# Advanced Machine Learning Subsidary Notes

Lecture 16: Bayesian Inference

Adam Prügel-Bennett

January 28, 2021

## 1 Keywords

- Bayes, Conjugate Priors, Uninformative Priors

## 2 Main Points

** Bayesian Statistics

- Background
  - Bayesian statistics is an approach to making statistical inference
  - It was first proposed by the Rev. Thomas Bayes but published after his death in 1761
  - It was championed by the great French polymath Pierre-Simon Laplace
  - It fell out of favour at the beginning of the $20^{th}$ Century due to philosophical prejudice
    - It was deemed to be unscientific because it requires specifying a prior probability which seemed objective
  - It was championed from the 1950's by Ed Jaynes who eventually convinced most people that Bayesian inference is consistent
  - Perhaps more importantly with the raise of computers it was possible to demonstrate that Bayesian methods work
- Bayes' Rule
  - Bayes rule follows from a simple identity

$$\mathbb{P}[h_i|\mathcal{D}] = \frac{\mathbb{P}[\mathcal{D}|h_i] \; \mathbb{P}[h_i]}{\mathbb{P}[\mathcal{D}]}$$

    - $\mathbb{P}[h_i|\mathcal{D}]$ is the *posterior* probability of hypothesis $h_i$ given data $\mathcal{D}$
    - $\mathbb{P}[\mathcal{D}|h_i]$ is the *likelihood* of the data given the hypothesis $h_i$
    - $\mathbb{P}[h_i]$ is our *prior* probability (our belief in hypothesis $h_i$ before seeing the data
    - $\mathbb{P}[\mathcal{D}]$ is a normalisation terms I will call the *evidence* and is just equal to

$$\mathbb{P}[\mathcal{D}] = \sum_{i=1}^{n} \mathbb{P}[h_i, \mathcal{D}] = \sum_{i=1}^{n} \mathbb{P}[\mathcal{D}|h_i] \; \mathbb{P}[h_i]$$

- Although this is just a mathematical identity, to use it in inference requires specifying a prior belief
  - I argued from the first lecture that for machine learning to work we need to make assumptions about the data
  - In Bayesian approach these assumptions our built into our likelihood and prior

- A nice property of Bayesian inference is that it turns the problem from an inverse problem (computing the posterior $\mathbb{P}[h_i|\mathcal{D}]$) into a forward problem of computing the likelihood $\mathbb{P}[\mathcal{D}|h_i]$
  - This solves problems of missing data (we only need to calculate the probability of the data we see)
  - It is also much more straightforward to model the forward process (from physics) rather than the inverse process
  - That said, it can still be technically very challenging
- Bayes' rule has the same form whether we are working with discrete variables (where we have probability masses) or continuous variables (where we have probability densities)

## 2.1 Conjugate Priors

- Technically performing Bayesian inference can be difficult because the posterior can be very ugly to work with
- But for a few of the classic likelihood functions there exist *conjugate prior distributions* such that the posterior is in the same class of distributions as the prior
  - Suppose that the likelihood of observing data $\boldsymbol{X}$ given some parameters $\boldsymbol{\theta}$ is $p(\boldsymbol{X}|\boldsymbol{\theta})$ (this may be a probability mass or density depending on whether $\boldsymbol{X}$ is continuous or not)
  - We want to infer $\boldsymbol{\theta}$ which we do through the posterior $p(\boldsymbol{\theta}|\boldsymbol{X})$
  - If we are lucky there is a family of distributions, $\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi})$ parameterised by $\boldsymbol{\phi}$, that describe the parameters of the likelihood $\boldsymbol{\theta}$ and satisfy

$$\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}') \propto p(\boldsymbol{X}|\boldsymbol{\theta})\,\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}')$$

  - Thus if we start with a prior $\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}_0)$ then after observing data $\boldsymbol{X}_1$ we would end up with a posterior $\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}_1)$ where $\boldsymbol{\phi}_1$ depends on observation $\boldsymbol{X}_1$ and $\boldsymbol{\phi}_0$
  - We can repeat this
    - When we make a second observation $\boldsymbol{X}2$ our posterior now becomes our prior (we have updated our beliefs) so now we obtain a new posterior

$$\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}_2) \propto p(\boldsymbol{X}_2|\boldsymbol{\theta})\,\text{Conj}(\boldsymbol{\theta}|\boldsymbol{\phi}_1)$$

