MODEL SOLUTIONS

SETTER: Trevor Cohn and Neil Lawrence

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2013–2014

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE 2 hours

Answer THREE of the four questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

1. This question concerns general concepts in machine learning.

a) Overfitting is a common problem in machine learning, in both regression and classification.

(i) Explain the problem, when it might occur, and how you can measure the extent of overfitting. [10%]

(ii) With reference to the regression or classification models you have studied, describe two different regularisation techniques for addressing the overfitting problem. Provide a sentence or two on each, explaining how they limit overfitting. [15%]

ANSWER:

1. Overfitting is where we model the training set so accurately that generalisation accuracy is compromised. Effectively we end up modelling idiosyncrasies in the training data (e.g., noise) rather than the underlying signal, which are what matters when applying the model to unseen data. Typically overfitting occurs when the model has more parameters than there are data points, e.g., fitting a nth order polynomial to n points. Overfitting can be measured using held-out validation, that is, by measuring the accuracy on unseen data. If this differs greatly from training accuracy, this suggests overfitting (but could be other effects, e.g., non-iid).

2. Using a prior is one method, either as part of a MAP estimate or in fully Bayesian inference. This biases the posterior away from extreme solutions. E.g., in logistic regression, a Gaussian can penalise high magnitude weights, in regression the noise model can allow for loose data fit, in the Bayes classifier a prior over likelihood terms is often used to avoid problems of zeros.

Other techniques are early stopping in iteratively trained models, such as the perceptron (stopping early stops the model being fit properly, hoping that overfitting is limited to the final stages). There's also margin maximisation in the SVM, regularising the model behaviour on linearly separable data, and allowing margin violations in the soft-margin SVM, such that the margin criteria can be prioritised over training accuracy. Finally setting the k parameter in k-nearest neighbour allows for smoother decision surfaces (with larger k), thus reducing training fit (1-NN has a perfect training fit) to allow for better generalisation.

b) Model training for probabilistic models often involves taking point estimates for the model parameters, such as the *maximum a posteriori* (MAP) estimate.

(i) Define the objective the MAP is optimising, making reference to Bayes' rule. [10%]

(ii) *Bayesian inference* is an alternative inference technique which also makes use of a prior. Outline the Bayesian inference technique, and contrast it with the use of point estimates. [20%]

(iii) Why might you choose to use Bayesian inference instead of a point estimate, or vice-versa? In what circumstances would Bayesian inference be preferable?

[15%]

ANSWER:

1. The MAP estimate is optimising the log-posterior of the training data. The posterior is

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{P(D)}$$

using Bayes' rule, where θ are the model parameters and D is the training data. The log posterior is

$$\log p(\theta|D) = \log p(D|\theta) + \log p(\theta) - \log P(D)$$

and as we're only interested in maximising this wrt θ, then the last constant term is omitted.

2. Using a MAP estimate means solving for θ (the mode of the posterior), then using this to make predictions. In contrast Bayesian inference involves reasoning under the full posterior in making predictions, by integrating out θ. In other words, we compute the expectation of the prediction under the posterior. I.e.,

$$p(y_{test}|\mathbf{x}_{test}, D) = \int p(y_{test}|\mathbf{x}_{test}, \theta)p(\theta|D)d\theta$$

where we consider the predictions, y_{test}, for a test example $\mathbf{x}_{test}$.

3. The Bayesian approach is justified if there is uncertainty, i.e., the posterior is not sharply peaked. Otherwise (a spiked posterior) the two approaches are equivalent. More concretely, we tend to have more uncertainty when the training set is small, or when the modelling assumption are invalid. In either case, Bayesian inference is preferable. A caveat is that Bayesian inference can be more complex for some models, and is often more computationally demanding.

c) Several pairs of distributions are said to be *conjugate*, such as the Binomial and Beta; Multinomial and Dirichlet; and Normal (mean) and Normal. Explain the notion of *conjugacy*, and how this might be practically important in a classification or regression scenario. [15%]

ANSWER:

Conjugacy means that when the likelihood and prior distributions are combined to form the posterior (i.e., applying Bayes rule), the resulting distribution is in the same family as the prior. This means that we can perform exact Bayesian inference, by working with the posterior distribution. This is important in classification and regression, as we often use Binomial, Normal and other distributions and often want to include a prior to represent our initial beliefs and avoid pathological behaviours. Moreover, we gain the ability to reason under uncertainty.

d) Logistic regression is a probabilistic model for binary classification. It defines the probability of class 1 as

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}$$

and $p(\mathcal{C}_1|\mathbf{x}) = 1 - p(\mathcal{C}_2|\mathbf{x})$. Show how this gives rise to a linear discriminant function. You may want to start by formulating the log-odds ratio for predicting $\mathcal{C}_1$ versus $\mathcal{C}_2$.

