DE424 Week 9 Learning: Bayesian estimation

In Week 8, we estimated the parameters of a model by finding the set of parameter values that maximises the likelihood function, $p(\mathcal{D}|\boldsymbol{\Theta})$; however, this maximum-likelihood estimation was an ad-hoc approach that does not work well in all scenarios. A more principled approach – the Bayesian approach – is to model *everything* you are uncertain about as RVs. According to this approach, the parameters, $\boldsymbol{\Theta}$, are now considered RVs, and estimating the parameters from data, $\mathcal{D}$, amounts to calculating the posterior distribution,

$$p(\boldsymbol{\Theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}),$$

using standard inference techniques. We call this approach to parameter estimation the *full Bayesian approach*. For this approach, we have to specify a prior distribution $p(\boldsymbol{\Theta})$ for the parameters; this prior distribution also contains parameters, which are called *hyperparameters*.

In the full Bayesian approach, we represent our knowledge of the parameters as probability distributions, not specific values. This could make inference difficult and slow. An approach that avoids the complexity of the full Bayesian approach, but one that is more principled than MLE is *maximum a posteriori (MAP) estimation*, where we find the set of parameter values that maximise the posterior distribution over the parameters, or

$$\begin{aligned}\boldsymbol{\Theta}^{\text{MAP}} &= \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{\Theta}|\mathcal{D}) \\ &= \arg\max_{\boldsymbol{\Theta}} p(\mathcal{D}|\boldsymbol{\Theta})p(\boldsymbol{\Theta}).\end{aligned}$$

In the practical part of this assignment, we will use both MAP estimation and the full Bayesian approach. For both approaches, we have to choose a prior distribution, $p(\boldsymbol{\Theta})$. For a specific likelihood function, $p(\mathcal{D}|\boldsymbol{\Theta})$, we try to choose the prior distribution such that the posterior distribution, $p(\boldsymbol{\Theta}|\mathcal{D})$, is in a convenient form and easy to calculate; this requires a number of new distributions (e.g. the Dirichlet and gamma distributions; see also the concept of a *conjugate distribution* below). It is important to have an idea of how to choose prior parameter distributions as well as how to work with them; however, we will sidestep this complexity for the practical part by looking at a problem for which we can use a Gaussian prior distribution for the parameters.

1 Preparation: Watch the Koller videos on Bayesian estimation

Watch the following videos in the Coursera course Probabilistic Graphical Models 3: Learning, Week 1:

- *Bayesian Parameter Estimation for BNs: Bayesian Estimation (15:27)*
- *Bayesian Parameter Estimation for BNs: Bayesian Prediction (13:40)*
- *Bayesian Parameter Estimation for BNs: Bayesian Priors for BNs (17:02)*

Watch the following videos in the Coursera course Probabilistic Graphical Models 3: Learning, Week 2:

- *Parameter Estimation in MNs: MAP Estimation for MRFs and CRFs (9:59)*

Notes on these videos:

- The 3 videos about *Bayesian Parameter Estimation for BNs* follow the full Bayesian approach, whereas the video *Parameter Estimation in MNs: MAP Estimation for MRFs and CRFs (9:59)* discusses MAP estimation.
- The video *Bayesian Parameter Estimation for BNs: Bayesian Estimation (15:27)* introduces the idea of *conjugate prior distributions*, which are specific choices for the parameter prior distribution such that the posterior distribution is in same parametric family (e.g. the Dirichlet distribution is a conjugate prior distribution for the categorical distribution). Take note of the idea; however, we will not use it in the practical parts of this assignment.

