# MODEL SOLUTIONS

SETTER: Trevor Cohn and Neil Lawrence

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE Autumn Semester 2012–2013

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) *Bayes' rule* is used in many contexts in machine learning.

(i) Provide a definition of Bayes' rule. You should include mathematical formulae, and define any variables used. [10%]

ANSWER:

Bayes' rule is defined as $P(y|x) = \frac{P(y)P(x|y)}{P(x)}$. In this formulation $y$ is the latent quantity we wish to infer and $x$ is the observed evidence (in machine learning – more generally, they're both just random variables). Lose 2 marks if missing denominator. Lose 2 if $x$ and $y$ aren't defined (loosely).

(ii) Define the following components in Bayes' rule – *posterior*, *likelihood*, *prior*, *marginal likelihood* – and briefly describe their purpose (1-2 sentences for each). [20%]

ANSWER:

The components are posterior $= \frac{\text{prior} \times \text{likelihood}}{\text{marginal likelihood}}$.

- the prior embodies our initial intutions about what latent values we expect to see, before observing any data
- the likelihood is the probability of the observed data given the latent values
- the marginal likelihood is the probability of the data under any setting of the latent values, i.e., marginalising out the latent values
- the posterior is the quantity we're interested in, which combines both the prior and the likelihood. This way our initial beliefs (prior) are moderated by the evidence seen in the data (likelihood).

Equal marks each.

(iii) Provide two examples where Bayes' rule is used in machine learning, and describe why Bayes' rule is used in this setting. [15%]

ANSWER:

Here are some possibilities:

- model estimation, where we're interested in the posterior $p(\theta|x)$, where $\theta$ are the model parameters and $x$ are the training examples. This is used as it's much more straightforward to model the likelihood and prior terms (i.e., using Bayes' rule) than to model the posterior directly. Here straightforward means modelling convenience and also mathematical tractability.
- generative models of classification, e.g., mixtures of gaussians, naive Bayes or similar. In these settings we model $p(\mathcal{C}|x)$ where $\mathcal{C}$ is the class and $x$ is the data instance. In this case it's simpler to model the posterior as a

likelihood and a prior term, which leads to a simple form of the posterior and a closed-form training algorithm. Conversely, modelling the posterior directly (e.g., as done in logistic regression) requires the use of a non-linear activation function, and necessitates an iterative gradient based training algorithm.

- there are others, but we haven't covered these in any depth in class. Full marks for an alternative feasible answer.

7.5 marks for each

b) What does the term *marginalise* mean in relation to probability distributions? You should consider both the *discrete* and *continuous* settings, and provide mathematical formulae to support your answer. [15%]

ANSWER:

This means to remove the effect of a random variable from a joint distribution over multiple RVs, resulting in the marginal distribution. For discrete distributions (PMFs), this means summing over the variable,

$$p(x) = \sum_y p(x, y)$$

and for continuous distributions (PDFs), this means integrating,

$$p(x) = \int_y p(x, y) dy$$

-3 if only consider one setting, but otherwise correct.

c) The Binomial-Beta is said to be an example of a *conjugate* prior relationship.

(i) Give a definition of a *conjugate prior*, and motivate why conjugate priors are desirable. [15%]

(ii) Prove that conjugacy holds for the Binomial and Beta distributions. Show your working. [25%]

For your reference, the Binomial distribution is defined as

$$P(k, n|u) = \binom{n}{k} u^k (1-u)^{n-k}$$

where $k$ is the number of successes after $n$ trials (both positive integers), and $u$ is the binomial parameter (real number between 0 and 1). The Beta distribution is defined as

$$P(u|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} u^{\alpha-1} (1-u)^{\beta-1}$$

where $u$ is a real number between 0 and 1 and $\alpha$ and $\beta$ are the Beta parameters. The function $B(\cdot)$ is a normalising constant.

ANSWER:

A conjugate prior is a distribution which has the property that the posterior distribution (which you get by combining the prior and likelihood) has the same form as the prior. This is desirable as it means that (usually) inference is straightforward, as we only need to deal with one distribution. (8 marks for definition, 7 marks for reason)

The proof goes as follows – express the likelihood and the prior, take their product to get the posterior, and then express as a beta distribution.

$$\begin{aligned}
\mathcal{B}(k,n|p) &\propto p^k(1-p)^{n-k} && \text{likelihood}\\
\beta(p|\alpha,\beta) &\propto p^{\alpha-1}(1-p)^{\beta-1} && \text{prior}\\
P(p|k,n,\alpha,\beta) &\propto \mathcal{B}(k,n|p)\times\beta(p|\alpha,\beta) && \text{posterior}\\
&\propto p^{k+\alpha-1}(1-p)^{n-k+\beta-1}\\
&\propto \beta(p|k+\alpha,n-k+\beta)
\end{aligned}$$

Note that we use proportional to in order to discard irrelevant scaling factors, including the denominator term (marginal likelihood) in Bayes' rule.

