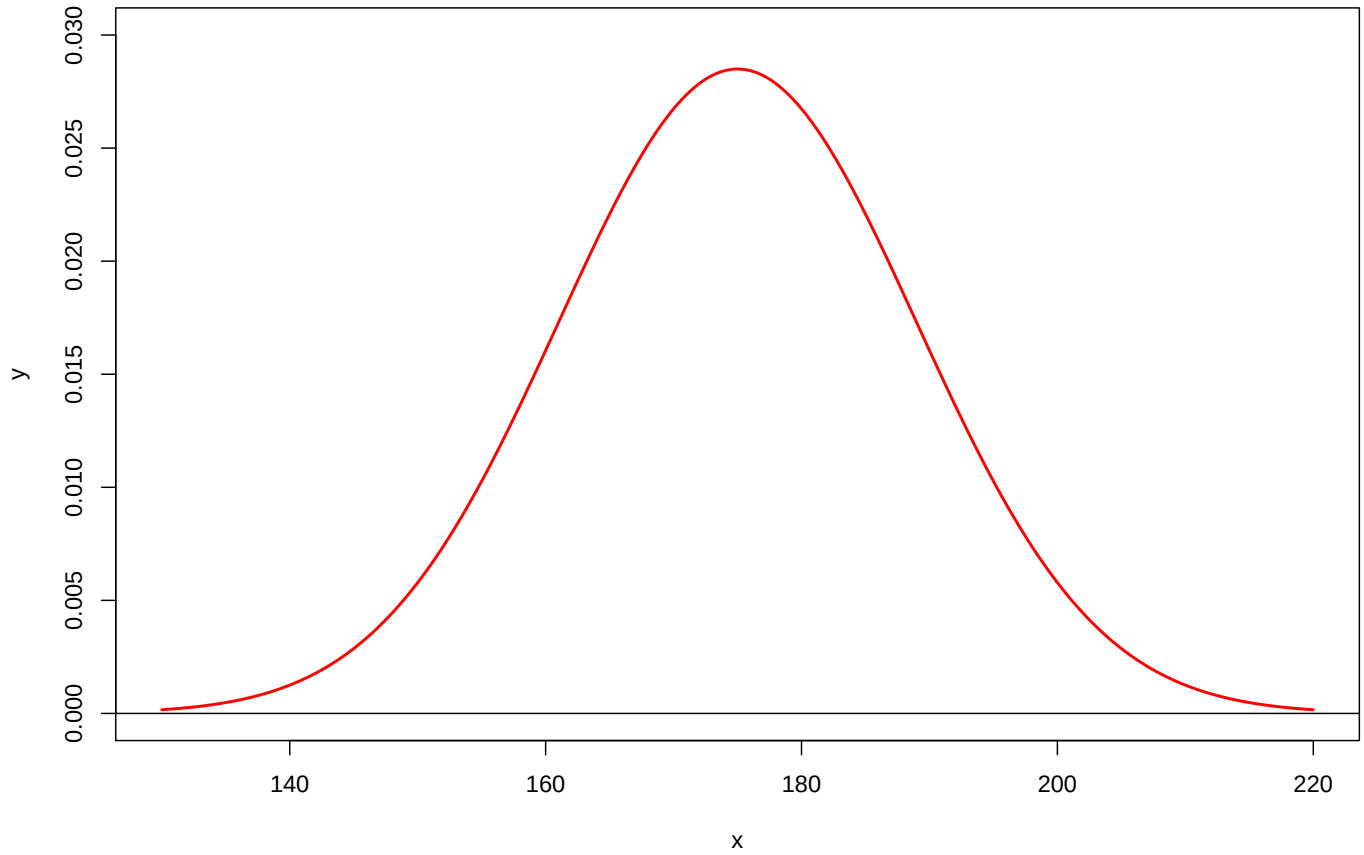
- **Binomial distribution** $y \sim Bin(n, p)$
  conjugate prior for proportion p: $p \sim Beta(a, b)$
- **Multinomial distribution**
  $y \sim Multi((n_1, \ldots, n_m), (p_1, \ldots, p_m))$
  conjugate prior for proportions $p_1, \ldots, p_m$:
  $p \sim Dir(\alpha, (\theta_1, \ldots, \theta_m))$
- **Poisson distribution** $y \sim Poi(\lambda)$
  conjugate prior for rate λ: $\lambda \sim Ga(a, b)$

