MLPR Assignment Part 2: General comments

Below are comments on some recurring issues I came across when marking the second part of the assignment, which I thought it would help to explain in more detail than was possible in the individual feedback sheets.

1 Using standard errors when comparing estimated values

When evaluating the performance of models and comparing them you were asked to compute estimates of two performance metrics, the classification accuracy and mean log probability of the true y labels. The first time you were asked to do this you were also explicitly instructed to calculate the standard error values of your mean estimates[1] which give an indication of the uncertainty in our estimates of these quantities due to calculating them with a finite subset of the assumed infinite possible set of examples.

As well as calculating and reporting the standard error values it was important to bear them in mind when making comparisons between the point estimates. A lot of people despite correctly calculating the point estimates and standard error values then went on to make comparisons only considering the point estimates in isolation, and therefore made statements along the line 'the performance of classifier X is better than classifier Y on metric Z' despite the difference between the point estimates being a lot smaller in magnitude than the standard errors on those estimates, and so the values reported being insufficient to strongly support such a conclusion.

Another common mistake was to view the standard error values (often in this case calculated incorrectly and/or reported only for one of the two performance metrics) as being some sort of direct indicator of the classification performance of the models in themselves and making statements along the lines of 'the standard error of classifier X is lower than that of classifier Y so classifier X has the better performance'. This did not make sense because as described above the standard error values give an indication of the uncertainty in the estimates of the performance metrics, they should not be compared in isolation from the estimate they refer to.

2 Standard error versus standard deviation

A common error when calculating the standard error of a mean was just to report the sample standard deviation of the values used to compute the mean. The standard error is this value divided by the square root of the number of values. While the standard deviation is a measure of the spread of the distribution of the values themselves, the standard error is a measure of the spread of the distribution of *mean estimates* calculated from these values. Intuitively as we get more data points we expect our mean estimate to become less uncertain and so the standard error should reflect this.

There was also some confusion in the opposite direction. In question 3b you were asked to compare your fitted value of the noise level ϵ to your scatter plot of posterior samples of ϵ against $\log \lambda$ and state whether the fitted value seemed reasonable. Some people here estimated the posterior mean and calculated its standard error and commented that the fitted value of ϵ was outside these error bars. Here we did not expect the maximum likelihood value of ϵ to necessarily exactly coincide with the posterior mean (which will also depend on the choice of prior) - the question was intended instead to ask if the fitted value of ϵ seemed reasonable under the posterior *distribution*. It would therefore be more relevant here to estimate the standard deviation of the marginal posterior on ϵ and check how many standard deviations the fitted value of ϵ was from the posterior mean. It was also fine here to just eyeball the plot however to establish a rough plausible range for ϵ under the posterior.

[1] While this was only explicitly asked for in part 2 question 1b, it was a good idea to report these values whenever calculating such estimates and if you sensibly encapsulated the code you used to calculate the mean and standard error from a finite set values into a function, or used the `errorbar_str` provided as part of the solutions to tutorial 5, this should have required no extra effort.

3 Checking your answers make sense

It is always worth checking that the value you calculate for a specific quantity is consistent with what you know about that quantity. For example sometimes quantities can only take values in a certain range. Here the accuracy estimates (if expressed as a proportion) should always have been between 0 and 1. The mean log probability of true label estimates could only ever take non-positive values - the probabilities of binary labels can only take values between 0 and 1 and so the logarithm of these is less than or equal to zero. Therefore if you calculate a negative accuracy, an accuracy more than one or a positive mean log probability (all these were reported) this should be an indication to you that you have made some error in your calculation.

Also try to check if the values you calculate seems sensible given what they are measuring. If you are computing the accuracy for a model fitted to predict binary labels it should be quite surprising if the accuracy is significantly below 50% - this would mean the model is worse than randomly guessing and we would actually get an improvement by predicting the opposite of what it says.

4 Definitions of likelihood, prior and posterior

In question 3 on the hierarchical model there was a lot of confusion between the differing meanings of likelihood, log-likelihood, prior and posterior.