Mark breakdown – half marks for formulating the posterior correctly; full marks for the final correct answer (which can be in text). Lose 5 marks for each simple mistake, e.g., with normalisation constants.

2. This question is based on the following data. We are trying to predict whether or not it will rain, and have identified two features which might be important—whether the sky is clear or cloudy, and whether or not we hear birds singing. Over three days, we observed the following:

   1. cloudy sky; birds singing; rain
   2. clear sky; birds singing; no rain
   3. clear sky; birds quiet; rain

   a) We have decided to model the data using the perceptron algorithm.

      (i) Illustrate this data using a graph, denoting each day as a point. Draw a separating hyperplane on the graph and label the regions for the two classes. [10%]

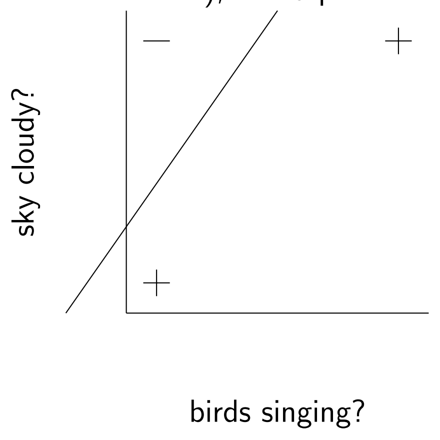
      (ii) State the weight update rule used in the perceptron algorithm. [5%]

      (iii) Now apply the perceptron algorithm for training the model parameters. First, represent your training data as a matrix, $\mathbf{X}$, for the data points and vector $\mathbf{t}$ for their target values (+1 = rain and -1 = no rain). Now perform just one pass of the perceptron algorithm over the training set, showing your working and the final parameter values. Don't forget to include a bias term. [15%]

      (iv) The following day there is a cloudy sky and the birds are quiet. What is your model's prediction (i.e., rain or not)? Include your working. [5%]

ANSWER:

1. Something like the figure below – 5 points for drawing the points (n.b. flipping either axis is ok), and 5 points for drawing a line that splits the +s from the -s.



2. The update rule is $\mathbf{w} \leftarrow \mathbf{w} + t_i \mathbf{x}_i$.

3. First, we represent the data and the target values. We incorporate the bias term using the first columns of 1s.

| $x_0$ | $x_1$ | $x_2$ | $t$ |
|---|---|---|---|
| 1 | 1 | 1 | 1 |
| 1 | 0 | 1 | -1 |
| 1 | 0 | 0 | 1 |

Now starting with parameters $\mathbf{w} = [000]$.

(a) instance 1, $\mathbf{w} = [0\ 0\ 0]$, $\mathbf{w}^T\mathbf{x} = 0$, $y(x) = 1$, no error hence no update

(b) instance 2, $\mathbf{w} = [0\ 0\ 0]$, $\mathbf{w}^T\mathbf{x} = 0$, $y(x) = 1$, an error hence update $\mathbf{w}+ = [-1\ 0\ -1]$

(c) instance 3, $\mathbf{w} = [-1\ 0\ -1]$, $\mathbf{w}^T\mathbf{x} = -1$, $y(x) = -1$, an error hence update $\mathbf{w}+ = [1\ 0\ 0]$

The final value is $\mathbf{w} = [0\ 0\ -1]$.

This can be presented in a number of ways. 5 marks = final answer correct. 5 marks = show input data in a sensible way. 5 marks = show iterative updates with changing $w$. No marks docked if using a sign algorithm that assigns 0 to -1, although this will lead to a different end result.

4. The data point is $x = [1\ 1\ 0]$, and therefore $y(x) = \text{sign}(0) = 1$. So it will predict rain. No marks docked if using a sign algorithm that assigns 0 to -1, nor if the answer here is consistent with the answer to part 2 above.