Important Conjugate prior combinations for continuous likelihoods

- **Normal distribution** $y \sim N(\mu, \sigma^2)$
  - conjugate prior for mean μ: $\mu \sim N(m, s^2)$
  - conjugate prior for variance σ^2: $\sigma^2 \sim IG(a, b)$ which is equivalent to
  - conjugate prior for precision $\lambda = \frac{1}{\sigma^2}$: $\lambda \sim G(a, b)$
- **Exponential distribution** $y \sim Ex(\lambda)$
  conjugate prior for rate λ: $\lambda \sim Ga(a, b)$

Simple Example - body sizes - prior

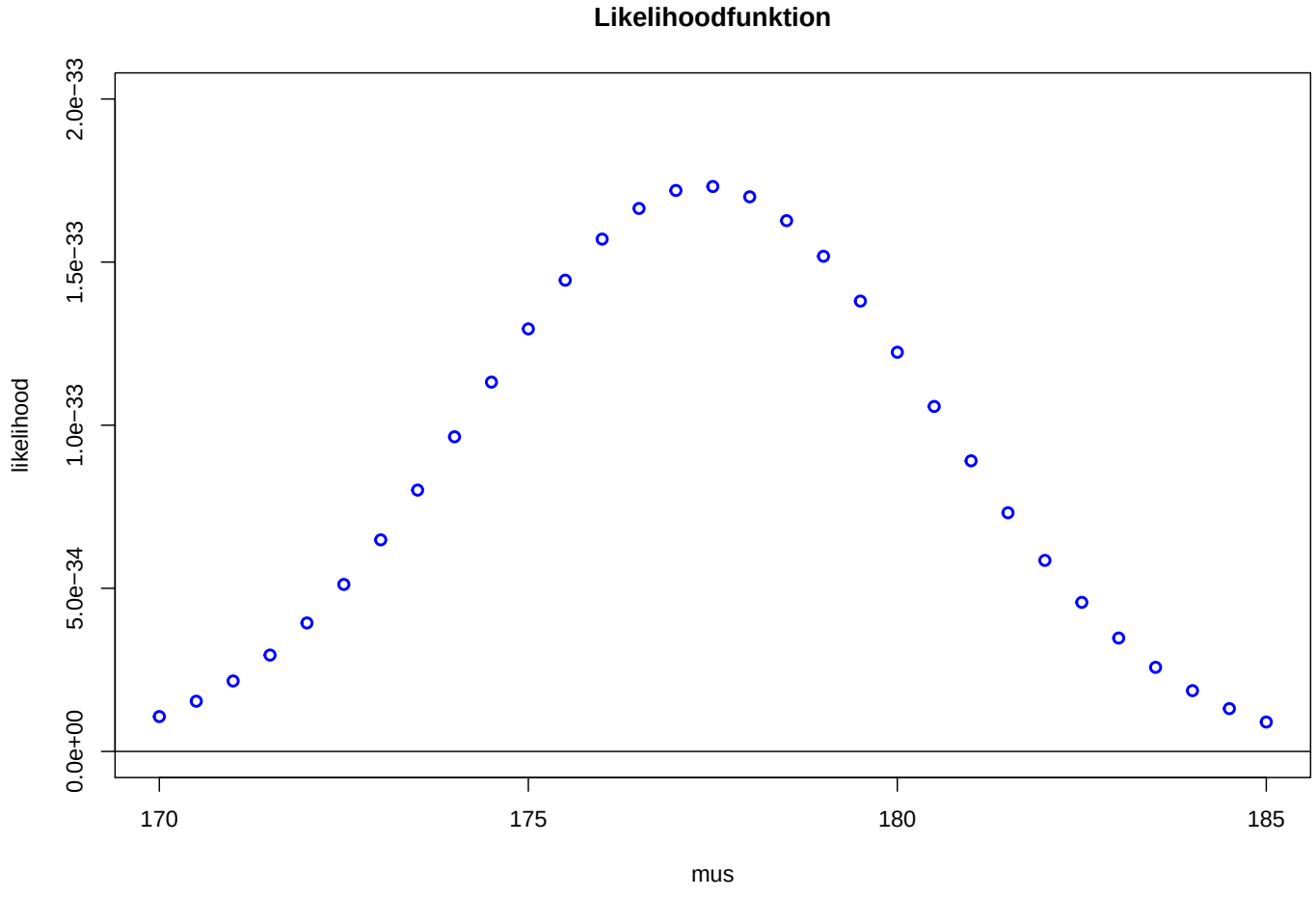
The prior information is that body sizes are normally distributed with a mean of 175 cm and a standard deviation of 14 cm.