[15%]

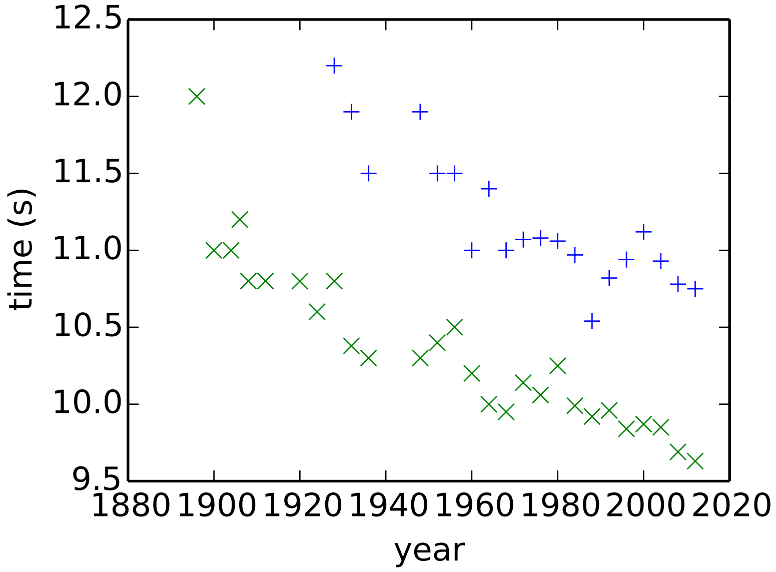
ANSWER:

Start by defining the discriminant $y(\mathbf{x})$ as the log-odds ratio, such that $y(\mathbf{x}) > 0$ corresponds to $\mathcal{C}_1$ being more likely than $\mathcal{C}_2$, and vice-versa. Then simplify as follows:

$$\begin{aligned} y(\mathbf{x}) = \log\frac{p(\mathcal{C}_1|\mathbf{x})}{p(\mathcal{C}_2|\mathbf{x})} &= \log\frac{\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}}{1-\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}} \\ &= \log\frac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})-1} \\ &= \log\frac{1}{\exp(-\mathbf{w}^\top\mathbf{x})} \\ &= \mathbf{w}^\top\mathbf{x} \end{aligned}$$

The final result defines a linear discriminant, $y(\mathbf{x}) = \mathbf{w}^\top\mathbf{x}$. QED.

2. This question is based on classifying the gender of Olympic 100m sprint winners. Shown below is a plot of the winning times in each year for the women's and men's events.



The women's results are shown with plus symbols $(+)$ and the men's results with crosses (x). We would now like to develop a classifier to predict the gender ($t =$ male or female) automatically based on a two dimensional data point, $\mathbf{x} =$ (year, winning time). Note that we are seeking to model this as a classification dataset, **not** regression as used for your class work.

a) We decide to model this data using a linear binary classifier.

(i) Draw a rough diagram to illustrate a decision boundary that you might hope to learn from this data with a linear classifier. Label the regions corresponding to the two classes, $\mathcal{C}_1 =$ female, $\mathcal{C}_2 =$ male. [10%]

(ii) Is a linear classifier an appropriate choice for this data? You should consider both *interpolation* and *extrapolation* settings, and explain these terms in your answer. [15%]

ANSWER:

1. They should replicate the plot (coarsely) and draw a diagonal straight line from top-left to bottom right such that the crosses are below and the pluses are all above the line. The region above should be labelled female (C_1) and below male (C_2).

2. Yes and no. Yes in the interpolation setting, where we consider points only in the range of our training data, namely from years 1888–2012 and times 9.5–12 seconds. This is because the data is clearly linearly separable. Outside these ranges, i.e., the extrapolation setting, we would expect the linear model to fall down. I.e., the problem becomes non-linearly separable. This is because running times cannot continue to improve linearly, otherwise the event would be run in 0 or negative time. Using this as a threshold for assessing gender is clearly unsound. We would need some kind of non-linear method capable of representing an asymptotic decline.

b) Basis functions can be used to develop non-linear models. Give an example of a basis function that would be appropriate for this dataset, and explain why. Based on your choice, state the basis vector $\phi(\mathbf{x})$ used to represent a data-point $\mathbf{x}$. [15%]

ANSWER:

There are a few possible answers here. RBFs are ok, but they're best suited for interpolation, for which we already have a reasonable solution (linear). A polynomial would be better, or better yet, a tailored function based on a geometric decay with time such as $\phi(x) = [e^{x_{year}}, x_{time}, 1]$. (You might need some more degrees of freedom to control the rate of decay etc.)

