Advanced Data Analytics
Lecture week 6: Bayesian principles

Ian T. Nabney

University of
BRISTOL

Overview

- Understand motivation for Bayesian inference
- Overview of different techniques for achieving effective Bayesian algorithms
- Selection of data science tasks that Bayesian inference can be applied to

Further reading: Bishop sections 1.3.2, 3.3, 3.4, introduction to Chapter 10. Nabney section 9.1.

Bayesian interpretation of probability

- We have viewed probabilities in terms of the frequencies of random, repeatable events. This is the classical or frequentist interpretation of probability.
- Now we turn to the more general Bayesian view, in which probabilities provide a quantification of uncertainty.
- Consider an uncertain event, such as whether the Arctic ice cap will have disappeared by the end of the century. This is not an event that can be repeated numerous times in order to define a notion of probability as we did earlier in the context of boxes of fruit.
- Nevertheless, we will generally have some idea, for example, of how quickly we think the polar ice is melting. If we now obtain fresh evidence, for instance from a new Earth observation satellite gathering novel forms of diagnostic information, we may revise our opinion on the rate of ice loss.
- Our assessment of such matters will affect the actions we take, for instance the extent to which we endeavour to reduce the emission of greenhouse gasses. We would like to be able to quantify our expression of uncertainty and make precise revisions of uncertainty in the light of new evidence.

Bayesian inference

- We capture our assumptions about $\mathbf{w}$, before observing the data, in the form of a prior probability distribution $p(\mathbf{w})$.
- Bayes' theorem in the form

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \tag{1}$$

  allows us to evaluate the uncertainty in $\mathbf{w}$ after we have observed $\mathcal{D}$ in the form of the posterior probability $p(\mathbf{w}|\mathcal{D})$
- In both the Bayesian and frequentist paradigms, the likelihood function $p(\mathcal{D}|\mathbf{w})$ plays a central role. However, the manner in which it is used is fundamentally different in the two approaches.
- In a frequentist setting, $\mathbf{w}$ is considered to be a fixed parameter, whose value is determined by some form of 'estimator', and error bars on this estimate are obtained by considering the distribution of possible data sets $\mathcal{D}$.
- From the Bayesian viewpoint there is only a single data set $\mathcal{D}$ (namely the one that is actually observed), and the uncertainty in the parameters is expressed through a probability distribution over $\mathbf{w}$.

Using the posterior

- Bayesian inference is very simple in principle. In equation 1, $p(\mathcal{D})$ (known as the evidence is a normalisation factor that ensures that the posterior integrates to 1. It is given by an integral over the parameter space.

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w}')p(\mathbf{w}')\,\mathrm{d}\mathbf{w}'. \tag{2}$$

- Once the posterior has been calculated, every type of inference is made by integrating over this distribution. For example, to make a prediction at a new input $\mathbf{x}^*$, we need to calculate the prediction distribution

$$p(\mathbf{y}|\mathbf{x}^*, \mathcal{D}) = \int p(\mathbf{y}|\mathbf{x}^*, \mathbf{w})p(\mathbf{w}|\mathcal{D})\,\mathrm{d}\mathbf{w}. \tag{3}$$

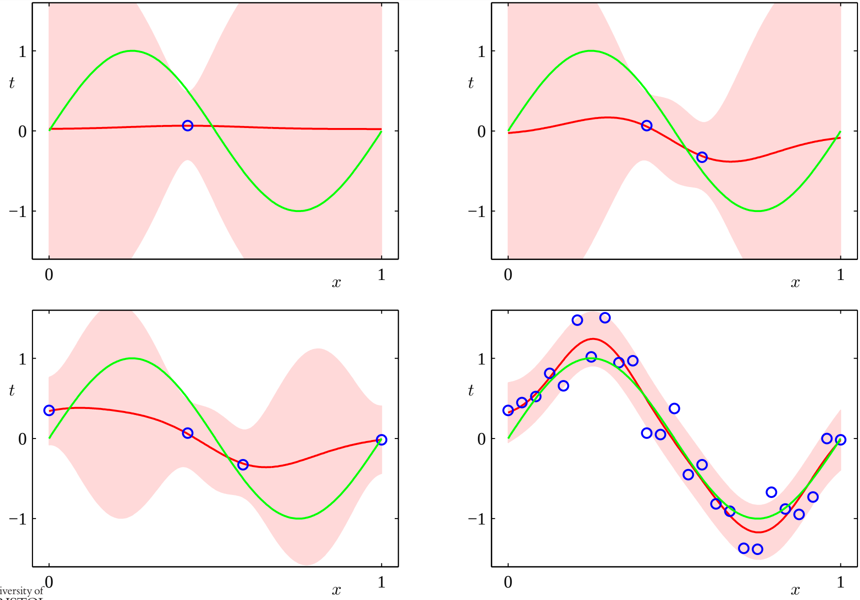
  A point prediction uses the mean of this distribution, given by

$$E(\mathbf{y}|\mathbf{x}^*, \mathcal{D}) = \int \mathbf{y}\; p(\mathbf{y}|\mathbf{x}^*, \mathcal{D})\,\mathrm{d}\mathbf{y}, \tag{4}$$

  while the variance of the prediction distribution (given by a similar integral) can be used for error bars.