Simple Example - body sizes - Likelihood

We start out with body sizes of persons measured in cm: 167, 169, 189, 182, 187, 173, 184, 181, 178, 182, 187, 184, 187, 187, 162, 179, 163, 159, 169, 179

This results in the following Likelihood function $\prod_{i=1}^{n} f(x_i|\mu)$ for some selected values of μ



Simple Example - body sizes - posterior

We obtain the posterior distribution by combining the **prior distribution**

$$\mu \sim \mathcal{N}(175, 14)$$

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}14} e^{-\frac{1}{2}\left(\frac{\mu-175}{14}\right)^2}$$

and the **Likelihood function**

$$L(x|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}15} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{15}\right)^2}$$

calculating

$$\pi(\mu|x) \propto L(x|\mu) \cdot \pi(\mu)$$

As the marginal distribution is simply a constant for specific data we leave it out of the calculations like all other constants which will have to be adapted in such a way that the posterior distribution is actually a probability distribution with area under the curve $= 1$.

Simple Example - body sizes - posterior

$$\pi(\mu|x) \propto \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}15} e^{-\frac{1}{2}\left(\frac{x_i-\mu}{15}\right)^2} \cdot \frac{1}{\sqrt{2\pi}14} e^{-\frac{1}{2}\left(\frac{\mu-175}{14}\right)^2}$$

is expanded to $\pi(\mu|x) \propto e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i-\mu}{15}\right)^2 - \frac{1}{2}\left(\frac{\mu-175}{14}\right)^2}$ and reordering the elements and dropping constants leads us to

$$\pi(\mu|x) \propto e^{-\frac{1}{2}\left(\frac{1}{15^2}\sum_{i=1}^{n}\left(x_i^2-2\mu x_i+\mu^2\right)-\frac{1}{14^2}\left(\mu^2-2\mu 175+175^2\right)\right)}$$