- With continuous features, the required integration required for marginalisation in the full Bayesian approach often makes this impossible to do directly. One of our alternatives is to switch to max-sum inference instead of sum-product inference. The maximisation can then be done by finding the maximal points on the respective distributions – this can be done via differentiation/optimisation, an easier alternative to the integrations originally required. In effect this finds the most likely parameter values for the model given the observed data.
- In Bayesian estimation, equations often take the form where it appears as if the training data is supplemented with a number of extra "ghost" features which correspond to the characteristics of the prior distribution– this is the so-called "equivalent sample size" introduced by the prior distribution, as discussed in *Bayesian Parameter Estimation for BNs: Bayesian Prediction (13:40)*. As the number of data points increase, ML and MAP estimation converge to the same values.
- In *Parameter Estimation in MNs: MAP Estimation for MRFs and CRFs (9:59)*, Koller makes the connection between using a parameter prior and regularisation – regularisation can be viewed as choosing a prior distribution for the paramaters. Being a distribution, the prior distribution over the parameters in turn is governed by its own parameters, now known as *hyperparameters*. Closer inspection should show that these hyperparameters are exactly what we tune when we set regularisation parameters. But now they have physical meaning in terms of the constraints they place on the model parameters.

2 Preparation: Read sections of Barber Chapters 8 and 9 on Bayesian estimation

Read the following sections of Barber Chapter 8 about statistics for machine learning:

- Section 8.3: This section covers several distributions that are useful to know of. By now you should be able to appreciate the relevance of most of these distributions: some of these distributions are used for "normal" RVs (e.g. Bernoulli, categorical, Gaussian), whereas others are used as prior distributions over parameters (e.g. gamma, beta, Dirichlet, Laplace). For this module, we'll stick to familiar discrete and Gaussian distributions though.
- Section 8.5: This section discusses the exponential family, which is a general description of many common distributions. It also introduces the idea of a conjugate prior distribution – every likelihood that is a member of the exponential family has a conjugate prior distribution. It is often mathematically convenient to choose the form of the prior to match that of the posterior - this makes of it simply another term similar to the others in the likelihood product. This is called a conjugate prior. For categoricals (i.e., probability tables) we use the Dirichlet prior. The joint prior for the mean and precision of a 1-dimensional Gaussian is the Gaussian-gamma distribution, and for a multi-dimensional Gaussian it is the Gaussian-Wishart distribution. Take note of these concepts, but we will not use them futher in this module.
- Section 8.6: Reread this section – it is important that you understand it.
- Section 8.8: In Week 9, you would have read Subsection 8.8.1; now study Subsection 8.8.2 and read through Subsection 8.8.3.

Read the following sections of Barber Chapter 9 about learning as inference:

- Section 9.1: Here Barber introduces the perspective that learning (parameters) is just a form of inference – this idea is important to understand. You can leave out the parts about making decisions (Subsections 9.1.2 and 9.1.4).
- Section 9.2: Barber introduces ML-II here, which learns the hyperparameters of a model using MLE. Of course, nothing prevents us – in principle – from considering the hyperparameters to be random variables in their turn too, and instead tune their hyper-hyperparameters on a validation set[1]. However, we have to stop somewhere, and at this level of abstraction, we simply use MLE or choose them to be something sensible.
- Section 9.4: This section is about Bayesian parameter learning for BNs, which in essence is just inference. It is important that you understand the ideas (modelling the problem, decomposing the problem given complete data, hyperparameters), but do not get bogged down in the calculations with the beta and Dirichlet prior distributions.

[1] Turtles all the way down.

3 Practical: MAP estimation for a simple regression example

For this question, we will revisit the line fitting problem of Question 4 of Week 8, and use MAP estimation to determine the y-coordinates of the two endpoints of the line, y_I and y_F. In this part you do not necessarily have to use `emdw`'s abilities directly (although you can). You can therefore do your coding in whichever environment that suits you (e.g. Python). To help you check your work, the answer at an intermediate point is given at the end of this question in blue.

(a) (On paper) Redraw the BN for this model with M measurements using plate notation.