For an RBF, the datapoint would be represented as $\phi(\mathbf{x}) = [N(\mathbf{x}, \mu_1), N(\mathbf{x}, \mu_2), \ldots]$ where N is the normal density with a given wdth σ, and there are several RBFs with centres μ_1, μ_2 etc which are points in 2D (time, year) space.

For a quadratic, $\phi(\mathbf{x}) = [1, x_1, x_1^2, x_2, x_2^2, x_1x_2]$ where the last entry is optional. Higher order polynomials follow.

c) Various techniques can be used for *validation*, such as using a fixed held-out validation set, or cross-validation.

(i) Explain the purpose of validation. Define fixed held-out validation and cross-validation. [10%]

(ii) Imagine we were to perform k-fold cross-validation on this data. We do so by assigning the data points from the 100m dataset based on the year of each race, such that each 20 year period forms a fold. Explain why this form of evaluation might not give a reliable estimate of the generalisation error, and how you might fix this. [15%]

ANSWER:

1. Validation measures the error of the model on unseen data, i.e., estimates its generalisation ability. A fixed set reserves some data for evaluation, that is not used in training. The cross-fold technique splits the data into k equal parts 'folds' and then uses $k-1$ folds for training and 1 for testing, which is repeated k times such that each fold is used as the test set once. The error of the classifier is then reported on the testing folds aggregated together.

2. Assigning the data to folds in such a way would mix evaluation of extrapoliation and interpolation settings. In some cases data points from events before and after will be available, at other times only events before or only after are available. It is unlikely that we want this arbitrary mixing of applications, i.e., this would not be considered a fair evaluation. Instead we might train only on the earlier folds ($< k$) when testing on k (to target extrapolation to the future). Or randomly assign data to folds, to test the easier interpolation setting. A second less important issue is that the number of women events in the first fold would be zero, and may be difficult to learn to evaluation on. Random assignment or stratified cross-validation might help

here.

d) Kernel methods extend the basis functions technique to allow for richer and more flexible data representations.

(i) Explain what is meant by a *kernel function*, and how they relate to basis functions. [10%]

ANSWER:

A kernel is a function that compares two training instances, and provides a measure their similarity. A kernel is equivalent to an inner product under some basis.

(ii) Using a linear perceptron with basis function ϕ, the perceptron update takes the form

$$\mathbf{w} \leftarrow \mathbf{w} + \eta t_i \phi(\mathbf{x}_i)$$

after an error is made on the i^{th} training instance. Show how the weights can be represented as a linear combination of the training samples,

$$\mathbf{w} = \sum_i \alpha_i t_i \phi(\mathbf{x}_i)$$

and show the dual form of the update rule, in terms of α. [15%]

ANSWER:

If the weights start as zero, then $\alpha = 0$: makes the two representations equivalent. Thereafter an update changes the weights as follows

$$\begin{aligned}\mathbf{w} &\leftarrow \mathbf{w} + \eta t_i \phi(\mathbf{x}_i) \\ &= \sum_j \alpha_j t_j \phi(\mathbf{x}_j) + \eta t_i \phi(\mathbf{x}_i) \\ &= \sum_j \left(\alpha_j + \eta\delta(i,j)\right) t_j \phi(\mathbf{x}_j)\end{aligned}$$

where $\delta(i,j)$ is the Kronecker delta function. From this we can see that the weights will always remain in the dual form, as a linear combination of the training samples.

Consequently the update rule becomes $\alpha_i \leftarrow \alpha_i + \eta$.

(iii) Using the above reparameterisation, derive the kernel perceptron. This requires you to prove that the perceptron discriminant function $y(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})$ can be expressed such that the basis functions occur solely as inner products $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$. Justify why this is important. [10%]

ANSWER:

Given $\mathbf{w} = \sum_i \alpha_i t_i \phi(\mathbf{x}_i)$, the discriminant is now:

$$\begin{aligned} y(\mathbf{x}) &= \left(\sum_i \alpha_i t_i \phi(\mathbf{x}_i) \right)^T \phi(\mathbf{x}) \\ &= \sum_i \alpha_i t_i \phi(\mathbf{x}_i)^T \phi(\mathbf{x}) \end{aligned}$$

Note the last factor, $\phi(\mathbf{x}_i)^T \phi(\mathbf{x})$, which is paired as required. The importance of this is that we can replace these calls with a kernel function $k(\mathbf{x}_i, \mathbf{x})$ which allows for reasoning without explictly calculating the basis representation (e.g., which may be infinite, or otherwise untractable).

