Advanced Machine Learning

Bayesian Inference

Bayes, Conjugate Priors, Uninformative Priors

Outline

Dealing with Uncertainty

- In machine learning we are attempting to make inference under uncertainty
- The natural language for discussing uncertainty is probability
- The natural framework for making inferences is Bayesian statistics
- However, this requires that we encode our prior knowledge of the problem and specify a likelihood
- In consequence, probabilistic methods tend to be bespoke, rather then general purpose black boxes

Revision on Bayes

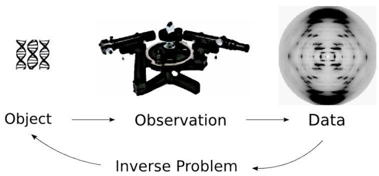
- Bayes' rule

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$$

  - ⋆ $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$ is the **posterior** probability of a hypothesis $\mathcal{H}_i$ (i.e. the probability of $\mathcal{H}_i$ **after** we see the data)
  - ⋆ $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$ is the **likelihood** of the data given the hypothesis. Note, that we calculated this from the forward problem
  - ⋆ $\mathbb{P}(\mathcal{H}_i)$ is the **prior** probability (i.e. the probability of $\mathcal{H}_i$ **before** we see the data)
  - ⋆ $\mathbb{P}(\mathcal{D})$ is the **evidence** or **marginal likelihood**

$$\mathbb{P}(\mathcal{D}) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{H}_i, \mathcal{D}) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{D}|\mathcal{H}_i)\,\mathbb{P}(\mathcal{H}_i)$$

Solving Inverse Problems



- We want the posterior $\mathbb{P}(\mathcal{H}_i|\mathcal{D})$ (i.e. the probability of what happened given some evidence)
- The Bayesian formalism converts this into the forward problem

$$\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$$

Bayesian Inference

- Bayes' rule says $\mathbb{P}(\mathcal{H}_i|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\mathcal{H}_i)\mathbb{P}(\mathcal{H}_i)}{\mathbb{P}(\mathcal{D})}$
- We calculate the likelihood $\mathbb{P}(\mathcal{D}|\mathcal{H}_i)$ (i.e. assuming the hypothesis, what is the chance of obtaining the data?)
- We consider the process of how the data is generated
- This uses the data we have (doesn't care about missing data)
- But we also need to know the prior $\mathbb{P}(\mathcal{H}_i)$
- Also, this can get difficult when we have many hypotheses

Evidence

- The normalisation term

$$\mathbb{P}(\mathcal{D}) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{H}_i, \mathcal{D}) = \sum_{i=1}^{n} \mathbb{P}(\mathcal{D}|\mathcal{H}_i)\,\mathbb{P}(\mathcal{H}_i)$$

  tells you how likely the data is (given the prior and likelihood function)
- It is called the **marginal likelihood** or **evidence**
- If we have two models M_1 and M_2 we can do **model selection** by choosing the model with the largest evidence $\mathbb{P}(\mathcal{D} \mid M_1)$ or $\mathbb{P}(\mathcal{D} \mid M_2)$
- This also allows us to select hyperparameters for a model

Probability Density

- When we are working with continuous variables it is more natural to work with probability densities

$$f_X(x) = \lim_{\delta x \to 0} \frac{\mathbb{P}(x \le X < x + \delta x)}{\delta x}$$

- Note that densities are non-negative, but can be greater than 1 (they are not probabilities)
- However

$$\mathbb{P}(a \le X \le b) = \int_a^b f_X(x)\,\mathrm{d}x$$

  is a probability and is less than or equal to 1

Densities and Bayes

- Bayes' rule also applies to densities

$$\mathbb{P}(x \le X < x + \delta x | Y) = \frac{\mathbb{P}(Y|x)\,\mathbb{P}(x \le X < x + \delta x)}{\mathbb{P}(Y)}$$

