Some Introductory Remarks on Bayesian Inference

Mikio L. Braun

Seminar on Bayes Theory, TU Berlin, SS07

Overview

1. Introduction
   - Bayes Rule
2. Conjugacy
   - Example: Bernoulli distribution
   - Example: Gaussian random variables
   - Exponential Families
3. Philosophical Background
   - On Interpretations of Probability Theory
   - Bayesianism vs. Frequentism in Terms of Modelling

Bayes Rule

Ingredients:

- Model M
- Data D
- Prior $P(M)$
- Conditional Probability $P(D|M)$
- Bayes Rule $P(M|D) = \dfrac{P(D|M)P(M)}{P(D)} = \dfrac{P(D|M)P(M)}{\int P(D|M)P(M)dM}$

Bayes Rule

Ingredients:

- Model M
- Data D
- Prior $P(M)$
- Conditional Probability $P(D|M)$
- Bayes Rule $P(M|D) = \frac{P(D|M)P(M)}{P(D)} = \frac{P(D|M)P(M)}{\int P(D|M)P(M)dM}$

$\rightsquigarrow$ multiple data points by independence assumption:

$$P(D_1, \ldots, D_n|M) = \prod_{i=1}^{n} P(D_i|M).$$

An example

- $M \in \{\text{evolution, intelligent design, the Matrix}\}$.
- $D \in \{\text{fossils, the bible, déjà vu}\}$

An example

- $M \in \{\text{evolution, intelligent design, the Matrix}\}$.
- $D \in \{\text{fossils, the bible, déjà vu}\}$

So what is $P(\text{evolution}|\text{the bible})$?

An example

- $M \in \{$evolution, intelligent design, the Matrix$\}$.
- $D \in \{$fossils, the bible, déjà vu$\}$

So what is $P(\text{evolution}|\text{the bible})$?

Or $P(\text{intelligent design}|\text{fossils})$?

An example

- $M \in \{\text{evolution, intelligent design, the Matrix}\}$.
- $D \in \{\text{fossils, the bible, déjà vu}\}$

So what is $P(\text{evolution}|\text{the bible})$?

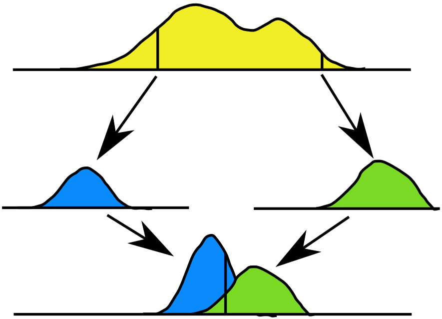
Or $P(\text{intelligent design}|\text{fossils})$?

Or $P(\text{the Matrix}|\text{fossils})$ vs. $P(\text{the Matrix}|\text{many déjà vus})$?

Alternatively...

$\int P(D|M)P(M)dM$ looks like one step in a Markov chain.

$\rightsquigarrow$ models are weighted according to their contribution to D



Why choose different priors?

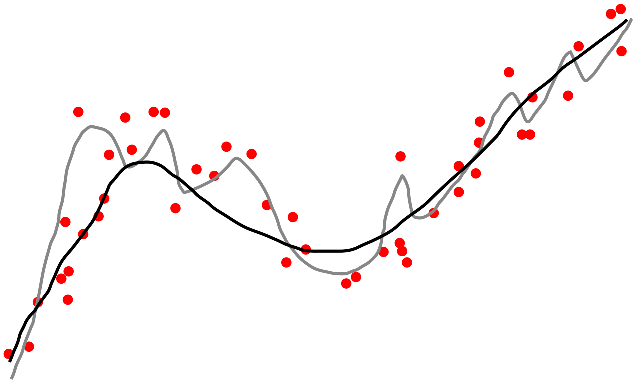
Shouldn't we be open to all possibilities?

And be free from prejudice?

Why choose different priors?

Shouldn't we be open to all possibilities?