If we have a model with parameters $\boldsymbol{\theta}$ of N i.i.d. data points then the likelihood will generally refer to the *joint probability of all the data points* given a set of model parameters

$$\text{likelihood}(\boldsymbol{\theta}) = \mathbb{P}\left[\text{all data points} \,|\, \boldsymbol{\theta}\right] = \prod_{i=1}^{N} \mathbb{P}\left[\text{data point } i \,|\, \boldsymbol{\theta}\right]$$

The log-likelihood is then

$$\text{log-likelihood}(\boldsymbol{\theta}) = \log \mathbb{P}\left[\text{all data points} \,|\, \boldsymbol{\theta}\right] = \sum_{i=1}^{N} \log \mathbb{P}\left[\text{data point } i \,|\, \boldsymbol{\theta}\right]$$

A common mistake was to neglect to take logarithms of each probability when deriving expressions for the log-likelihood in question 3a and just summing up the probabilities of each data point given the parameters. A number of people also expressed the likelihood as the probability of only one data point given parameters - in general this is a valid interpretation of likelihood however here you were specifically asked about the (log-)likelihood of a model on N data points and more importantly $\mathcal{L}(\boldsymbol{w}, \epsilon)$ had specifically been defined as the log-likelihood all data points so it was not valid to then reuse this notation to refer to log-likelihood of a single data point (and then go on to use this in the expression of the log-posterior which some did).

Another common mistake was to confuse the log-posterior and log-likelihood, with a number of people giving expressions in the first two sections of question 3a for what they described as log-likelihood which included terms involving λ which only appeared in the posterior and prior not the likelihood.

There was also some confusion between the posterior and prior distributions when answering question 3b. Here you were asked to state if our posterior beliefs about $\log \lambda$ and ϵ are independent. A number of people assumed that as our prior beliefs on $\log \lambda$ and ϵ are independent then the answer to this question must be that they are independent. However the posterior is conditioned on the data and under the posterior marginal distribution on $\log \lambda$ and ϵ the two variables are not independent. This is due there being dependencies between the weights and both $\log \lambda$ and ϵ in the joint posterior. Therefore is we do not fix the weights then $\log \lambda$ and ϵ depend on each other via the weights. This was visible as a negative correlation in the scatter plot you were asked to produce of the samples of $\log \lambda$ and ϵ.

5 Correlations

Some people incorrectly stated that the negative correlation evident between $\log \lambda$ and ϵ in the scatter plot of MCMC samples from the posterior was due to the correlation between successive samples when using a MCMC method. While it is correct to say that generally successive samples generated using a MCMC method will be correlated (often highly so), this is distinct from whether the variables of the target distribution we are trying to sample from have correlations between them.

