PGMs week 10 Learning: Introduction to Bayesian Techniques

With this week's topic we can only very briefly glance at a large and important approach to learning. The scope is well beyond what we can cover in such limited time, I recommend that you initially just aim to understand the broad ideas involved.

We have already in the last two weeks encountered the concept that the parameters of a model can indeed also be represented as random variables, thereby implying that they are also governed by a probability distribution. However, up till now we always ultimately reduced that distribution to a single set of parameter values. For instance with the MAP techniques we searched for that single set of parameter values where its posterior distribution achieved a maximum.

Bayesian techniques do not try to reduce or summarise that distribution by a single set of values, but endeavour to use the full spectrum over which the parameters vary.

There are a few different situations in which this idea is applied:

1 Parameter distribution estimation

Study material: Watch the MacKay video 10 as well as Koller 15.3 and 15.5 videos (focuses on discrete variables). Read MacKay Chapter 3 and Barber Sections 8.8, 9.2, 9.4.

Here we want to estimate the *distribution* of the model parameters:

$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta)p(\Theta|\lambda)}{p(\mathcal{D})}, \tag{1}$$

with $\mathcal{D}$ the training data, Θ the parameters, λ the hyper parameter(s), $p(\Theta|\lambda)$ the prior on Θ, $p(\mathcal{D}|\Theta)$ the likelihood of Θ and $p(\Theta|\mathcal{D})$ the posterior of Θ given $\mathcal{D}$. The data $\mathcal{D}$ typically consists of observed training pairs $\{(x_n, y_n)\}, n = 1 \ldots N$.

- We first need to consider the estimation of the hyper parameters λ:
  - In the full Bayesian approach we simply choose it. If we somehow have a good external indication on what is reasonable we might want to stick with this, otherwise we may choose a so-called 'uninformative' prior.
  - We can vary them systematically and set them by evaluating on a validation set.
  - Optimise them by making use of Maximum Likelihood (no validation set required). This is known as ML-II / Empirical Bayes / Evidence Approximation:
    - Use the EM algorithm – the model parameters Θ are now the latent variables (together with possible other latent variables such as missing data etc), and the hyper parameters are the system parameters to be estimated. We make an initial choice for them, use that to find a distribution over Θ (this might be tricky – see some of the techniques mentioned later), use that distribution to ML update our choice of the hyper-parameters and iterate till convergence.
    - Use an optimisation technique to directly determine the most likely hyper-parameter values. This typically involves a marginalisation over Θ which might be tricky (see later). For gradients, see Barber Eq 11.6.3 – note how that is related to the M step function in the EM algorithm.
    - Question: What about the danger of overspecialising on the training set? Yes, there is an increased risk, but we also see the elegance of the Occam factor (see below) – specialising needs parameters that are unusual in terms of the prior. If we set the prior parameters wide to allow this, the prior itself is lowered to keep the volume at unity. This penalises specialisation. There is a dynamic balance between the system parameters and its hyper-parameters.
- How do we estimate the model parameters? We don't, we only need to estimate the hyper-parameters (see above). The observed data $\mathcal{D} = \{(x_n, y_n)\}$ can be combined with the hyper-parameters λ to infer a distribution over the parameters Θ (and possibly also other latent variables) as $p(\Theta|\mathcal{D}, \lambda)$.

2 Model comparison

Study material: Read MacKay Chapter 28 and Barber Chapter 12

This one is closely related to the above, but the purpose is to compare which one of several models is more likely to explain the observed data.

$$p(\mathcal{H}_i|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{H}_i)p(\mathcal{H}_i)}{p(\mathcal{D})}, \tag{2}$$

$$p(D|\mathcal{H}_i) = \int p(\mathcal{D}|\Theta, \mathcal{H}_i)p(\Theta|\mathcal{H}_i, \lambda)d\Theta, \tag{3}$$

where $\mathcal{H}_i$ is the hypotheses that a specific model i is the appropriate model explaining the data, $p(D|\mathcal{H}_i)$ is the evidence for $\mathcal{H}_i$, $p(\mathcal{D}|\Theta, \mathcal{H}_i)$ is the best fit likelihood and $p(\Theta|\mathcal{H}_i, \lambda)$ is the Occam factor. Where-as the likelihood term prefers more complex models, the Occam factor automatically penalises complexity. Typically in model comparison the ratio between competing models is determined:

$$\frac{p(\mathcal{H}_i|\mathcal{D})}{p(\mathcal{H}_j|\mathcal{D})} = \frac{p(\mathcal{D}|\mathcal{H}_i)}{p(\mathcal{D}|\mathcal{H}_j)}\frac{p(\mathcal{H}_i)}{p(\mathcal{H}_j)}. \tag{4}$$

This ratio sometimes is given on a dB scale.