And be free from prejudice?



Conjugacy

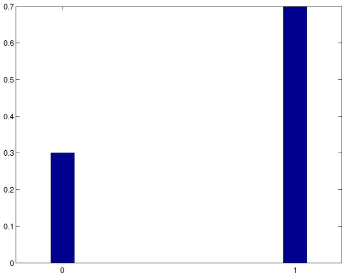
Depending on the probabilities involved, computing Bayes formular requires one integration which may be infeasible.

However, for many probability distributions, it is possible to choose a prior such that

- Bayes rule can be applied exactly,
- the posterior has the same functional form as the prior.

This is called *conjugacy*, and the prior is called the *conjugate prior*.

Bernoulli distribution



$$P(x = 1|\mu) = \mu \qquad P(x = 0|\mu) = 1 - \mu$$

$$\rightsquigarrow P(x|\mu) = \mu^x (1-\mu)^{1-x}$$

Guessing the prior

$$P(M|D) = \frac{PD|MP(M)}{P(D)} \propto P(D|M)P(M).$$

Approach: Forget the normalization, look for a $P(M|\theta)$ such that

$$P(D|M)P(M|\theta) \propto P(M|\theta')$$

For example: $\mu^a(1-\mu)^b$:

$$\mu^x(1-\mu)^{1-x}\mu^a(1-\mu)^b = \mu^{x+a}(1-\mu)^{b+1-x}$$

Finding the normalization

Fortunately, this has already been carried out and the correct prior distributions can be found (somewhere)...

Beta distribution:

$$\mathrm{Beta}(\mu|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1}.$$

Expectation: $a/a+b$

($\Gamma(n)$ interpolates the factorial, $\Gamma(n) = (n-1)!$).

Interpreting the prior: Pseudo-counts

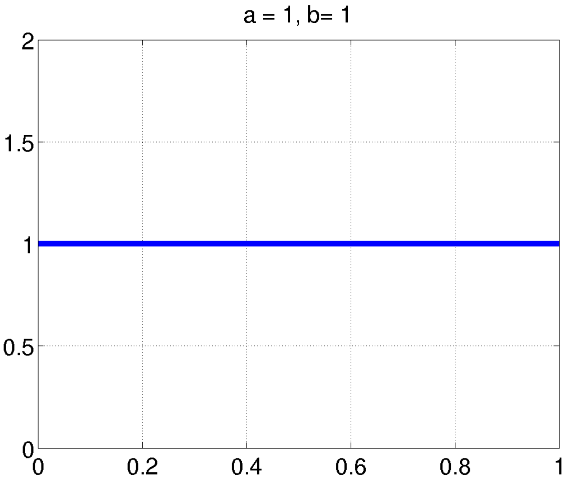
a, b are "pseudo-counts":

$$\mu^x(1-\mu)^{1-x}\mu^{a-1}(1-\mu)^{b-1} = \mu^{a+x-1}(1-\mu)^{b+(1-x)-1}$$

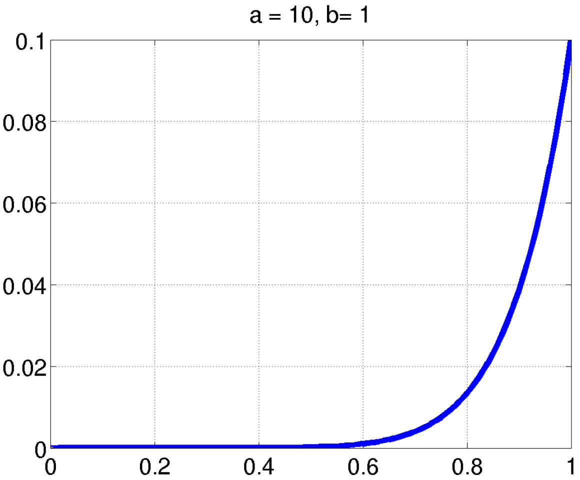
Therefore:

$$\begin{aligned} a &\to a+1 \quad \text{when} \quad x = 1 \\ b &\to b+1 \quad \text{when} \quad x = 0 \end{aligned}$$