- The problem is that evaluating integrals such as (2) and (4) is very difficult because they are only analytically tractable for a small class of prior and likelihood distributions. The dimensionality of the integrals is given by the number of network parameters, so simple numerical integration algorithms break down.

Predictive distributions



Benefits of Bayesian approach

1. In principle, by taking account of parameter uncertainty, overfitting is not a problem.
2. Regularisation can be given a natural interpretation in the Bayesian framework. Being able to reason consistently with regularisation parameters makes it easier to find good values for them. For example, it is possible to optimise regularisation parameters as part of the training process.
3. Parameter uncertainty can be accounted for in network predictions. For regression problems, error bars, or prediction intervals, can be assigned to network predictions. For classification problems, output class probabilities are moderated to less extreme values.
4. There is a principled framework for deciding questions of model complexity.
5. The relative importance of different input variables can be determined using automatic relevance determination (ARD) by choosing a specific weight prior.

Bayesian inference in practice

- For many models of practical interest, it will be infeasible to evaluate the posterior distribution or indeed to compute expectations with respect to this distribution.
  - The dimensionality of the latent space may be too high to work with directly or because the posterior distribution has a highly complex form for which expectations are not analytically tractable.
  - For discrete variables, the marginalizations involve summing over all possible configurations of the hidden variables, and though this is always possible in principle, we often find in practice that there may be exponentially many hidden states so that exact calculation is prohibitively expensive.
- In such situations, we need to resort to approximation schemes, and these fall broadly into two classes, according to whether they rely on stochastic or deterministic approximations.
  - Stochastic techniques such as Markov chain Monte Carlo have enabled the widespread use of Bayesian methods across many domains.
  - Deterministic approximation schemes some of which scale well to large applications. These are based on analytical approximations to the posterior distribution, for example by assuming that it factorizes in a particular way or that it has a specific parametric form such as a Gaussian.

Bayesian linear regression

- We shall consider zero-mean isotropic Gaussian weight prior governed by a single precision parameter α so that

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{5}$$

and the corresponding posterior distribution over **w** is then given by

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{6}$$

with

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^{\mathrm{T}} \mathbf{t} && (7) \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Phi}. && (8) \end{aligned}$$

- Consider a single input variable x, a single target variable t and a linear model of the form $y(x, \mathbf{w}) = w_0 + w_1 x$. Because this has just two adaptive parameters, we can plot the prior and posterior distributions directly in parameter space.
- The figure demonstrates the sequential nature of Bayesian learning in which the current posterior distribution forms the prior when a new data point is observed.

Bayesian model comparison

- The over-fitting associated with maximum likelihood can be avoided by marginalizing (summing or integrating) over the model parameters instead of making point estimates of their values.
- Models can then be compared directly on the training data, without the need for a validation set. This allows all available data to be used for training and avoids the multiple training runs for each model associated with cross-validation.
- It also allows multiple complexity parameters to be determined simultaneously as part of the training process.

Comparision framework

- Suppose we wish to compare a set of L models $\{\mathcal{M}_i\}$ where $i = 1, \ldots, L$.
- We shall suppose that the data is generated from one of these models but we are uncertain which one. Our uncertainty is expressed through a prior probability distribution $p(\mathcal{M}_i)$. Given a training set $\mathcal{D}$, we then wish to evaluate the posterior distribution

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i). \tag{9}$$

- Once we know the posterior distribution over models, the predictive distribution is given, from the sum and product rules, by

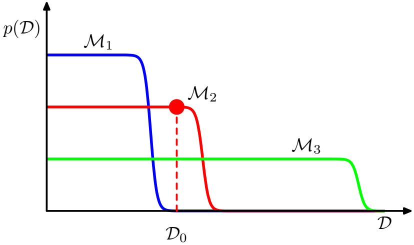
$$p(t|\mathbf{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x}, \mathcal{M}_i, \mathcal{D})p(\mathcal{M}_i|\mathcal{D}). \tag{10}$$

Model evidence

- The model evidence $p(\mathcal{D}|\mathcal{M}_i)$ (also known as the marginal likelihood) expresses the preference shown by the data for different models.
-
$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i)p(\mathbf{w}|\mathcal{M}_i)\,\mathrm{d}\mathbf{w}. \tag{11}$$

  From a sampling perspective, the marginal likelihood can be viewed as the probability of generating the data set $\mathcal{D}$ from a model whose parameters are sampled at random from the prior.

- Consider three models $\mathcal{M}_1$, $\mathcal{M}_2$ and $\mathcal{M}_3$ of successively increasing complexity.
- Imagine running these models generatively to produce example data sets, and then looking at the distribution of data sets that result.
- To generate a particular data set from a specific model, we first choose the values of the parameters from their prior distribution $p(\mathbf{w})$, and then for these parameter values we sample the data from $p(\mathcal{D}|\mathbf{w})$.



Summary

- Understand motivation for Bayesian inference
- Overview of different techniques for achieving effective Bayesian algorithms
- Selection of data science tasks that Bayesian inference can be applied to

Next we shall consider two approximation frameworks (the evidence procedure and variational inference) and also look at stochastic methods briefly.