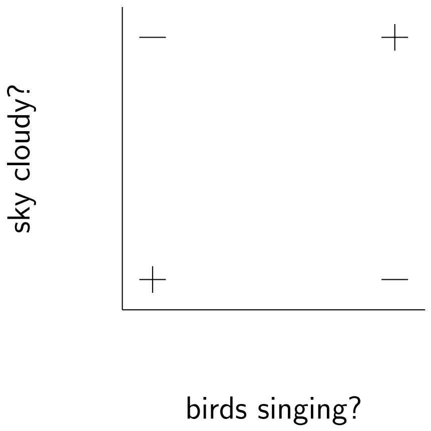
b) It turns out on the fourth day that it doesn't rain. We now elect to include this new example (cloudy sky; birds quiet; no rain) into our training set and re-train our model.

(i) The above model will no longer be appropriate for this dataset. Justify why this is the case. [10%]

(ii) Would you be able to solve the problem using radial basis functions? If so, how many RBFs are needed and where could they be placed? Please justify your answer, either way. [20%]

ANSWER:

1. The data set is no longer linearly separable (5 marks). If we draw the new dataset, we have



which is the classic XOR problem. There exists no straight line that can discriminate between the classes. We will always get one point incorrect (5 marks).

2. Yes (5 marks). RBFs allow for non-linear decision boundaries by measuring the euclidean distance to a given point. This gives rise to curved contours and generally curved decision boundaries based on which centre is closer to each point. (7 marks)

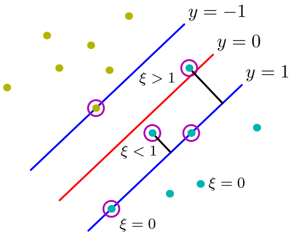
You can solve it using two RBFs, which need to be placed at or near two points of the same class (either both +s or both -s). (8 marks)

c) *Support Vector Machines (SVMs)* refine the perceptron by including the notion of margin of separation.

(i) Illustrate the concept of the separating margin using a diagram assuming binary classification with two dimensional input data. Now highlight all the support vectors, and annotate the margin. State which points violate the margin constraints (ensure you include a few). [15%]

ANSWER:

Something like this:



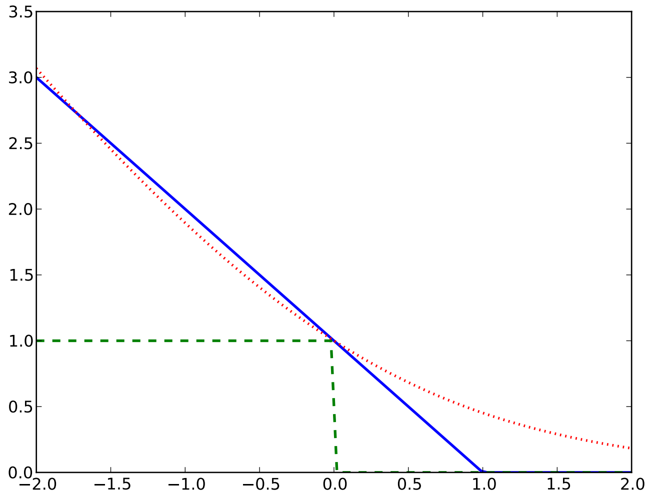
(Bishop, PRML, Chapter 7)

Note that the circled points are the SVs, $\xi > 0$ shows various margin violations, and the parallel lines for $y = \pm 1$ show the decision boundary.
Marks assigned: correctly identifies SVs (5 marks); margin annotated between SVs and boundary (5 marks); shows points on the wrong side of the decision boundary annotated as violating the constraints (5 marks).

(ii) What is the loss function being minimised by the soft-margin SVMs? Include the mathematical formula. How does this differ from the zero-one loss and logistic loss? Provide a diagram to illustrate your answer. [20%]

ANSWER:

It's the hinge loss, which is defined as $-\min(0, 1 - ty(x))$ where $t$ is the target class, and $y(x) = w^T x$ is the discriminant value. See figure, it's the blue solid line. It differs from 0-1 loss (green dashes) at it penalises correct but unconfident classifications, and has a rising loss based on the magnitude of mistake. Note that 0-1 loss is discontinuous, where hinge and log loss are not. Compared to logistic loss (red dotted), $\log(1 + \exp(ty(x)))$, it is quite similar, but is less harsh on bad misclassifications and is rewards all confident classifications $(y > 1)$ equally while logistic loss gradually reduces towards zero at the asymptote. This means logistic loss will take (slightly) more account of outliers on the correct side of the decision boundary.