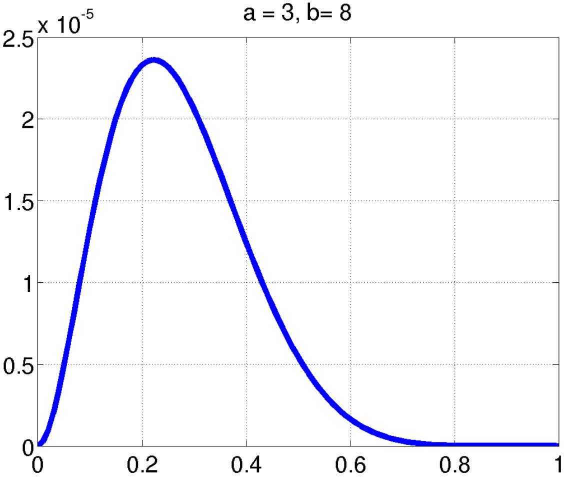
The Beta-Distribution



The Beta-Distribution



The Beta-Distribution



In a similar manner...

Binomial distribution:

$$\binom{n}{k}\mu^k(a-\mu)^{n-k} \quad \Longrightarrow \quad \mathrm{Beta}(\mu|a,b).$$

Multionimal distribution:

$$\binom{n}{n_1 n_2 \dots n_K}\prod_{k=1}^{K}\mu_k^{n_k} \quad \Longrightarrow \quad \text{Dirichlet distribution}$$

$$\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\dots\Gamma(\alpha_K)}\prod_{k=1}^{K}\mu_k^{\alpha-1}.$$

The Gaussian

The Gaussian distribution:

$$p(x|\mu, \sigma^2) \propto e^{-(x-\mu)^2/2\sigma^2}$$

Let us guess the correct prior for μ: it should be a quadratic function x:

$$p(\mu|a, b) \propto e^{-a(x-b)^2}$$

... which is basically again a Gaussian distribution.

Posterior for n data points:

$$\mu_n = \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0 + \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2}\mu_{ML}$$
$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}.$$

The Gaussian

Prior for σ^2: Rewrite $\lambda = 1/\sigma^2$, then

$$p(x|\mu, \lambda) \propto \lambda^{1/2} e^{-\lambda(x-\mu)^2/2}.$$

Guessing the prior:

$$\rightsquigarrow \qquad \lambda^b e^{-b\lambda}$$

This leads to the Gamma-distribution:

$$\Gamma(\lambda|a, b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} e^{-b\lambda}.$$

Posterior for n data points:

$$a_N = a_0 + \frac{n}{2}$$
$$b_N = b_0 + \frac{n}{2}\sigma^2_{ML}.$$

Exponential Families

In general, conjugate priors exist for distributions from the *exponential family*.

$$p(x|\theta) = h(x)e^{\langle \theta, x \rangle - \psi(\theta)}.$$

Guessing the prior...

$$p(\theta|a, b) \propto e^{\langle \theta, a \rangle - b\psi(\theta)}.$$

Because:

$$e^{\langle \theta, x \rangle - \psi(\theta)} e^{\langle \theta, a \rangle - b\psi(\theta)} = e^{\langle \theta, a+x \rangle - (b+1)\psi(\theta)}$$

Exponential Families (cont'd)

Likelihood	Prior/Posterior
Gaussian (mean)	Gaussian
Gaussian (variance)	Gamma
Poisson	Gamma
Gamma	Gamma
Binomial	Beta
Negative Binomial	Beta
Multinomial	Dirichlet

Bayesianism vs. Frequentism

Frequentism: Maximum-likelihood, Hypothesis Testing, Unbiased Estimates, Support Vector Machines, etc.

Bayesianism: Bayesian estimation, Gaussian Processes, Belief Networks, Factor Graphs, etc.