- Dividing by δx and taking the limit $\delta x \to 0$

$$f_{X|Y}(x|Y) = \frac{\mathbb{P}(Y|x)\, f_X(x)}{\mathbb{P}(Y)}$$

- Similarly if X is discrete and Y continuous

$$\mathbb{P}(X|y) = \frac{f_{Y|X}(y|X)\,\mathbb{P}(X)}{f_Y(y)}$$

- If both X and Y are continuous

$$f_{X|Y}(x|y) = \frac{f_{Y|X}(y|x)\, f_X(x)}{f_Y(y)}$$

Practical Bayesian Inference

- Often consider learning parameters $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

- This can be hard for large data sets as the posterior, $p(\boldsymbol{\theta}|\mathcal{D})$, is often a mess
- If we are lucky and have a simple likelihood then if we choose the right prior we end up with a posterior of the same form as the prior
- This occurs in some classic probabilistic inference problems, but as we will see soon it is also true for Gaussian Processes

Outline

1. Bayes' Rule
2. **Conjugate Priors**
3. Uninformative Priors

Learning a Probability

- Suppose we have a coin and we want to establish the probability of a head
- We want to learn this from a series of independent trials
- (Independent trials with two possible outcomes are known in probability theory as Bernoulli trials)
- Let X_i equal 1 if the i^{th} trial is a head and 0 otherwise
- If the probability of a head is p then the **likelihood** of a X_i is

$$\mathbb{P}(X_i|p) = p^{X_i}(1-p)^{1-X_i} = \begin{cases} p & \text{if } X_i = 1 \\ (1-p) & \text{if } X_i = 0 \end{cases}$$

Prior

- We may have a prior belief (e.g. we have made a few trials or we see the coin looks like a normal penny)
- We will suppose we can model our prior belief in terms of a **Beta distribution**

$$f(p) = \mathrm{Beta}(p|a,b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}$$

- $B(a,b)$ is just a normalisation constant

$$B(a,b) = \int_0^1 p^{a-1}(1-p)^{b-1}\mathrm{d}p = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

- This is a useful function for modelling the distribution of a random variable in the range 0 to 1

Uninformative Prior

- Suppose we have no idea about p what should we do?
- Laplace (one of the first Bayesian's) suggested giving equal weighting to all values of p
- This corresponds to a beta distribution with $a = b = 1$
- (Surprisingly other arguments suggest using $a = b = 0$ which provides a strong bias towards $p = 0$ and $p = 1$)
- Given enough data the prior is not so important and we will stick with Laplace for now

Independent Trials

- Using Bayes' rule

$$f(p|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|p)\, f(p)}{\mathbb{P}(\mathcal{D})}$$

- Assuming the trials are independent (a reasonably fair assumption for tossing coins) then the likelihood factorises

$$\begin{aligned}\mathbb{P}(\mathcal{D}|p) &= \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i} \\ &= p^{X_1}(1-p)^{1-X_1} p^{X_2}(1-p)^{1-X_2} \ldots p^{X_n}(1-p)^{1-X_n} \\ &= p^{\sum_i X_i}(1-p)^{\sum_i (1-X_i)} = p^s(1-p)^{n-s}\end{aligned}$$

$s = \sum_i X_i$ (number of successes/heads)

Posterior

- Plugging in a prior $f(p) = \mathrm{Beta}(p|a_0,b_0)$

$$f(p|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|p)\, f(p)}{\mathbb{P}(\mathcal{D})} = \frac{p^s(1-p)^{n-s} \times p^{a_0-1}(1-p)^{b_0-1}}{\mathbb{P}(\mathcal{D})\, B(a_0,b_0)}$$

- The denominator is a normalising factor

$$\begin{aligned}\mathbb{P}(\mathcal{D}) &= \int_0^1 \mathbb{P}(\mathcal{D}|p)\, f(p)\,\mathrm{d}p = \int_0^1 \frac{p^{s+a_0-1}(1-p)^{n-s+b_0-1}}{B(a_0,b_0)}\mathrm{d}p \\ &= \frac{B(s+a_0, n-s+b_0)}{B(a_0,b_0)}\end{aligned}$$

Conjugate Priors

- The posterior distribution is Beta distribution

$$f(p|\mathcal{D}) = \frac{p^{s+a_0-1}(1-p)^{n-s+b_0-1}}{B(s+a_0, n-s+b_0)} = \text{Beta}(p|s+a_0, n-s+b_0)$$

