Bayesian Machine Learning

Instructor: Tim G. J. Rudner

Homework 2
Due: Thursday, October 6, 11:59pm via NYU Brightspace

Model Comparison, Occam's Razor, and the Laplace Approximation
(49 marks)

The *evidence* $p(\mathcal{D}|\mathcal{M})$, also known as the *marginal likelihood*, is the probability that if we were to randomly sample parameters $\boldsymbol{\theta}$ from $\mathcal{M}$ that we would create dataset $\mathcal{D}$:

$$p(\mathcal{D}|\mathcal{M}) = \int p(\mathcal{D}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta} \tag{1}$$

Simple models $\mathcal{M}$ can only generate a small number of datasets, but because the marginal likelihood must normalise, it will generate these datasets with high probability. Complex models can generate a wide range of datasets, but each with typically low probability. For a given dataset, the marginal likelihood will favour a model of more appropriate complexity, as illustrated in Figure 1.

1. (12 marks): Consider the Bayesian linear regression model,

$$y = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}, \mathbf{z}) + \epsilon(\mathbf{x}) \tag{2}$$

$$\epsilon(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2) \tag{3}$$

$$p(\mathbf{w}) = \mathcal{N}(0, \alpha^2 I) \tag{4}$$

   where the data $\mathcal{D}$ consist of N input-output pairs, $\{\mathbf{x}_i, y_i\}_{i=1}^N$, $\mathbf{w}$ is a set of linear weights which have a Gaussian prior with zero mean and covariance $\alpha^2 I$, $\mathbf{z}$ is a set of deterministic parameters of the basis functions $\boldsymbol{\phi}$, and σ^2 is the variance of additive Gaussian noise. Let $\mathbf{y} = (y_1, \ldots, y_N)^\top$ and $X = \{\mathbf{x}_i\}_{i=1}^N$.

   (a) (2 marks): Draw the directed graphical model corresponding to the joint distribution over all parameters.

   (b) (2 marks): Derive an expression for the log marginal likelihood $\log p(\mathbf{y}|\mathbf{z}, X, \alpha^2, \sigma^2)$ showing all relevant steps.

   (c) (8 marks): Derive expressions for the derivatives of this log marginal likelihood with respect to *hyperparameters* $\mathbf{z}$, α^2, and σ^2. You can make reference to matrix derivative identities.

2. (14 marks): The posterior $p(\boldsymbol{\theta}|\mathcal{M}, \mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ will often be sharply peaked around its maximum value, as in Figure 2. The evidence in Eq. (1) can thus be approximated by its height times its width $\sigma_{\theta|\mathcal{D}}$:

$$\overbrace{p(\mathcal{D}|\mathcal{M})}^{\text{evidence}} \approx \overbrace{p(\mathcal{D}|\hat{\boldsymbol{\theta}}, \mathcal{M})}^{\text{data fit}} \overbrace{p(\hat{\boldsymbol{\theta}}|\mathcal{M}) \sigma_{\theta|\mathcal{D}}}^{\text{Occam factor}} \tag{5}$$

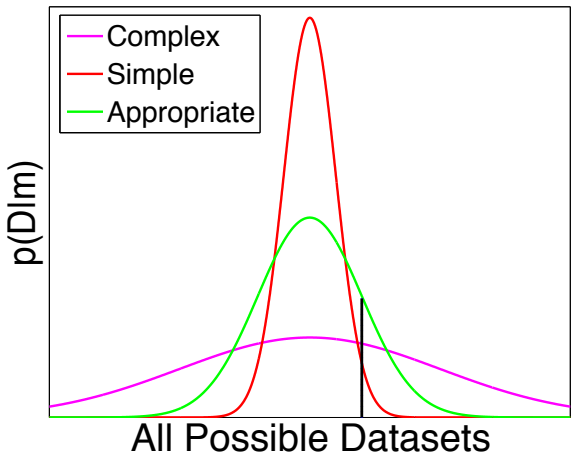

Figure 1: Bayesian Occam's Razor. The marginal likelihood (evidence) vs. all possible datasets $\mathcal{D}$. The vertical black line corresponds to an example dataset $\mathcal{D}$.

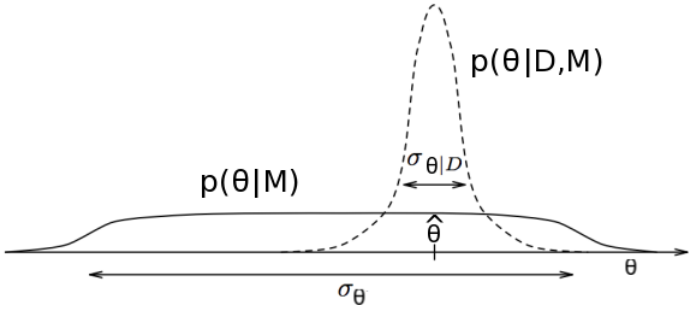

Figure 2: The posterior $p(\theta|\mathcal{D},\mathcal{M})$ and prior $p(\theta|\mathcal{M})$ over parameters θ under model $\mathcal{M}$. $\hat{\theta} = \text{argmax}_\theta p(\theta|\mathcal{D},\mathcal{M})$.

The evidence thus naturally compartmentalizes into data fit and Occam factor terms. Suppose for simplicity that the prior is uniform on a large interval such that $p(\hat{\boldsymbol{\theta}}|\mathcal{M}) = 1/\sigma_\theta$. The Occam's factor then becomes $\frac{\sigma_{\theta|\mathcal{D}}}{\sigma_\theta}$.