Irreconcilable Differences or Two Sides of the Same Coin?

Interpretations of Probability Theory

P-Theo does not provide any linkage to the world.

It's basically just this:

$$P(\emptyset) = 0, \quad P(\overline{A}) = 1 - P(A), \quad P(\dot{\bigcup}_i A_i) = \sum_i P(A_i)$$

$$E(X), \quad P(A \cap B) = P(A)P(B) \quad P(A|B) = P(A \cap B)/P(B)$$

From this, everything else is derived, including laws of large numbers, etc.

Bayesianism vs. Frequentism

Use P-Theo to...

Frequentism: ... model independent repeatable experiments

$\rightsquigarrow$ if I sum up many realizations, they will be close to the expectation.

Bayesianism: ... model computations on belief distributions

$\rightsquigarrow$ if I model the data correctly, my belief will be updated accordingly.

Bayesianism vs. Frequentism

Use P-Theo to...

Frequentism: ... model independent repeatable experiments

$\rightsquigarrow$ if I sum up many realizations, they will be close to the expectation.

Bayesianism: ... model computations on belief distributions

$\rightsquigarrow$ if I model the data correctly, my belief will be updated accordingly.

Compete only in terms of real-world performance, but not over what is the correct way to use P-Theo.

Bayesianism vs. Frequentism

Use P-Theo to...

Frequentism: ... model independent repeatable experiments

$\rightsquigarrow$ if I sum up many realizations, they will be close to the expectation.

Bayesianism: ... model computations on belief distributions

$\rightsquigarrow$ if I model the data correctly, my belief will be updated accordingly.

Compete only in terms of real-world performance, but not over what is the correct way to use P-Theo.

Except for: Bayesian approaches result in posterior distribution while Frequentist methods usually just return a single solution.

B vs F—in terms of modelling

Machine learning methods can roughly be decomposed in terms of

- Modelling (what is it I want to learn)
- Regularization (make sure we don't overfit)
- Inference (actually compute the solution given the data)

B vs F—in terms of modelling

Machine learning methods can roughly be decomposed in terms of

- Modelling (what is it I want to learn)
- Regularization (make sure we don't overfit)
- Inference (actually compute the solution given the data)

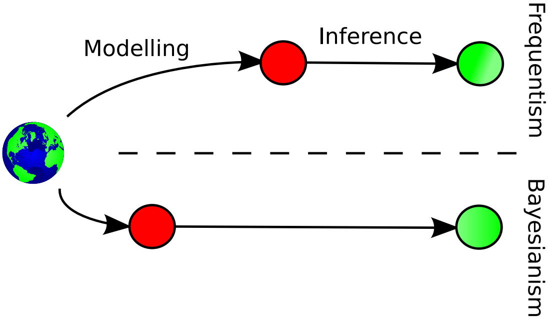
And this holds for both:

	Bayesians	Frequentist
modelling	$P(D\|M)$	loss function
regularization	$P(M)$	regularization
inference	Bayes-rule	optimization

B vs F—different kinds of uncertainty

Frequentism: modelling is kind of inexact, but at least inference is exact.

Bayesianism: modelling is clear, but inference is kind of inexact.



B vs F—irreconcilable differences?

Maybe, since tools are very different:

Frequentist: know which computations on samples converge/concentrate, optimization theory (convex optimization, gradient descent, interior point methods...), etc.

Bayesians: probability distributions, which priors make sensible computations, sampling methods like MCMC, approximation methods.

B vs F—irreconcilable differences?

Maybe, since tools are very different:

Frequentist: know which computations on samples converge/concentrate, optimization theory (convex optimization, gradient descent, interior point methods...), etc.

Bayesians: probability distributions, which priors make sensible computations, sampling methods like MCMC, approximation methods.

At least: You don't have to choose! You can learn both. And of course, you can combine both ;)

Summary

- Bayes rule
- Conjugate priors
- Bayesianism and Frequentism