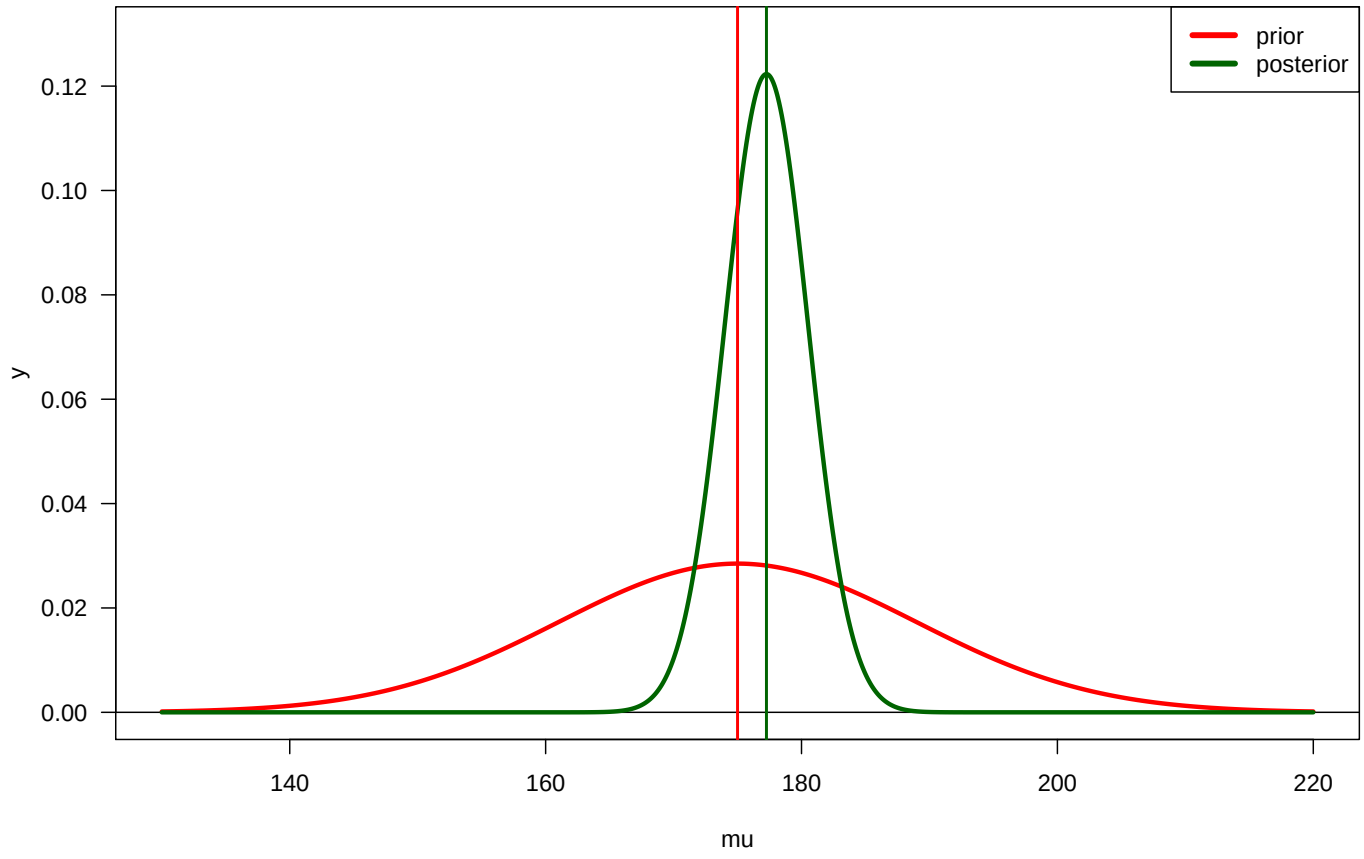
$$\pi(\mu|x) \propto e^{-\frac{1}{2}\left(\frac{1}{15^2}\mu^2 n-2\mu\frac{1}{15^2}\sum_{i=1}^{n} x_i+\frac{1}{14^2}\mu^2-2\mu\frac{1}{14^2}175\right)}$$

$$\pi(\mu|x) \propto e^{-\frac{1}{2}\left(\mu^2\left(\frac{1}{15^2}n+\frac{1}{14^2}\right)-2\mu\left(\frac{1}{15^2}\sum_{i=1}^{n} x_i-\frac{1}{14^2}175\right)\right)}$$

The trained eye can identify the structure of a normal distribution again

$$\boldsymbol{\pi(\mu|x)} \sim \mathcal{N}\left(\frac{\frac{1}{15^2}\sum_{i=1}^{n} x_i - \frac{1}{14^2}175}{\frac{1}{15^2}n+\frac{1}{14^2}}, \left(\frac{1}{15^2}n+\frac{1}{14^2}\right)^{-1}\right)$$

Updating


prior
posterior
0.12
0.10
0.08
0.06
0.04
0.02
0.00
y
140
160
180
200
220
mu

How to get information out of this posterior?

