Machine Learning - Michaelmas Term 2017
Lecture 7 : Bayesian Approach to Machine Learning

Lecturers: Christoph Haase & Varun Kanade

1 Bayesian Approach to Machine Learning

In this section, we briefly describe the Bayesian approach to machine learning and it's connections to Ridge Regression and Lasso. For the most part in this course, we'll not adopt the Bayesian approach. However, it is well worth understanding at least the basic aspects of this approach. The description of the Bayesian approach and its relation to the "frequentist" approach provided here is far from adequate and those interested should refer to Murphy (2012, Chap. 5, 6) and beyond.

In the "frequentist" approach, the assumption is that there are "true" parameters, unknown though they may be to us. The goal is to make use of data, which depends on the true parameters and which we observe through some random process, to infer the true "unknown" parameters. In the Bayesian approach, in the absence of any data, a belief about what the parameters may be is represented by a *prior* distribution on the parameters; let us denote this prior on the parameters by $p(\mathbf{w})$. As in the frequentist setting, the data will depend on parameters and will be observed through some random process. When the data, denoted by $\mathcal{D}$ is observed, the belief about the parameters is updated and represented using what is called the *posterior* distribution. This distribution is obtained using Bayes' Rule using the prior distribution $p(\mathbf{w})$ and the (probabilistic) data model, denoted by $p(\mathcal{D} \mid \mathbf{w})$. Then, the posterior on the model parameters, given the data, is given by:

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

Thus, the posterior reflects the updated belief about the parameters after observing some data. In the limit of infinite data, the posterior distribution will become a point mass at the maximum likelihood estimate (as long as the prior has non-zero mass everywhere).

Let us now discuss how a prediction is made using this approach. To make things a bit more concrete, let's suppose that the new input point is $\mathbf{x}_{\text{new}}$ and we wish to predict the output y_{new} (or in general a distribution over the output). There are two approaches to this, the first is to use a point-estimate (or a plugin estimate) which uses a single set of parameters that are chosen to represent the posterior distribution. For example, this may be the posterior mean, median or mode. The second approach is to use the entire posterior distribution to make the prediction, sometimes referred to as the full Bayesian approach, by integrating out the parameters $\mathbf{w}$. Thus, we may express:

$$p(y \mid \mathbf{x}_{\text{new}}, \mathcal{D}) = \int_{\mathbf{w}} p(y \mid \mathbf{w}, \mathbf{x}_{\text{new}}) \cdot p(\mathbf{w} \mid \mathcal{D}) \, \mathrm{d}\mathbf{w}.$$

While the full Bayesian approach is certainly desirable as it accounts for all our prior beliefs as well as the observed data, for all but the simplest of models this can be computationally expensive (or even intractable!). There is a lot of research on developing approximate methods in the case of computational intractability, which we will not cover in the course.

Let us now return to the first approach, which is to obtain a point estimate. The mode, however unrepresentative of the distribution as whole it may be, stands out for one reason. In order to compute the median or mean of the posterior distribution it is necessary to compute

the denominator of (1). This denominator represents just the probability of observing the data, obtained by integrating out the parameters $\mathbf{w}$, *i.e.,*

$$p(\mathcal{D}) = \int_{\mathbf{w}} p(\mathcal{D} \mid \mathbf{w}) \cdot p(\mathbf{w}) \, \mathrm{d}\mathbf{w}.$$

However, except in relatively simple cases, even this integral may be computationally expensive to evaluate. In order to obtain the mode though, the denominator is unnecessary, it can be obtained by simply looking for $\mathbf{w}$ where the numerator of (1) achieves the maximum value. Thus, it is often common to express the posterior as,

$$p(\mathbf{w} \mid \mathcal{D}) \propto p(\mathcal{D} \mid \mathbf{w}) \cdot p(\mathbf{w}) \tag{2}$$

The mode of the posterior is a point estimate known as the *maximum a posteriori* or MAP estimate. This can be obtained by finding $\mathbf{w}$ that maximises the RHS of (2). For most of this section, we'll focus on computing the MAP estimate.

1.1 Bayesian Linear Regression

Let us now look specifically at linear regression through the Bayesian approach. We'll still consider the linear model given by,

$$p(y \mid \mathbf{x}, \mathbf{w}) = \mathcal{N}(\mathbf{w} \cdot \mathbf{x}, \sigma^2)$$

Throughout this section, we'll think of σ^2 as fixed and known. Thus, we're only representing $\mathbf{w}$ as the parameters. We need to define a prior distribution over $\mathbf{w}$; let us assume that this is a spherical Gaussian distribution with mean $\mathbf{0}$ and variance τ^2 in each direction,

$$\begin{aligned} p(\mathbf{w}) &= \mathcal{N}(\mathbf{0}, \mathbf{\Lambda}), && \text{where } \ \mathbf{\Lambda} = \tau^2 \mathbf{I}_D \\ &= \frac{1}{(2\pi\tau^2)^{D/2}} \cdot \exp\left(-\frac{\mathbf{w}^\mathsf{T}\mathbf{w}}{2\tau^2}\right) \end{aligned}$$

As we've been doing so far, given data $\mathcal{D} = \langle (\mathbf{x}_i, y_i) \rangle_{i=1}^N$, we can represent $\mathbf{y}$ given model parameters $\mathbf{w}$ and inputs $\mathbf{X}$ as,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \frac{1}{(2\pi\sigma^2)^{N/2}} \cdot \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w})}{2\sigma^2}\right)$$