Note that in the figure logistic loss has been scaled by a factor of $\log(2)$. This isn't important, it just makes the graph easier to interpret.
10 marks for stating hinge loss including formula. 5 marks for comparison, stating at least one valid difference each between HL and 01 and HL and LL.
5 marks for diagram.

3. This question concerns regression and the Bayesian approach to regression with basis functions.

   a) For each pair of terms below define and contrast the two terms. Use at least one example to illustrate your answer.

      (i) overdetermined and underdetermined systems [15%] ANSWER:

      Overdetermined systems have more data than parameters. In the case of generalised linear models, if you try and solve the system exactly, each data point leads to a set of simultaneous equations and each parameter an unknown. In overdetermined systems you have more equations than unknowns. A one dimensional linear regression problem, without noise, and having more than two observations is over determined. Underdetermined systems have more parameters than data. When trying to solve the system exactly you have fewer equations than unknowns. A one dimensional linear regression problem with only one data point is an underdetermined system.

      (ii) epistemic and aleatoric uncertainty [15%] ANSWER:

      Epistemic uncertainty is our uncertainty about events the outcome of which could be in principle known. For example watching a recording of a football match which has already finished involves epistemic uncertainty about the result. Aleatoric uncertainty is uncertainty about events which is not knowable. For example, watching a football match live leads to aleatoric uncertainty about the result.

   b) A typical linear model could have the form

$$t_i = mx_i + c + \epsilon_i$$

   where $t_i$ is the regression target observation, $x_i$ is the input location and $\epsilon_i$ is the noise. All are associated with the $i$th observation.

      (i) Write the form of the basis set for the $i$th data point, $\boldsymbol{\phi}_i$, such that this model can be written:

$$t_i = \mathbf{w}^\top \boldsymbol{\phi}_i + \epsilon_i.$$

      [5%] ANSWER:

      The basis here is simply $\boldsymbol{\phi}_i = [1 \quad x_i]^\top$ where $w_1 = c$ and $w_2 = m$.

      (ii) The linear model is a 1st order polynomial. What would the basis, $\boldsymbol{\phi}_i$, be for a 4th order polynomial? [5%] ANSWER:

      The basis here is simply $\boldsymbol{\phi}_i = [1 \quad x_i \quad x_i^2 \quad x_i^3 \quad x_i^4]^\top$.

c) In a regression problem we are given a vector of real valued targets, $\mathbf{t}$, consisting of $N$ observations $t_1 \ldots t_N$ which are associated with a unidimensional input $x_1 \ldots x_N$. We are to perform the regression by minimizing the following error function with respect to $\mathbf{w}$

$$E(\mathbf{w}) = \sum_{i=1}^{N} (t_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2.$$

Write down the *likelihood* that corresponds to this error function, introducing any additional parameters as necessary. Describe how the error function is related to the likelihood. [20%]

ANSWER:

The error function corresponds to a *Gaussian* likelihood. The additional parameter that is required is the variance of the Gaussian which corresponds to the noise level, denoting this by $\sigma^2$ gives us the following likelihood

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

The error function, up to a scaling and a value that is constant in $\mathbf{w}$, is the negative logarithm of the likelihood.

d) Consider the following Gaussian prior density for $\mathbf{w}$,

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

where $k$ is the length of the vector $\mathbf{w}$ and $\alpha$ is the variance of the prior.

(i) Multiply the prior by the likelihood from (c). Show that the result is of the form of an exponentiated quadratic, and describe why that means the posterior density for $\mathbf{w}$ is Gaussian. [20%]

ANSWER:

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

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) \propto p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \sigma^2)p(\mathbf{w}),$$

the logarithm of which can be written as

$$\log p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(t_i - \mathbf{w}^\top \boldsymbol{\phi}_i)^2 - \frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w} + \text{const}$$

where the constant represents terms which don't include $\mathbf{w}$. Even at this stage it is clear that this expression contains terms that are only quadratic or linear in $\mathbf{w}$. When we re-exponentiate to get the posterior density, the only density that is the exponential of a quadratic is the multivariate Gaussian density. So this must be the multivariate Gaussian.