- As an example let us consider a number of Bernoulli trials (we assume independent observations $X_i \in \{0, 1$ occurring with probability $p$
  - But we don't know $p$
  - We assume our prior distribution is a beta distribution $\text{Beta}(p|a_0, b_0) \propto p^{a_0-1}\,(1-p)^{b_0-1}$
  - Our likelihood is $\mathbb{P}[X|p] = p^X\,(1-p)^{1-X}$
  - Using Bayes' rule our posterior is

$$f(p|X) \propto p^X\,(1-p)^{1-X}\,p^{a_0-1}\,(1-p)^{b_0-1} = p^{a_0+X-1}\,(1-p)^{b_0+1-X-1}$$

  - But this has the same functional form as a beta-distribution
  - It will be a beta-distribution as it has to be normalised (posteriors are always normalised)
  - Thus $f(p|X) = \text{Beta}(p|a_1, b_1)$ where $a_1 = a_0 + X$ and $b_1 = b_0 + 1 - X$
- The Poisson distribution is defined as

$$\text{Pois}[N|\mu] = \frac{\mu^N\,\mathrm{e}^{-\mu}}{N!}$$

  - It describes the probability of observing $N$ events where each event is assumed independent and the expected number of events is $\mu$
  - Suppose we have made observations $N_1$, $N_2$, etc. and we are trying to infer $\mu$
  - The Poisson distribution has a conjugate distribution which is the gamma distribution

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

  - Assuming a prior $\mathrm{Gamma}(\mu|a_0,b_0)$ and assume we make an observation $N_1$ then the posterior is given by

$$f(\mu|N_1) \propto \mu^{N_1}\,\mathrm{e}^{-\mu}\,\mu^{a_0-1}\,\mathrm{e}^{-b_0\,\mu} = \mu^{a_0+N_1-1}\,\mathrm{e}^{-(b_0+1)\,\mu}$$

  - Thus $f(\mu|N_1) = \mathrm{Gamma}(\mu|a_1,b_1)$ where
    * $a_1 = a_0 + N_1$
    * $b_1 = b_0 + 1$

- Conjugate priors are the exception rather than the rule but they do occur for some classic distributions. E.g.

| **Likelihood** | **Prior** |
|---|---|
| Binomial/Bernoulli | Beta |
| Poisson | Gamma |
| Multinomial | Dirchlet |
| Univariate Normal | Gamma-Normal |
| Multivariate Normal | Wishart |

### 2.2 Uninformative Priors

- What should we do if we have no prior information

- This disturbed statisticians who felt that there might not be a subjective answer to this

- However, Ed Jaynes argued that there is a unique answer that we get by requiring invariance

- Scale parameters

  - Often we are trying to infer scale parameters
  - These would include the rate in the Poisson distribution or the standard deviation in a normal distribution
  - They are non-negative numbers that determine the scale (i.e. what units we should measure in)
  - For most problems we have some idea about this (otherwise if would be difficult to do the measurement)
  - But we may still not no the order of magnitude of what we are measure in
  - What should we use as a prior?
  - The answer to this was given by Harold Jeffreys and is known as the *Jeffreys' prior*
  - Let's see the argument
  - If we have not idea about what scale an observable $x$ takes then we should expect there to be equal probability to be in the intervals $[A,B]$ or the interval $[A/c,B/c]$
  - This may seem strange as these two intervals are different lengths (but non-overlapping if $c \neq 1$) but if this wasn't the case we would know something about which scale to use

- If our prior for $x$ is $p(x)$ then we require

$$\int_A^B p(x)\,\mathrm{d}x = \int_{A/c}^{B/c} p(x)\,\mathrm{d}x$$

- Now we can make a change of variables
- $y = c\,x$ so this last integral becomes

$$\int_{A/c}^{B/c} p(x)\,\mathrm{d}x = \int_A^B \frac{1}{c}\,p\Big(\frac{y}{c}\Big)\,\mathrm{d}y$$