- All Inference about the parameter of interest θ is based on the posterior distribution (and therefore also on the prior).
- The information contained in the posterior distribution can be summarised in different ways as appropriate to the inference goal, e.g.
  - Means, standard deviations, medians. (point estimation)
  - Probability of exceeding a certain threshold, say θ_0, $\Pr(\theta > \theta_0 \mid \mathbf{y})$. (hypothesis tests)
  - Credibility intervals. (interval estimation)

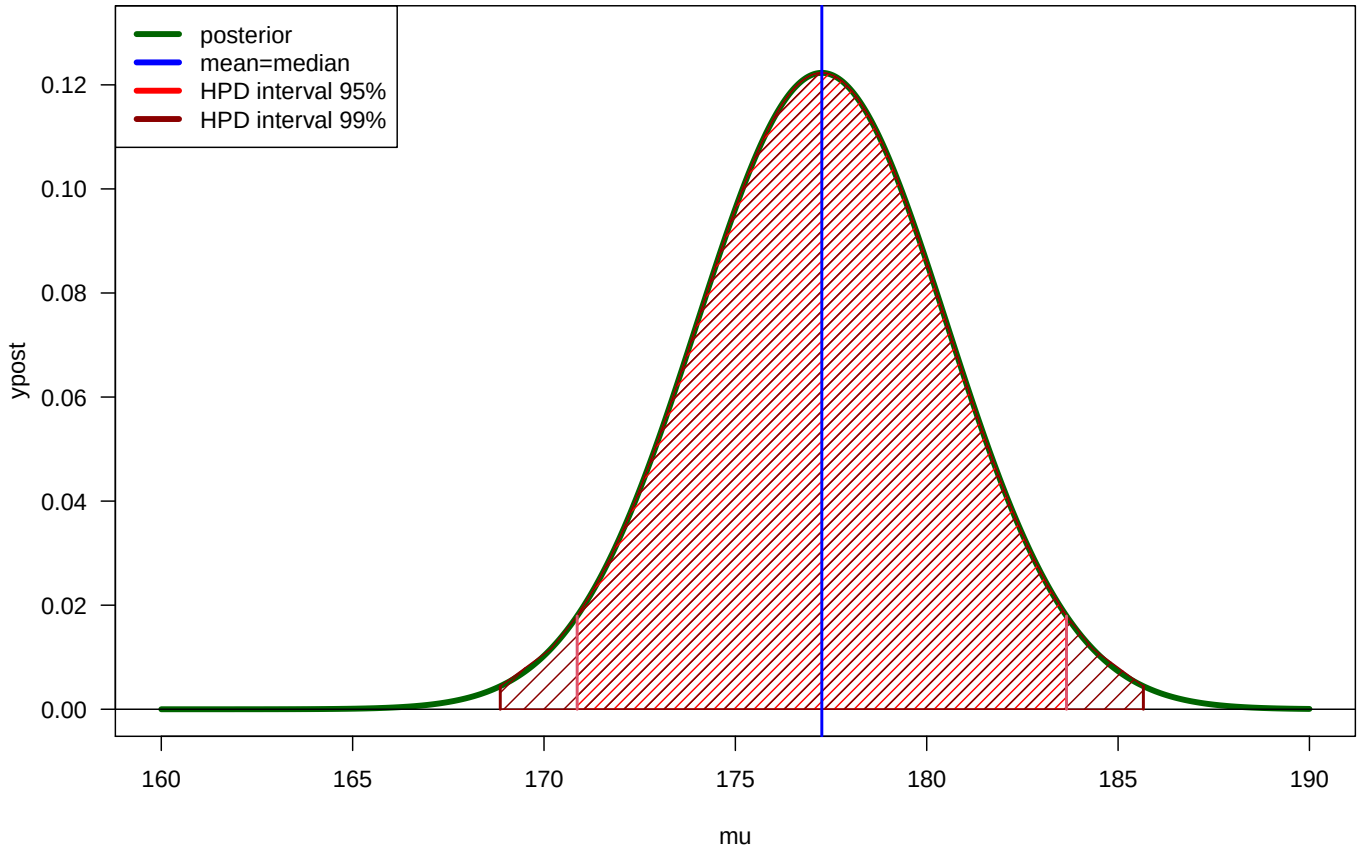
Bayesian estimators and HPD-Intervals

posterior distribution $\rightarrow$ many ways of defining point and interval estimators

The basis for this are "**Loss functions**""