(ii) Show that the mean, $\boldsymbol{\mu}_w$, and covariance, $\mathbf{C}_w$, of the posterior density are given by

$$\mathbf{C}_w = \left[\alpha^{-1}\mathbf{I} + \sigma^{-2}\sum_i^N \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right]^{-1}$$

$$\boldsymbol{\mu}_w = \mathbf{C}_w\sigma^{-2}\sum_{i=1}^N t_i\boldsymbol{\phi}_i$$

[20%]

ANSWER:

When the brackets from the previous part are multiplied out and we collect quadratic and linear terms in $\mathbf{w}$ we have

$$\log p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2) = -\frac{1}{2}\mathbf{w}^\top\left[\alpha^{-1}\mathbf{I} + \sigma^{-2}\sum_i^N \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right]\mathbf{w}+\sigma^{-2}\sum_{i=1}^N t_i\boldsymbol{\phi}_i^\top\mathbf{w}+\text{const}$$

The exponent of a multivariate Gaussian density takes the form

$$-\frac{1}{2}(\mathbf{w}-\boldsymbol{\mu}_w)^\top\mathbf{C}_w^{-1}(\mathbf{w}-\boldsymbol{\mu}_w)$$

which multiplies out to

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

This needs to be matched to the log posterior we multiplied out above. Matching the quadratic forms between the two expressions implies that

$$\mathbf{C}_w = \left[\alpha^{-1}\mathbf{I} + \sigma^{-2}\sum_i^N \boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right]^{-1}$$

and

$$\mathbf{C}_w^{-1}\boldsymbol{\mu}_w = \sigma^{-2}\sum_{i=1}^N t_i\boldsymbol{\phi}_i$$

so

$$\boldsymbol{\mu}_w = \mathbf{C}_w\sigma^{-2}\sum_{i=1}^N t_i\boldsymbol{\phi}_i$$

4. This question concerns the concepts behind data modelling such as generalization and model selection. In your answers, when it is appropriate, you may want to make use of the regression example we saw in the lectures and lab class involving the gold medal winning 100m times from the Olympic games between 1896 and 2008.

a) Give short definitions for the following terms associated with a model fitting and generalisation capability.

(i) overfitting [5%]

ANSWER:

Overfitting is when a model with too greater complexity is applied to a data set. The result is that the training error is low, but when the model is applied to previously unseen test data, the error is high. [For example in the olympics data if a too higher order polynomial is fitted the regression line goes through all the data points but doesn't generalise well between them].

(ii) extrapolation [5%]

ANSWER:

Extrapolation is error that arises when moving beyond the region of the data. [For example, in the Olympics 100 m data set we have data up until the year 2008, predicting forward in time from this data (2012, 2016 etc) is extrapolation — example not required for full marks]

(iii) interpolation [5%]

ANSWER:

Predicting between training data points. [For example in the Olympics data predicting the missing Olympics during the Second World War (1940 and 1944) would be interpolation because they are between 1936 and 1948].

b) In this part we will cover approaches to model selection.

(i) What is a validation set? [10%]

ANSWER:

A validation set is a portion of the training data which is not used for training the model, but is used to estimate the test error for the purposes of model selection. For example in the Olympics data we could hold out from 1980 to the present day from the training data and test on these examples to evaluate the quality of the model.

(ii) What is the difference between hold out validation and cross validation? [20%]

ANSWER:

Hold out validation is when a portion of the training data is selected as a validation set and the validation set is simply 'held out' at the training stage and used to estimate the test error. In cross validation the validation set is alternated across the training data, a portion of the data is used for training, and a portion for validation. The portions are then swapped (perhaps several times, or even $N$ times for leave one out cross validation) and the errors are averaged. For example, in the olympics data hold out validation involves training only on e.g. the first portion of the data set and testing on a later portion

(iii) What are the relative advantages and disadvantages of leave-one-out cross validation and five fold cross validation? [25%]

ANSWER:

Leave one out cross validation gives the best estimate of the generalization error as it uses almost all the training data to compute each model [10 marks], and computes the test error as an average over all data [5 marks]. Unfortunately it increases computational complexity by $N$ where $N$ is the number of training data because it needs to be done $N$ times. Five fold cross validation is quicker to perform as it only needs to be done five times, but it doesn't use all the training data to estimate the model so it can give model selection for a model that is less complex than could be used.

c) What is the Bayesian approach to model selection and why is it less susceptible to overfitting? [30%]

ANSWER:

The Bayesian approach to model fitting involves integrating over the parameters rather than optimising them out. The model selection is then done using the marginal likelihood (the likelihood with the parameters marginalised) rather than the likelihood. This typically has less parameters and therefore is less susceptible to overfitting. The reason this works is because Bayesian approaches perform model averaging, they average over many plausible solutions to the problem and give error bars, rather than optimising one solution that can overfit.

**END OF QUESTION PAPER**