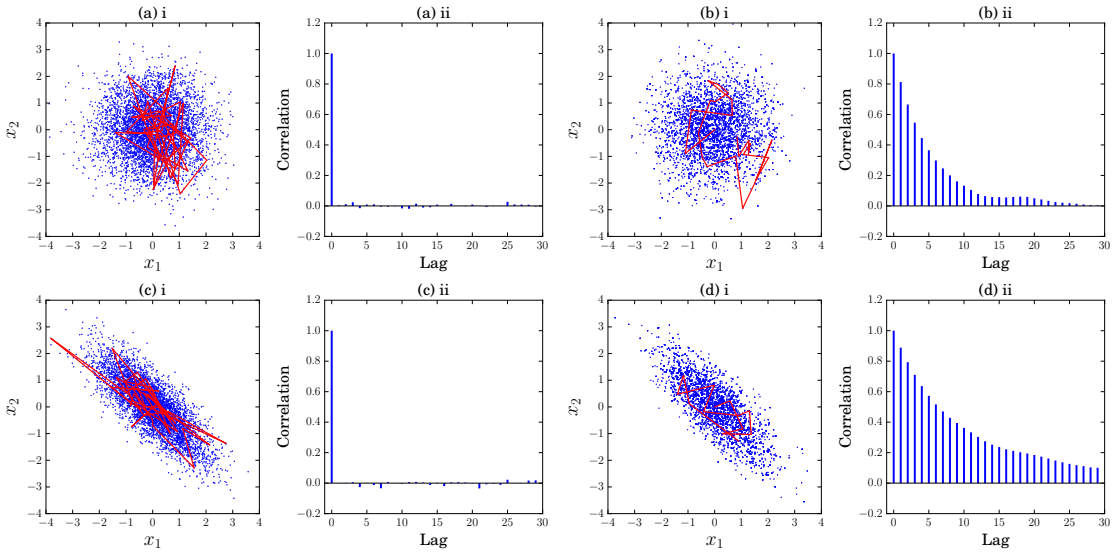

Figure 1: 2D sample scatter plots and between sample correlation plots for sets of independent and Metropolis random walked based samples from two zero-mean Gaussian distributions. (a) i. independent samples of a pair of zero mean unit variance independent Gaussian variables. (a) ii. correlations between successive x_1 samples from (a) i. at different lags. (b) i. Random-walk Metropolis samples of a pair of zero mean unit variance independent Gaussian variables. (b) ii. correlations between successive x_1 samples from (b) i. at different lags. (c) i. independent samples of a pair of zero mean unit variance Gaussian variables with correlation $\Sigma_{12} = -0.8$. (c) ii. correlations between successive x_1 samples from (c) i. at different lags. (d) i. Random-walk Metropolis samples of a pair of zero mean unit variance Gaussian variables with correlation $\Sigma_{12} = -0.8$. (d) ii. correlations between successive x_1 samples from (d) i. at different lags. In the sample plots blue points are individual samples while the red paths are the trajectories of set of 40 successive samples (including rejected updates for Metropolis samples).

This is illustrated in figure 1. Here we have plots visualising the samples from two different zero-mean bivariate Gaussian distributions - the top row for a bivariate Gaussian with covariance $\Sigma_{11} = \Sigma_{22} = 1 \;\; \Sigma_{12} = 0$ i.e. independent variables, and the bottom row for a bivariate Gaussian with covariance $\Sigma_{11} = \Sigma_{22} = 1 \;\; \Sigma_{12} = -0.8$ i.e. negatively correlated variables. The left two columns are based on samples drawn independently from the respective distributions. The right two columns are based on samples generated using a MCMC method (random-walk Metropolis updates with proposal width 1). For each distribution / sampling method combination, both a scatter plot of samples is shown (i.) and a plot of the correlations measured between successive x_1 samples at different lags (ii.).

It can be seen here in panels (b) i. and (b) ii. that we can have a target distribution in which our beliefs about the variables are independent, but that successive samples from the MCMC method used are highly correlated. Conversely in panels (c) i and (c) ii we can see an example of a distribution in which we our beliefs about the variables are highly correlated but successive samples are not at all correlated. What we can see from panel (d) ii. however is that strong correlations between variables in the target distribution can lead to the successive MCMC samples

being more correlated (compare the decay of the correlations in (d) ii. to (b) ii.) - this is due to the random-walk Metropolis MCMC dynamic here being not very well suited to highly correlated distributions.

Another correlation related issue that a few people ran into was the difference between correlation and statistical dependence. It was correctly pointed out by several students that if two variables are uncorrelated this does not necessarily mean they are statistically independent (unless they are jointly Gaussian). This caused some people to decide to try to assess if there was any dependence between $\log\lambda$ and ϵ under the posterior in question 3b by estimating the mutual information between them, which is a more general measure of the statistical dependence between two random variables. This is a valid approach to take however there are several caveats that need to be born in mind. Unlike correlation, estimating the mutual information between a pair of continuous random variables from a finite set of samples if very difficult, with estimators tending to be either very noisy or biased. Mutual information estimates are also a lot more expensive to compute than correlations and more care needs to be taken with their interpretation - ideally we would also calculate error bars on our estimate to allow us to decide more rigorously if the mutual information appears to be distinct from zero or not, it is not immediately obvious what a typical mutual information value between two dependent variables would be. Although a correlation coefficient estimate may be insufficient to declare two variables independent if it is indistinguishable from zero, if it is significantly different from zero this is sufficient to show statistical dependence between the two variables without the complications of invoking mutual information.

6 Transforming variables in probability density functions

A relatively subtle error that was made by a number of people in question 3b when performing the slice sampling was to apply a non-linear transformation to the variables the posterior density was defined on without accounting for this change of variables correctly. In particular several people defined their log-posterior function in terms of λ rather than $\log\lambda$ and some used $a = \sigma^{-1}(a)$ instead of ϵ (`slice_sample` expects that the arguments used in the function handle specified to calculate the target density are the parameters the density is defined on).

The correction needed involves calculating the determinant of the Jacobian (matrix of first partial derivatives) of the transformation - there was an example of applying this in the Background Work tutorial sheet. What follows is a slightly mathematically involved explanation of why we need to account for a change of variables like this applied specifically to the problem here.