(b) (On paper) Choose the prior parameter distribution as $p(\boldsymbol{\Theta}) = \mathcal{N}(y_I; \mu_I, \sigma_I^2) \cdot \mathcal{N}(y_F; \mu_F, \sigma_F^2)$. Use the likelihood function $p(\mathcal{D}|\boldsymbol{\Theta})$ from Week 8 to write $\log p(\boldsymbol{\Theta}|\mathcal{D})$ as a sum of quadratic terms in terms of $\hat{y}^{[m]}$, y_I and y_F (except for a few extra terms, your answer should be the same as for Question 4(b) of Week 8).

(c) (On paper) Now derive $\boldsymbol{\Theta}^{\text{MAP}} = \arg\max_{\boldsymbol{\Theta}} \log p(\boldsymbol{\Theta}|\mathcal{D})$ using a similar procedure to that of Question 4(c) of Week 8 (you should be able to reuse much of that answer).

(d) (In code) Use the same procedure as Question 4(d) of Week 8 to generate data (or reuse the same data). Choose appropriate parameters for the prior distribution $p(\boldsymbol{\Theta})$ and calculate $\boldsymbol{\Theta}^{\text{MAP}}$ for the generated data. Plot the actual line, data points, MAP estimate as well as ML estimate on the same graph. How similar are the MAP and ML estimates?

(e) (In code) Repeat the sampling of measurements and subsequent paramater estimation, but vary the number of measurements. What is the smallest number of measurements that you need to calculate a MAP estimate? What is the effect of the number of measurements on the difference between the MAP and ML estimates?

(f) (In code) Repeat the sampling of measurements and subsequent parameter estimation, but vary the measurement noise variance σ_v^2 while keeping the number of measurements fixed. How does this parameter influence the difference between the MAP and ML estimates?

Answers:

(b)

$$\begin{aligned}\log p(\Theta|\mathcal{D}) = & -\frac{1}{2\sigma_I^2}(y_I - \mu_I)^2 - \frac{1}{2\sigma_F^2}(y_F - \mu_F)^2 \\ & + \sum_{m=1}^{M}\left[-\frac{1}{2\sigma_v^2}\left(\hat{y}^{[m]} - (1-\alpha^{[m]})y_I - \alpha^{[m]}y_F\right)^2\right] \\ & + \text{ terms that do not contain } y_I \text{ or } y_F\end{aligned}$$

4 Practical: Full Bayesian estimation for a simple regression example

Finally, let's do full Bayesian estimation of the same line fitting example we have used for MLE and MAP estimation. For full Bayesian estimation, we are interested in calculating the posterior *distribution* over the parameters y_I and y_F instead of calculating *values* for them. Specifically, we want to determine the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D}) = p\left(y_I, y_F|\hat{y}^{[1]}, \ldots, \hat{y}^{[M]}\right)$.

Although we can use `emdw` (you are welcome to try!), it is easier for this simple problem to symbolically derive an expression for the posterior distribution and then implement it in software that can perform matrix operations and visualise data. You can therefore again do your coding in whichever environment that suits you (e.g. Python). To help you check your work, the answer at an intermediate point is given at the end of this question in blue.

(a) (On paper) Redraw the BN for this model with M measurements using plate notation, but with y_I and y_F drawn as random variables, not as parameters. Now write the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$ in terms of the factors in this model.

(b) (On paper) Write each factor in the model as a canonical Gaussian factor in terms of the random variables y_I and y_F. Use the same distributions as in Question 3(b) for the prior distributions over y_I and y_F. For the factors describing the relationship between the measurement $\hat{y}^{[m]}$ and the y-coordinates of the endpoints y_I and y_F, use the measurement model as described in Question 4 of Week 8, and write it as a conditional distribution in the mean-covariance form as in Question 4 of Week 7. Then convert it into the canonical form in terms of random variables y_I and y_F using similar manipulations as for the example in Week 7's summary.

(c) (On paper) Now calculate the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$ by combining the canonical Gaussian factors of (b). The parameters of the posterior distribution should be very similar to the matrices and vectors you calculated for the MAP estimation in Question 3.