3. This question concerns regression and maximum likelihood fits of regression models with basis functions.

   a) What role do the basis functions play in a regression model? [10%]

ANSWER:

Basis functions turn linear models into non linear models.

   b) The polynomial basis with degree d computed for a one dimensional input has the form

$$\boldsymbol{\phi}(x_i) = \left[1 \; x_i \; x_i^2 \; x_i^3 \; \dots \; x_i^d\right]^\top .$$

Give a disadvantage of the polynomial basis. Suggest a potential fix for this disadvantage and propose an alternative basis. [20%]

ANSWER:

For large inputs, like in the olympic data we studied in class, the entries of the basis vector will become very large, leading to numerical problems. One fix for this is to map the inputs between -1 and 1. That prevents this happening. A basis that doesn't have this problem is the radial basis, where each basis is given by a exponentiated quadratic form, centred at different locations.

   c) The likelihood of a single data point in a regression model is given by,

$$p(y_i|\mathbf{w}, \mathbf{x}_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}\right).$$

Assuming that each data point is independent and identically distributed, derive a suitable *error function* that should be minimized to recover $\mathbf{w}$ and σ^2. Explain your reasoning at each step. [25%]

ANSWER:

The independence assumption means that

$$\begin{aligned} p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) =& \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}\right). \\ =& \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left(-\frac{\sum_{i=1}^{N}(y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}\right). \end{aligned}$$

The logarithm is a monotonic function. This allows us to apply it to the likelihood and maximize the log likleihood.

$$\log p(\mathbf{y}|\mathbf{w}, \mathbf{X}, \sigma^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{N}(y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}.$$

When optimizing, we are able to drop the first term involving π because it was constant in $\mathbf{w}$ and σ^2. Finally, by convention we minimize in optimization, so we take the negative log likelihood and find the error as:

$$E(\sigma^2, \mathbf{w}) = \frac{N}{2}\log(\sigma^2) + \frac{\sum_{i=1}^N (y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}.$$

d) Now show that this error function is minimized with respect to the vector $\mathbf{w}$, at the following point,

$$\mathbf{w}^* = \left[\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

where $\boldsymbol{\Phi}$ is a *design matrix* containing all the basis vectors, and $\mathbf{y}$ is a vector of regression targets. [30%]

ANSWER:

Since we are only looking at $\mathbf{w}$, we can ignore terms associated with σ^2. First multiply out the brackets:

$$E(\mathbf{w}) = \frac{1}{2\sigma^2}\sum_{i=1}^N y_i^2 - \frac{1}{\sigma^2}\mathbf{w}^\top \sum_{i=1}^N \boldsymbol{\phi}(\mathbf{x}_i) y_i + \frac{1}{2\sigma^2}\mathbf{w}^\top \sum_{i=1}^N \boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w}$$

Now take derivatives with respect to $\mathbf{w}$

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -\frac{1}{\sigma^2}\sum_{i=1}^N \boldsymbol{\phi}(\mathbf{x}_i) y_i + \frac{1}{\sigma^2}\sum_{i=1}^N \boldsymbol{\phi}(\mathbf{x}_i)\boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w}$$

this can be rewritten in matrix form as

$$\frac{\mathrm{d}E(\mathbf{w})}{\mathrm{d}\mathbf{w}} = -\frac{1}{\sigma^2}\boldsymbol{\Phi}^\top \mathbf{y} + \frac{1}{\sigma^2}\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\mathbf{w}$$

finding a fixed point involves setting the gradient to zero, which gives

$$\mathbf{w}^* = \left[\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

as required.

e) What problem will arise as the number of basis functions we use increases to become larger than the number of the data points we are given? How can we perform a regression in this case? [15%]

ANSWER:

The matrix $\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ is not full rank and the system is underdetermined. This means we cannot find a unique solution for $\mathbf{w}$. A solution is to apply a Bayesian approach.

4. This question deals with Bayesian approaches to machine learning problems.

a) Machine learning deals with data. What do we need to combine with the data in order to make predictions? [10%]

ANSWER:

Assumptions, and these are often incorporated in the form of a model.

b) Bayes' Rule relates four terms: the *likelihood*, the *prior*, the *posterior* and the *marginal likelihood* or *evidence*.