- **posterior mean** $\leftrightarrow$ quadratic L_2 loss
- **posterior median** $\leftrightarrow$ absolute L_1 loss
- **posterior mode** $\leftrightarrow$ 0-1 loss
- **Highest Posterior Density Interval**
  shortest possible interval for a given coverage probability

Example body sizes



Code for graphics, posterior and HPD

```
x<-seq(160,190,by=0.05)
y<-dnorm(x,mean = 175,sd=14)  # prior
sdpost=1/sqrt(1/15^2*length(groessen)+1/14^2)
meanpost=(1/15^2*sum(groessen)+175/14^2)/(1/15^2*length(groessen)+1/14^2)
ypost<-dnorm(x,mean = meanpost,sd=sdpost) # posterior
plot(x,ypost,lwd=4,col="darkgreen",type="l",ylim=c(0,0.13),xlab="mu",las=1)
abline(h=0)
tabpost<-cbind(qnorm(seq(0.025,0.975,by=0.01),mean=meanpost,sd=sdpost),
               dnorm(qnorm(seq(0.025,0.975,by=0.01),mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost))
tabpost2<-cbind(qnorm(seq(0.005,0.995,by=0.01),mean=meanpost,sd=sdpost),
                dnorm(qnorm(seq(0.005,0.995,by=0.01),mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost))
polygon(rbind(c(qnorm(0.025,mean=meanpost,sd=sdpost),0),tabpost,c(qnorm(0.975,mean=meanpost,sd=sdpost),0)
polygon(rbind(c(qnorm(0.005,mean=meanpost,sd=sdpost),0),tabpost2,
              c(qnorm(0.995,mean=meanpost,sd=sdpost),0),c(qnorm(0.005,mean=meanpost,sd=sdpost),0)),
        col="darkred",density = 10)
abline(v=meanpost,col="blue",lwd=2)
lines(x=c(qnorm(0.025,mean=meanpost,sd=sdpost),qnorm(0.025,mean=meanpost,sd=sdpost)),
      y=c(0,dnorm(qnorm(0.025,mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost)),col=2,lwd=2)
lines(x=c(qnorm(0.975,mean=meanpost,sd=sdpost),qnorm(0.975,mean=meanpost,sd=sdpost)),
      y=c(0,dnorm(qnorm(0.975,mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost)),col=2,lwd=2)
lines(x=c(qnorm(0.005,mean=meanpost,sd=sdpost),qnorm(0.005,mean=meanpost,sd=sdpost)),
      y=c(0,dnorm(qnorm(0.005,mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost)),col="darkred",lwd=2)
lines(x=c(qnorm(0.995,mean=meanpost,sd=sdpost),qnorm(0.995,mean=meanpost,sd=sdpost)),
      y=c(0,dnorm(qnorm(0.995,mean=meanpost,sd=sdpost),mean=meanpost,sd=sdpost)),col="darkred",lwd=2)
legend("topleft",legend=c("posterior","mean=median","HPD interval 95%","HPD interval 99%"),lwd=4,
       col=c("darkgreen","blue","red","darkred"))
```

Bayes spam filtering

We consider the example of a simple spam filter which decides based on a specific word or combination of words whether an email is likely a spam email. A similar, yet somewhat more sophisticated idea is the basis for spam filters of large mailing service providers where instead of a single word or combination a "bag of words" is underlying the decision process.

For our simple example we focus on a single word. Our goal is to learn about the probability of an email which contains the word W is Spam $\mathbb{P}(S|W)$.

Bayes meets AI

We apply Bayes' theorem in its simplest form

$$\mathbb{P}(S|W) = \frac{\mathbb{P}(W|S) \cdot \mathbb{P}(S)}{\mathbb{P}(W|S) \cdot \mathbb{P}(S) + \mathbb{P}(W|S^C) \cdot \mathbb{P}(S^C)}$$

$\mathbb{P}(S)$ is the probability that a randomly selected email is spam. Statisitcs by mail service providers show rate of up to 0.8, the naive Bayesian Classifier assumes 0.5. We will later look at this rate and refine it based on received and classifed email data.