(d) (In code) Calculate the information matrix $\mathbf{K}$ and information vector $\mathbf{h}$ for the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$ using the same data and parameter choices as for Question 3. Convert the posterior distribution to the mean-covariance form by calculating the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Does this posterior distribution make sense given the data and prior distributions?

(e) (In code) Draw a number of lines $\begin{bmatrix} y_I & y_F \end{bmatrix}^T$ from the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$ and plot them with the actual line, data points, as well as the ML and MAP estimates on the same graph.

(f) (In code) Repeat the sampling of measurements and subsequent parameter estimation, but vary the number of measurements as well as the variance of the measurement noise. For which scenarios does the full Bayesian approach provide significantly more information about the parameters than ML and MAP estimation, and for which scenarios are there no significant differences?

(g) (In code) Can you retrieve the ML estimate from the posterior distribution $p(\boldsymbol{\Theta}|\mathcal{D})$? Can you retrieve the MAP estimate from the posterior distribution?

Answers:

(b)

$$p\left(y_I\right) = \mathcal{C}\left(y_I; \frac{1}{\sigma_I^2}, \frac{\mu_I}{\sigma_I^2}, -\right)$$

$$p\left(y_F\right) = \mathcal{C}\left(y_F; \frac{1}{\sigma_F^2}, \frac{\mu_F}{\sigma_F^2}, -\right)$$

$$p\left(\hat{y}^{[m]}|y_I, y_F\right) = \mathcal{C}\left(\begin{bmatrix} y_I \\ y_F \end{bmatrix}; \frac{1}{\sigma_v^2}\begin{bmatrix} \left(1-\alpha^{[m]}\right)^2 & \alpha^{[m]}\left(1-\alpha^{[m]}\right) \\ \alpha^{[m]}\left(1-\alpha^{[m]}\right) & \left(\alpha^{[m]}\right)^2 \end{bmatrix}, \frac{1}{\sigma_v^2}\begin{bmatrix} \left(1-\alpha^{[m]}\right)\hat{y}^{[m]} \\ \alpha^{[m]}\hat{y}^{[m]} \end{bmatrix}, -\right)$$

5 Evaluation: Write and submit a short report

Document the following in a short report:

- Present an overview of your derivation of Question 3(c) from the results obtained in Question 3(b), and the derivation of Question 4(c) from the results obtained in Question 4(b). Compare the results of Question 3(c) and 4(c) with that of Question 4(c) of Week 8, and discuss the similarities and differences. [4 marks]
- Present, compare and interpret the experimental results when varying the number of measurements (Question 3(e), Question 4(f), and Question 4(f) of Week 8). [3 marks]
- Present, compare and interpret the experimental results when varying the measurement noise variance (Question 3(f), Question 4(f), and Question 4(g) of Week 8). [3 marks]

Submit both the report (in PDF format) and all your code for this assignment on STEMLearn by **Sunday 27 April at 23:59**.

See the DE424 info sheet for general guidelines for the practical report. Also so the document about how to avoid academic misconduct in DE424 on STEMLearn. Specific instructions:

- The text in the report should be typed, the page limit is **5 pages**, and the report must be submitted in PDF format.
- At the top of the report, you should include a declaration that the report and the work presented in it are your own work. If you received any help or information from any source that contributed to the work or report, you should also mention the source and the extent of the help or information in the declaration.
- The report should be clear, but you do not have to spend much time on making it beautiful. You can, for instance, draw diagrams or write mathematical manipulations by hand, scan it in, and include it in your report as an image (instead of drawing diagrams in applications such as Inkscape or typing the mathematical manipulations as equations in LaTeX).

- You should give enough context in your report so that someone who is familiar with the module can understand how you have solved the problem and how well it worked without having to look at your code. Make sure that all your design decisions are clearly documented (and – if necessary – motivated). However, you should also be concise: avoid giving lengthy discussions of obvious matters, showing minute steps in mathematical manipulations, or copying large parts of your code into your report.