(i) Describe the role of each of these terms when modelling data. [20%]

ANSWER:

The prior is our belief about the parameters before we observe the data. The likelihood gives the relationship between our parameters and the observed data. The posterior gives our updated belief about the parameters after observing the data and the marginal likelihood or evidence is used for model comparison.

(ii) Write down the relationship between these four terms as given by Bayes' rule. [10%]

ANSWER:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}$$

c) In a regression problem we are given a vector of real valued targets, $\mathbf{y}$, consisting of N observations $y_1 \dots y_N$ which are associated with multidimensional inputs $\mathbf{x}_1 \dots \mathbf{x}_N$. We assume a linear relationship between y_i and $\mathbf{x}_i$ where the data is corrupted by independent Gaussian noise giving a likelihood of the form

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2\right).$$

Consider the following Gaussian prior density for the k dimensional vector of parameters, $\mathbf{w}$,

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

(i) This prior and likelihood can be combined to form the posterior. Explain why the resulting posterior will be Gaussian distributed. [10%]

ANSWER:

It will be Gaussian distributed because it depends on the product of the prior and the likelihood which are both exponentiated quadratic forms. The result of multiplying two exponential forms is that their argument is the sum of the arguments. Therefore, the sum of two quadratics. The posterior is therefore of exponentiated quadratic form, and the only density of that form is the Gaussian.

(ii) Show that the covariance of the posterior density for $\mathbf{w}$ is given by

$$\mathbf{C}_w = \left[\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbf{I}\right]^{-1},$$

where $\mathbf{X}$ is a *design matrix* of the input data. [35%]

ANSWER:

Here we need to use Bayes's rule. The posterior density is given by

$$p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) \propto p(\mathbf{y}|\mathbf{x},\mathbf{w},\sigma^2)p(\mathbf{w}),$$

the logarithm of which can be written as

$$\begin{aligned}\log p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) &= -\frac{1}{2\sigma^2}\sum_{i=1}^N (y_i - \mathbf{w}^\top\mathbf{x})^2 - \frac{1}{2\alpha}\mathbf{w}^\top\mathbf{w} + \text{const} \\ &= -\frac{1}{2\sigma^2}\sum_{i=1}^N y_i^2 - \frac{1}{2}\mathbf{w}^\top\left(\frac{1}{\sigma^2}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\top + \frac{1}{\alpha}\mathbf{I}\right)\mathbf{w} + \mathbf{w}^\top\frac{\sum_{i=1}^N(y_i\mathbf{x}_i)}{\sigma^2},\end{aligned}$$

where the constant represents terms which don't include $\mathbf{w}$. We can now use the equalities:

$$\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\top$$

and

$$\mathbf{X}^\top\mathbf{y} = \sum_{i=1}^N \mathbf{x}_i y_i$$

and substitute in to obtain

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^N y_i^2 - \frac{1}{2}\mathbf{w}^\top\left(\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbf{I}\right)\mathbf{w} + \mathbf{w}^\top\frac{\mathbf{X}^\top\mathbf{y}}{\sigma^2},$$

The next step is to recover the covariance of the Gaussian. To do that we complete the square of the quadratic form to obtain

$$\log p(\mathbf{w}|\mathbf{y},\mathbf{x},\sigma^2) = -\frac{1}{2}(\mathbf{w} - \boldsymbol{\mu}_w)\mathbf{C}_w^{-1}(\mathbf{w} - \boldsymbol{\mu}_w) + \text{const},$$

where when this term is multiplied out the linear and quadratic terms in $\mathbf{w}$ are given by,

$$-\frac{1}{2}\mathbf{w}^\top\mathbf{C}_w^{-1}\mathbf{w} + \boldsymbol{\mu}_w\mathbf{C}_w^{-1}\mathbf{w} + \text{const}$$

these need to be matched to the above equation, an that for the two quadratic forms to match we must have $\mathbf{C}_w = \left[\frac{1}{\sigma^2}\mathbf{X}^\top\mathbf{X} + \frac{1}{\alpha}\mathbf{I}\right]^{-1}$

(iii) Show that the mean of the posterior density for $\mathbf{w}$ is given by

$$\boldsymbol{\mu}_w = \mathbf{C}_w \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y}.$$

[15%]

ANSWER:

Now we match the linear term from the expanded quadratic form above. The important thing to note is that we need to introduce the covariance matrix to cancel the inverse. Once this has been noted the answer comes out as $\boldsymbol{\mu}_w = \mathbf{C}_w \frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{y}$.

END OF QUESTION PAPER