- But for this to equal the first integral for any interval $[A, B]$ we require

$$p(x) \propto \frac{1}{x}$$

- We note that if $p(x) = 1/x$ then

$$p\Big(\frac{x}{c}\Big) = \frac{c}{x} = c\,p(x)$$

  as required

- The strange thing about this distribution is it in improper in that

$$\int_0^\infty \frac{1}{x} = \infty$$

- Strangely a lot of uninformative priors turn out to be improper
- However, this doesn't seem to matter in Bayesian inference: after making some observations we end up with a proper prior

- **Benford's Law**

  - There is a strange consequence of Jeffreys prior which was first speculated upon in 1881 when Simon Newcomb noticed that logarithm tables were much more used for numbers beginning with 1 rather 2 and 2 rather than 3, etc. Frank Benford in 1938 looked at the occurrence of numbers in data set and found that the digit of first significant figure was more likely to be 1 than 2 and more likely to be 2 than 3.
  - If we assume these numbers are measured in an arbitrary scale and the occurrence of these numbers followed Jeffreys prior then the probability that real number between 1 and 10 actually took values between $n$ and $n+1$ is given by

$$\frac{\int_n^{n+1} \frac{1}{x}\mathrm{d}x}{\int_1^{10} \frac{1}{x}\mathrm{d}x} = \frac{\log(n+1) - \log(n)}{\log(10)} = \log\left(\frac{n+1}{n}\right)$$

  - Note that we would get the same result if the number was between 10 and 100 and we asked the probability that it was in the range $10\,n$ to $10\,(n+1)$
  - But this means the digit of the most significant figure will be $n$
  - So numbers beginning with 1 occur more often than numbers beginning with 2, and numbers beginning with 2 occur more often than those beginning with 3, etc.
  - This is weird but true (for most naturally occurring non-negative numbers)

# 3 Exercises

## 3.1 Throwing Dice

- There are three dice on the table. Two of them are normal dice, while the other dice has 6 on two faces and 1, 2, 3 and 5 on the other face. A dice is chosen at random and is thrown 10 times. On three occasions the top face is a 6. What is the probability that the dice chosen is the dishonest dice?
- See answers below

# 4 Experiments

## 4.1 Benford's Law

- Benford's law is so wacky that to convince yourself it is true you really need to test it out
- Find a set of data with features that are clearly scale parameters and have a go (you need enough data points to be convincing)
  - if you numbers are positive and could take any value then they are likely to be scale parameter

# 5 Answers

## 5.1 Throwing Dice

- We use Bayes' rule

$$P(\text{dishonest}|\text{data}) = \frac{P(\text{data}|\text{dishonest})\,P(\text{dishonest})}{P(\text{data}|\text{dishonest})\,P(\text{dishonest}) + P(\text{data}|\text{honest})\,P(\text{honest})}$$

- Now $P(\text{dishonest}) = 1/3$ and $P(\text{honest}) = 2/3$
- The probability of $k$ success out of $n$ trials with a success probability $p$ is given by the binomial distribution

$$\text{Binom}(k|n,p) = \binom{n}{k} p^k \,(1-p)^{n-k}$$

- The probability of 3 out of 10 rolls being a 6 is given by

$$P(\text{data}|\text{dishonest}) = \text{Binom}(3|10, \tfrac{1}{3}) = \binom{3}{10}\left(\frac{1}{3}\right)^3 \left(\frac{2}{3}\right)^7$$

$$P(\text{data}|\text{honest}) = \text{Binom}(3|10, \tfrac{1}{6}) = \binom{3}{10}\left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^7$$

- Thus

$$P(\text{dishonest}|\text{data}) = \frac{1}{1 + \frac{P(\text{data}|\text{honest})\,P(\text{honest})}{P(\text{data}|\text{dishonest})\,P(\text{dishonest})}}$$

$$= \frac{1}{1 + 2\left(\frac{1}{2}\right)^3 \left(\frac{5}{4}\right)^7} = 0.29549$$