- Something rather nice happened
- Starting with a beta distributed prior $f(p) = \text{Beta}(p|a_0, b_0)$ for a set of Bernoulli trials we obtain a beta distributed posterior $f(p|\mathcal{D}) = \text{Beta}(p|a_0+s, b_0+n-s)$
- This is not always the case (often the posterior will be very complicated) but it happens for a few likelihoods and priors
- When the posterior is the same as the prior then the likelihood and prior distributions are said to be **conjugate**

Incremental Updating

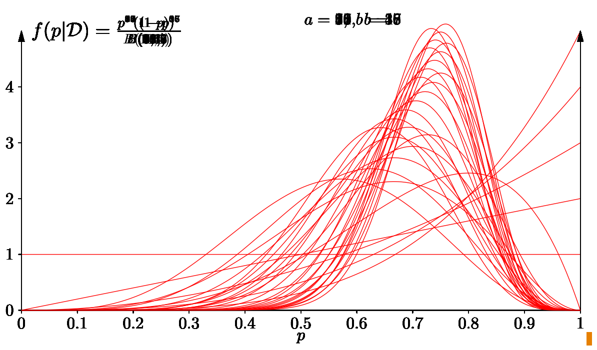
- For independent data we can update incrementally $\mathcal{D} = (X_1, X_2, \ldots, X_n)$

$$\begin{aligned} f(p|X_1) &= \frac{\mathbb{P}(X_1|p)\,f(p)}{\mathbb{P}(X_1)} \\ f(p|X_1, X_2) &= \frac{\mathbb{P}(X_2|p)\,f(p|X_1)}{\mathbb{P}(X_2)} \\ \vdots &= \vdots \\ f(p|X_1, X_2, \ldots, X_n) &= \frac{\mathbb{P}(X_n|p)\,f(p|X_1, \ldots, X_{n-1})}{\mathbb{P}(X_n)} \end{aligned}$$

- The posterior becomes the prior for the next piece of data
- For our problem the posterior is always Beta distributed

Example (p=0.7)

$\mathcal{D} = \{$H,H,H,T,T,H,H,H,H,T,H,T,H,H,H,H,H,T,H,H,H,H,T,H,H,H,T$\}$



Estimating Prediction Errors

- A full Bayesian treatment gives a prediction of its own error
- Assuming $f(p|\mathcal{D}) = \text{Beta}(p|a, b)$
- The expected value of p is given by $a/(a+b) = 23/32 = 0.719$
- The standard deviation is

$$\sqrt{\frac{ab}{(a+b)^2(a+b+1)}} = 0.078$$

Poisson Likelihoods

- Let's look at a second example of conjugate priors
- Suppose we want to find the rate of traffic along a road between 1:00pm and 2:00pm
- We assume the number of cars is given by a Poisson distribution

$$\mathbb{P}(N) = \text{Pois}(N|\mu) = \frac{\mu^N}{N!}\mathrm{e}^{-\mu}$$

- μ is the rate of traffic per hour which we want to infer from observation taken on different days

Using Bayes

- Let us assume a Gamma distributed prior

$$p(\mu) = \Gamma(\mu|a_0, b_0) = \frac{b_0^{a_0}\mu^{a_0-1}\mathrm{e}^{-b_0\mu}}{\Gamma(a)}$$

- We will assume that we know nothing. The uninformative prior is $a_0 = b_0 = 0$
- The data is $\mathcal{D} = \{N_1, N_2, \ldots, N_n\}$
- The likelihood is $\text{Pois}(N_i|\mu)$

Posterior

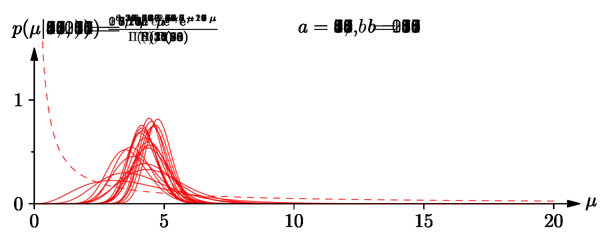
- The posterior after seeing the first piece of data is

$$\begin{aligned} p(\mu|N_1) &\propto \mathbb{P}(N_1|\mu)\,p(\mu) \\ &\propto \frac{\mu^{N_1}}{N_1!}\mathrm{e}^{-\mu}\mu^{a_0-1}\mathrm{e}^{-b_0\mu} \\ &\propto \mu^{N_1+a_0-1}\mathrm{e}^{-(b_0+1)\mu} \end{aligned}$$