$\mathbb{P}(W|S)$ is the probability that a word occurs in a spam email - "spamicity" ("spaminess").

$\mathbb{P}(S^C)$ is the probability that a message is not spam.

$\mathbb{P}(W|S^C)$ is the probability that the word occurs in a non-spam-mail.

combining the probailities of words

If a conditional probability of being a spam mail based on all words contained in the mail $p_i = \mathbb{P}(S|W_i)$ are combined, we obtain the joint probability of an email being a spam email.

$$p = \frac{p_1 p_2 \cdots p_N}{p_1 p_2 \cdots p_N + (1-p_1)(1-p_2)\cdots(1-p_N)}$$

How to do the less naive Bayesian Inference?

Let us consider $Y \sim Bin(n, \theta)$ with n being the total number of emails received and θ the unknown proportion of spam emails. Here Y counts the number of emails identified as spam emails.

We note that the likelihood

$$L(\theta|y) = \binom{n}{y}\theta^y(1-\theta)^{n-y} \propto \theta^y(1-\theta)^{n-y} = \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

which is the kernel of a $Beta(\alpha, \beta)$ distribution with $\alpha = y + 1$ and $\beta = n - y + 1$.

Therefore, the parameters a and b refer to the number of spam emails and non-spam emails respectively regularised by adding 1.

Assume that 50 emails were received and out of those, 30 were spam. Then, y=30 and n-y=20.

How did that happen? - Conjugate priors

We will show that

$$\theta \sim Beta(a, b) \quad prior \qquad ----->$$
$$\theta|y \sim Beta(a + y, b + n - y) \quad posterior$$

The trick here is that the prior distribution and the likelihood distribution have the same basic structure, as we have already seen in the normal distribution example for human sizes. Further assume that we start with **prior** information on the rates of spam emails just like the naive Bayesian filter and assume a 50% probability, then we can encode this for example as $Beta(5, 5)$ **prior** for θ. The parameters a and b of the Beta distribution mean that out of 10 "previous" emails a=5 have been identified as 'spam' and b=5 have been identified as 'not spam'.

Under these circumstances, given the observed sample, one could learn about the proportion of spam emails that it follows a Beta distribution $\theta|y \sim Beta(35, 25)$.

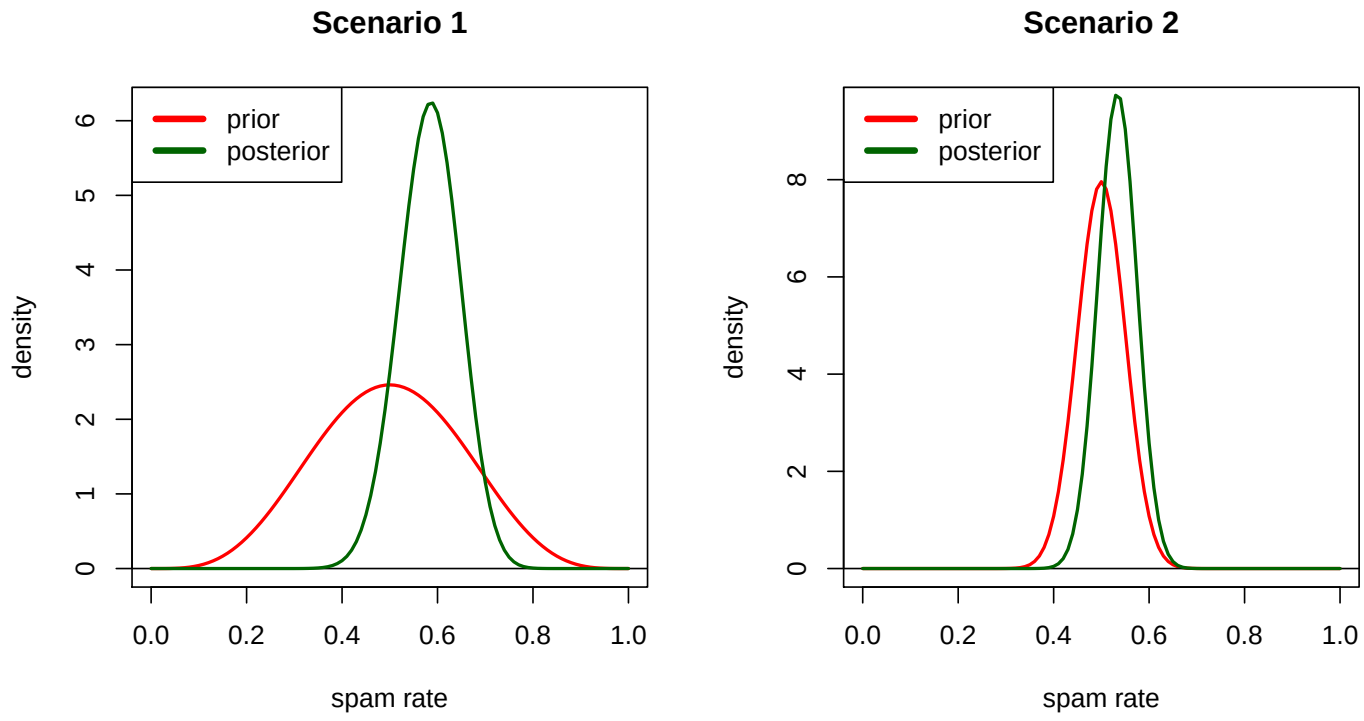
Changing the prior - Getting informaiton into the model