In question 3b you were asked to sample from a posterior distribution defined by a probability density function, specifically

$$p(\boldsymbol{w}, \epsilon, \log\lambda \,|\, \text{data}) = \frac{1}{Z} \exp\left\{\mathcal{L}(\boldsymbol{w}, \epsilon)\right\} \lambda^{\frac{D}{2}} \exp\left\{-\lambda \boldsymbol{w}^{\mathrm{T}} \boldsymbol{w}\right\}$$

with Z being an (unknown) normalising constant. A probability density function implicitly defines probabilities via an integral over regions of the space the density is defined on

$$\mathbb{P}\left[\widetilde{\boldsymbol{w}} \in \mathcal{A}, \widetilde{\epsilon} \in \mathcal{B}, \widetilde{\ell} \in \mathcal{C} \,|\, \text{data}\right] = \int_{\boldsymbol{w}\in\mathcal{A}} \int_{\epsilon\in\mathcal{B}} \int_{\ell\in\mathcal{C}} p(\boldsymbol{w}, \epsilon, \ell \,|\, \text{data}) \; \mathrm{d}\boldsymbol{w} \, \mathrm{d}\epsilon \, \mathrm{d}\ell$$

here the (non-standard) notation is being used that characters with tildes ($\sim$) above them represent random variables and for the sake of clarity ℓ has been used to denote $\log\lambda$.

If we want to redefine the variables in a probability density and ensure that the probabilities that the random variables take values in arbitrary ranges stay the same we need to apply the usual change of variables formula used when making a substitution in an integral. For instance if we want to make the change of variables $\lambda = \exp(\ell) \Rightarrow \ell = \log\lambda$ inside the integral first we need to calculate the Jacobian of this transform

$$\frac{\partial \ell}{\partial \lambda} = \frac{1}{\lambda}$$

and we also need to find the region in λ space to which corresponds to the region in ℓ space being integrated over

$$\mathcal{C}' = \{\exp\ell : \ell \in \mathcal{C}\}$$

This gives us that

$$\begin{aligned}
\mathbb{P}\left[\widetilde{\boldsymbol{w}} \in \mathcal{A}, \widetilde{\epsilon} \in \mathcal{B}, \widetilde{\lambda} \in \mathcal{C}' \,|\, \text{data}\right] &= \int_{\boldsymbol{w}\in\mathcal{A}} \int_{\epsilon\in\mathcal{B}} \int_{\lambda\in\mathcal{C}'} p(\boldsymbol{w}, \epsilon, \log\lambda \,|\, \text{data}) \left|\frac{\partial \ell}{\partial \lambda}\right| \,\mathrm{d}\boldsymbol{w}\,\mathrm{d}\epsilon\,\mathrm{d}\lambda \\
&= \int_{\boldsymbol{w}\in\mathcal{A}} \int_{\epsilon\in\mathcal{B}} \int_{\lambda\in\mathcal{C}'} \frac{1}{Z} \exp\left\{\mathcal{L}(\boldsymbol{w}, \epsilon)\right\} \lambda^{\frac{D}{2}} \exp\left\{-\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right\} \frac{1}{\lambda} \,\mathrm{d}\boldsymbol{w}\,\mathrm{d}\epsilon\,\mathrm{d}\lambda \\
&= \int_{\boldsymbol{w}\in\mathcal{A}} \int_{\epsilon\in\mathcal{B}} \int_{\lambda\in\mathcal{C}'} \underbrace{\frac{1}{Z} \exp\left\{\mathcal{L}(\boldsymbol{w}, \epsilon)\right\} \lambda^{\frac{D-2}{2}} \exp\left\{-\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right\}}_{p'(\boldsymbol{w},\epsilon,\lambda \,|\, \text{data})} \,\mathrm{d}\boldsymbol{w}\,\mathrm{d}\epsilon\,\mathrm{d}\lambda
\end{aligned}$$

and so we get that the new probability density function parametrised in terms of λ as opposed to $\log\lambda$ is

$$p'(\boldsymbol{w}, \epsilon, \lambda \,|\, \text{data}) = \frac{1}{Z} \exp\left\{\mathcal{L}(\boldsymbol{w}, \epsilon)\right\} \lambda^{\frac{D-2}{2}} \exp\left\{-\lambda \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}\right\}$$

Note that if $D \gg 1$ and so $\frac{D-2}{2} \approx \frac{D}{2}$ that this redefined density will be very similar to incorrect density found simply by swapping λ for $\log\lambda$ in the density above. This is why those who made an incorrect change of variables like this may have found their scatter plot looks visually indistinguishable from the correct version.

An intuition here is that not correctly accounting for the change of variables by including the Jacobian term corresponds to using a different prior density on a parameter - in this case the prior on $\log\lambda$ has a very weak effect on the posterior density and the change to a different prior makes little difference. In general however not accounting for change of variables properly can lead to quite significant changes in the density being sampled from and so it is vital to always correctly apply the change of variables formula when redefining the variables in a density.