(a) (2 marks): Provide an interpretation of the Occam's factor $\frac{\sigma_{\theta|\mathcal{D}}}{\sigma_\theta}$, wrt Figure 1.

(b) (4 marks): Show that if we use Laplace's method to approximate the posterior $p(\boldsymbol{\theta}|\mathcal{M},\mathcal{D})$ as a Gaussian, then the Occam's factor becomes $p(\hat{\boldsymbol{\theta}}|\mathcal{M})\det(A)^{-1/2}$ where $A = -\nabla\nabla \log p(\boldsymbol{\theta}|\mathcal{D},\mathcal{M})$. Use this expression to interpret each of the terms in the log marginal likelihood you derived for question 1(b).

(c) (6 marks): Derive an approximation for the log evidence $\log p(D|\mathcal{M})$ assuming a broad Gaussian prior distribution and iid observations, strictly in terms of the number of datapoints N, the number of parameters m (dimensionality of $\boldsymbol{\theta}$), and $\log p(D|\hat{\boldsymbol{\theta}})$. Show all of your work.

(d) (2 marks): Relate the Hessian A to the covariance matrix of a Gaussian prior over parameters.

3. (33 marks): Load the datasets $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively from `occam1.mat` and `occam2.mat` in the assignment files `a2files.zip`. Suppose we are considering three models to explain the data:

   (i) $\mathcal{M}_1$: The Bayesian basis regression model of Eq.(2)-(4), but with

   $$\boldsymbol{\phi}(x, \mathbf{z}) = \boldsymbol{\phi}(x) = (1, x, x^2, x^3, x^4, x^5)^\top \tag{6}$$

   (ii) $\mathcal{M}_2$: The Bayesian basis regression model of Eq.(2)-(4), but with

   $$\boldsymbol{\phi}(x, \mathbf{z}) = \left(\exp[-\frac{(x-1)^2}{z_1^2}], \exp[-\frac{(x-5)^2}{z_2^2}]\right)^\top, \quad \mathbf{z} = (z_1, z_2)^\top \tag{7}$$

   (iii) $\mathcal{M}_3$: The Bayesian basis regression model of Eq.(2)-(4), but with

   $$\boldsymbol{\phi}(x, \mathbf{z}) = \boldsymbol{\phi}(x) = (x, \cos(2x))^\top \tag{8}$$

   Parts of this question involve coding. Please hand in the Matlab, Octave, or Python code you used to solve this question, along with the plots of results generated by the code used to answer the questions. This code should be succinct and include comments. Your code should not exceed 5 pages in length. Answer all questions for both $\mathcal{D}_1$ and $\mathcal{D}_2$ unless the question explicitly states otherwise.

   (a) (20 marks): Using your work from question 1, write code to plot a histogram of the evidence for each of these three models, conditioned on the maximum marginal likelihood values of all the hyperparameters $\mathbf{z}$, α^2, and σ^2. To find these values, jointly optimize the log marginal likelihood with respect to these hyperparameters using a quasi-Newton method or non-linear conjugate gradients. Based on the histogram, which hypotheses do you believe generated the data?

   Hint 1: If you compute the Cholesky decomposition of $A = R^\top R$, where R is an upper right triangular matrix, then $A^{-1}\mathbf{b} = R^{-1}(R^{-1})^\top \mathbf{b}$. You can use this decomposition in conjunction with the helper function `solve_chol.m` for numerically stable solutions to linear systems involving A. Note also that $\log\det(A) = 2\sum_i \log(R_{ii})$.

   Hint 2: Use the function `checkgrad.m`, or compute a finite difference scheme, to numerically check the derivatives of the log marginal likelihood with respect to the model hyperparameters, as a means to debug and to check your answer for 1(c).

   Hint 3: Some of the relevant matrices may be very poorly conditioned. It is often useful to add a constant small amount of *jitter* to the diagonal of these matrices, $A \to A + \epsilon I$, where ϵ is on the order of 10^{-6}, before performing operations such as Cholesky decompositions. This procedure is advocated by Radford Neal in his 1996 PhD thesis, *Bayesian Learning for Neural Networks*.

   Hint 4: You may wish to constrain your hyperparameters to be positive, but perform an unconstrained optimization. To do so, you can optimize over the log hyperparameters in an unconstrained space. When computing derivatives of the log marginal likelihood with respect to the log parameters, you will find it helpful to use the chain rule.

   Hint 5: I have included `minimize.m`, a robust implementation of non-linear conjugate gradients you can use to optimize the marginal likelihood with respect to hyperparameters. You are, however, free to use another gradient based optimizer if you wish.

(b) (8 marks): Explain why the evidence histogram might disagree with the maximum likelihood ranking of models (in general and with respect to $\mathcal{D}_1$ and $\mathcal{D}_2$).

(c) (2 marks): Give the posterior mean and variance over the parameters of the model with highest evidence on $\mathcal{D}_2$.

(d) (3 marks): What values of the *hyperparameters* maximized the marginal likelihood?

(e) Optional (1 bonus mark): Plot the posterior over each set of model weights $\mathbf{w}$ (using extra coding space if required) for each dataset.

Markov chain Monte Carlo
(15 marks)

Follow Iain Murray's MCMC practical at
`http://homepages.inf.ed.ac.uk/imurray2/teaching/09mlss/handout.pdf`

Complete section 4, and answer the MCMC questions in section 5.

Hand in (1) your code for section 4, and (2) your answers to the 5 MCMC questions. The code is worth 10 marks, and the questions are worth 1 mark each. There are thus 15 marks for this part of the assignment.