3 Prediction / classification

Study material: Watch Koller 15.4 (focuses on discrete variables) and MacKay video 14. Read Barber 11.5, 12.5.1, 18.2.2-3 and Chapter 28.

How do we do Bayesian Classification? First of all we need to infer the distribution of $p(\Theta|\mathcal{D}, \lambda)$ as described above. To classify an newly observed x as $Y = y$, we marginalise over this distribution as follows:

$$\begin{aligned} p(Y|x, \mathcal{D}, \lambda) &= \int_\Theta p(Y, \Theta|x, \mathcal{D}, \lambda) \\ &= \int_\Theta p(Y|\Theta, x, \mathcal{D}, \lambda)p(\Theta|x, \mathcal{D}, \lambda) \\ &\approx \int_\Theta p(Y|\Theta, x, \mathcal{D}, \lambda)p(\Theta|\mathcal{D}, \lambda) \\ &= \langle p(Y|\Theta, x, \mathcal{D}, \lambda)\rangle_{p(\Theta|\mathcal{D},\lambda)} \end{aligned} \tag{5}$$

All of this seems pretty straightforward, but the devil is in the details. How do we infer the distribution for Θ, and how do we solve that marginalisation integral? Remembering that inference also requires marginalisation these two issues are not unrelated. Unfortunately for many (most?) interesting problems there are no exact solutions available. Broadly we have two major options to follow:

- We can make use of Monte Carlo based techniques, specifically the Markov Chain Monte Carlo (MCMC) techniques, to approximate these integrals. There are some Koller and MacKay videos on this subject, but we are not going to pursue this avenue.
- We can use the Laplace approximation to replace a problematic distribution with the best Gaussian sharing the same mode as the original distribution. To do this we need to find the MAP value of the problematic distribution, as well as its second derivative (the Hessian) at the mode. These two become the mean vector and covariance matrix of the Gaussian approximation. Note, this is not the *best* Gaussian fit to the distribution, but the best one that also shares the same mode.
- We can make use of variational inference to replace distributions with others that approximate them well, while also making the problem amenable to practical processing. As you might guess, the KL divergence is useful here.

4 Exercises

1. (Highly recommended to do.) What is the probability P_H that a flipped coin will show heads? Assume an uniform/uninformative prior for this parameter. Then toss a coin 9 times, noting the outcome of each toss. Use Eq 1 and draw a series of posterior distributions for P_H.

2. (Highly recommended to do.) Suppose that in the above 9 tosses you got 6 heads and 3 tails.
   (a) What is the maximum likelihood estimate of P_H?
   (b) Intuitively, which hypothesis do you think is more likely:
   $H_1 : P_H = 0.5$ or $H_2 : P_H \in \{[0 : 1]\}$?
   (c) Now use model comparison to formally compare H_1 to H_2 under the prior assumption that the two hypotheses are equally likely. Can you explain your results?

3. (Recommended to do.) Inferring the distribution for the parameters of a Gaussian: Generate a number of samples from some known Gaussian distribution. Consult Barber Section 8.8.2 and MacKay Chapter 24 and estimate the joint distribution for the mean and variance from the sampled data. For this case there is an analytic solution.

4. (This one is rather interesting due to the variational aspect. But it is more involved and will require some persistence of you.) In the previous exercise you would have found that the variance parameter is dependent on the mean parameter. Approximate this joint distribution as the product of two independent distributions. Use variational inference to infer these two distributions and compare their joint product with the actual distribution determined in the previous exercise. Consult MacKay Chapter 33 for more information.

5. (Deeper.) Revisit the logistic regression example from earlier exercises, but now implement a Bayesian version of it and compare with your previous results. Barber Section 18.2.3 as well as Chapter 28 should come handy for this.