- The posterior is also a Gamma distribution $\Gamma(\mu|a_1, b_1)$ with $a_1 = a_0 + N_1$, $b_1 = b_0 + 1$

Example ($\mu = 5$)

$\mathcal{D} = \{4, 4, 6, 4, 2, 2, 5, 9, 5, 4, 3, 2, 5, 4, 4, 11, 6, 2, 3, 11\}$



$$\mathbb{E}[\mu] = \frac{a}{b} = \frac{96}{20} = 4.8 \qquad \sqrt{\text{Var}(\mu)} = \sqrt{\frac{a}{b^2}} = 0.49$$

Outline

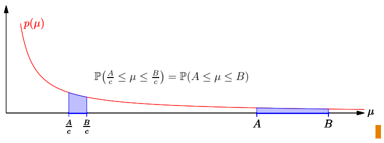
1. Bayes' Rule
2. Conjugate Priors
3. **Uninformative Priors**

Uninformative Priors

- What if we have no prior knowledge, what should we do?
- OK usually we know whether we should make a measurement using a micrometer, ruler or car mileage, but we might still know almost nothing
- This led to Bayesian statistics being labelled as *subjective*
- However Ed. Jaynes (the greatest proponent of Bayesian methods) argued that we could answer this using symmetry arguments

Uninformative Priors for Scale Parameter

- Why did we choose $a_0 = b_0 = 0$ implying a prior $p(\mu) = 1/\mu$?



- That is, we have no idea on what scale to measure μ

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

making a change of variables $\mu = \nu/c$

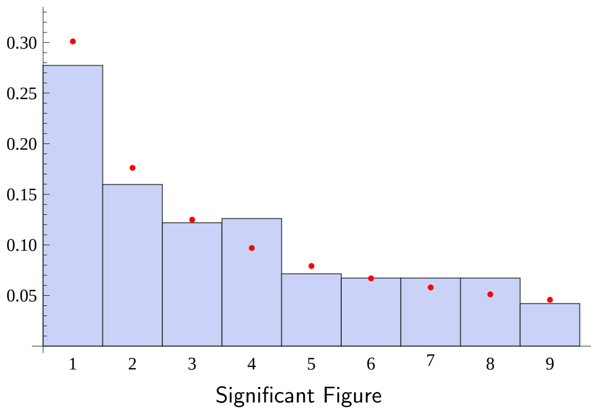
- Or $p(\mu) = \frac{1}{c} p(\frac{\mu}{c})$ implying $p(\mu) \propto \frac{1}{\mu}$

Benford's Law

- Numbers occurring in life (physical constants, amounts of money) should not depend on the units (scale) measuring them
- They should then be distributed as $p(x) \propto 1/x$
- A curious consequence of this is that the significant figure has a distribution

$$\mathbb{P}(\text{most s.f. of } x = n) = \frac{\int_n^{n+1} \frac{1}{x}\mathrm{d}x}{\int_1^{10} \frac{1}{x}\mathrm{d}x} = \frac{\int_{10n}^{10n+10} \frac{1}{x}\mathrm{d}x}{\int_{10}^{100} \frac{1}{x}\mathrm{d}x}$$
$$= \frac{\log(n+1) - \log(n)}{\log(10)} = \log_{10}\left(\frac{n+1}{n}\right)$$

Population Size of 238 Countries



Conclusion

- Bayesian inference provides a coherent framework which we can use for machine learning
- However, it requires a model of what is happening
- In practice Bayesian methods are easy if the data is generated from a likelihood with a conjugate prior distribution—we have to be clever to choose the right prior
- We will see in the next lecture that much more frequently we will have likelihoods with no conjugate prior and we have to work much harder
- When we have no knowledge there are consistent ways to express our ignorance