If we had used a different prior distribution which carried the same information such as a $Beta(50, 50)$ **prior** for θ. Then parameters a and b of the Beta distribution mean that out of 100 "previous" emails a=50 have been identified as 'spam' and b=50 have been identified as 'not spam'.

Combining this prior information with the observed sample, one could learn about the proportion of spam emails that it follows a Beta distribution $\theta|y \sim Beta(80, 70)$.

Obviously, we have much more prior evidence and therefore the prior observations have more influence compared to the data. This effect is called **informative** and such a prior is an **informative prior**.

Email example - Updating the spam rate



R Code for scenario

```
xrate<-seq(0,1,by=0.01)
prior1<-dbeta(xrate,shape1 = 5,shape2 = 5)
posterior1<-dbeta(xrate,shape1 = 35,shape2 = 25)
prior2<-dbeta(xrate,shape1 = 50,shape2 = 50)
posterior2<-dbeta(xrate,shape1 = 80,shape2 = 70)
par(mfrow=c(1,2))
plot(xrate,prior1,xlab="spam rate",ylab="density",main="Scenario 1",
     type="l",col="red",ylim=c(0,6.2),lwd=2)
abline(h=0)
lines(xrate,posterior1,col="darkgreen",lwd=2)
legend("topleft",legend=c("prior","posterior"),col=c("red","darkgreen"),lwd=4)
plot(xrate,prior2,xlab="spam rate",ylab="density",main="Scenario 2",
     type="l",col="red",ylim=c(0,9.52),lwd=2)
abline(h=0)
lines(xrate,posterior2,col="darkgreen",lwd=2)
legend("topleft",legend=c("prior","posterior"),col=c("red","darkgreen"),lwd=4)
```

Maximum Likelihood vs. Bayesian Estimation

Maximum Likelihood Estimation

- **Likelihood function** $L(\theta) = p(D|\theta)$ is basis for estimation
- **Point estimator**: Maximum-Likelihood estimator is an extreme value without a probability or distribution!
- **Confidence intervals** derived from ML estimator based on additional distribution assumptions for estimator or test statistics
- **Hypothesis tests**: asymmetry of hypotheses; p-value evaluates the type I error of incorrectly rejecting the null hypothesis

Bayesian Estimation

- **posterior distribution** derived from Likelihood function $L(\theta) = p(D|\theta)$ and prior information
- **Point estimator**: derived from posterior distribution dependent on a loss function
- **HPD Intervals** are the shortest intervals covering a given range of the parameter with a given probability which is naturally derived from the posterior distribution
- **Hypothesis tests**: symmetry of hypotheses; Bayes Faktor evaluates "how much more/less probable the alternative hypothesis is compared to the null hypothesis"