Thus, we can express the posterior as,

$$p(\mathbf{w} \mid \mathbf{X}, \mathbf{y}) \propto \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w})}{2\sigma^2}\right) \cdot \frac{1}{(2\pi\tau^2)^{D/2}} \exp\left(-\frac{\mathbf{w}^\mathsf{T}\mathbf{w}}{2\tau^2}\right)$$

The *maximum a posteriori* or MAP estimate is obtained by finding the value of $\mathbf{w}$ that maximises the RHS of the above expression. Since σ and τ are fixed, we can express this as:

$$\mathbf{w}_{\text{map}} = \operatorname*{argmax}_{\mathbf{w}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w})}{2\sigma^2} - \frac{\mathbf{w}^\mathsf{T}\mathbf{w}}{2\tau^2}\right)$$

Using the fact that log is monotone and converting argmax to argmin by flipping signs, we get

$$\begin{aligned} \mathbf{w}_{\text{map}} &= \operatorname*{argmin}_{\mathbf{w}} \left(\frac{(\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w})}{2\sigma^2} + \frac{\mathbf{w}^\mathsf{T}\mathbf{w}}{2\tau^2}\right) \\ \mathbf{w}_{\text{map}} &= \operatorname*{argmin}_{\mathbf{w}} \left((\mathbf{y} - \mathbf{X}\mathbf{w})^\mathsf{T}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\sigma^2}{\tau^2} \cdot \mathbf{w}^\mathsf{T}\mathbf{w}\right) \end{aligned} \tag{3}$$

Comparing the form of (**??**) and (3), we see that the MAP estimate is exactly that given by minimising the Ridge Regression objective with $\lambda = \sigma^2/\tau^2$.

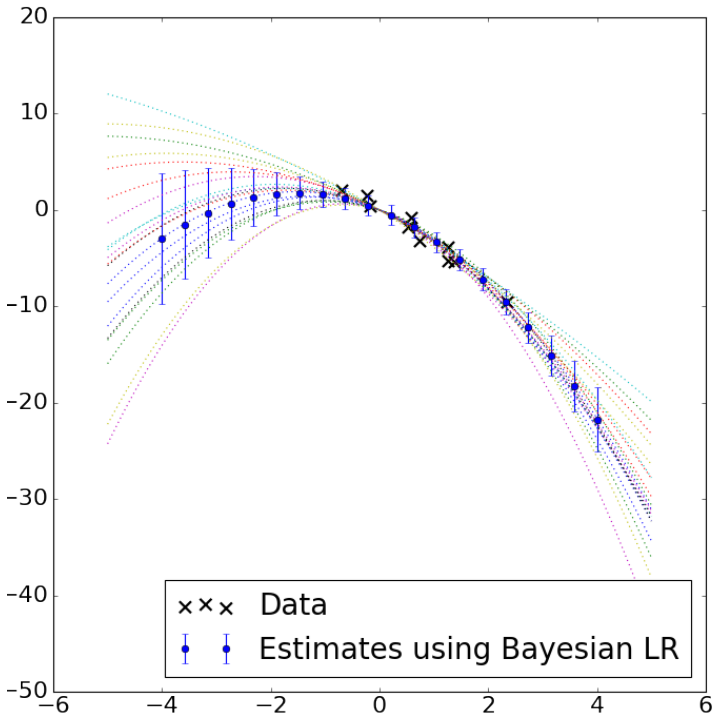

Figure 1: Full Bayesian Approach for polynomial regression in one dimension.

1.1.1 Full Bayesian Approach for Linear Regression

We mentioned earlier that the full Bayesian approach can be computationally expensive. In the case of Bayesian linear regression discussed in this section, everything can be expressed in closed form. The calculations are a bit tedious and the interested student is referred to Murphy (2012, Sec 7.6). However, let us see the advantage of this approach. For the setting described here, we can express the distribution over the output y of a new data point $\mathbf{x}_{\text{new}}$ as follows:

$$p(y \mid \mathbf{w}_{\text{new}}, \mathcal{D}) = \mathcal{N}(\mathbf{w}_{\text{map}} \cdot \mathbf{x}_{\text{new}}, \sigma^2 + \mathbf{x}_{\text{new}}^{\mathsf{T}} \mathbf{V}_N \mathbf{x}_{\text{new}}) \tag{4}$$

where

$$\mathbf{V}_N = \sigma^2 \left(\mathbf{X}^{\mathsf{T}}\mathbf{X} + \frac{\sigma^2}{\tau^2} \cdot \mathbf{I}_D \right)^{-1} \tag{5}$$

It can be shown (using singular value decomposition) that the variance in (4) is relatively small for $\mathbf{x}_{\mathbf{w}}$ that "look like" previously observed data and large for those that don't. Thus, the predictions include higher degree of uncertainty in parts of the input space where there is scarce data and less uncertainty where data is plenty. Figure 1 shows this for polynomial regression in one dimension. As shown in the figure, one way to think of the full Bayesian approach is to make prediction using $\mathbf{w}$ sampled from the posterior distribution; the figure shows models represented by several samples of $\mathbf{w}$ drawn from the posterior distribution as well as error bars representing the uncertainty. It can be seen that in the region where there is a lot of data almost all models drawn from the posterior make almost the same predictions, but in regions where data is scarce the predictions can be quite different.

Choosing a Prior

How to choose a prior in the Bayesian approach? That is one of the central questions and often a point of criticism of this approach. While in principle one should choose a prior that reflects

the true beliefs, often priors are chosen for mathematical convenience. In the absence of any definite beliefs about the prior, one should choose a prior that is as uninformative as possible. We'll not cover these aspects of the Bayesian approach in the course; the interested student may refer to the textbook by Murphy (2012).

References

Kevin P. Murphy. *Machine Learning : A Probabilistic Perspective*. MIT Press, 2012.