MODEL SOLUTIONS

SETTER: Neil Lawrence

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE **AUTUMN SEMESTER 2014–2015**

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE **1 hour**

Please note that the rubric of this paper is made different from many other papers.

Section A consists of THREE questions (Questions 1-3) and 40 marks in total.
Section B consists of TWO questions (Questions 4, 5) worth 60 marks each.

Answer ALL Questions 1-3 in Section A, and ONE Question from Section B (either Question 4 or Question 5).

Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

SECTION A

Answer all three questions in this section.

1. **Multiple Choice Questions: choose EXACTLY ONE answer to each part**

NoseyParker.com Ltd is an online retailer that records the music, film, games and books its users have bought. It uses this information to build a profile for each user, characterising their personality and interests.

Wizard Inc sells a massive multiplayer online role-playing game, WasteOfTimecraft (WoT). It wishes to sell these downloads through NoseyParker.com. NoseyParker.com wants to design a machine learning algorithm to target adverts at users who are most likely to buy its services.

NoseyParker.com has data about users who have bought downloads in the past. Each set of items that a user has bought in the past is in the form of a binary vector, $\mathbf{x}$. The elements of this vector are associated with items (either music, film or books). The entry in the vector contains a 1 if a given user has bought a given item.

a) For its historic data, NoseyParker.com then has a binary value, y which contains a 1 if a user downloaded the WasteOfTimecraft game, and a zero if the user didn't download the game. NoseyParker.com decides to try to use *logistic regression* to learn a set of classification weights, $\mathbf{w}$, for the data. Which of the following describes $\pi(\mathbf{x})$, the conditional probability of the positive class given $\mathbf{w}$ and $\mathbf{x}$? [3%]

(i) $\log(1+\exp(-\mathbf{w}^\top\mathbf{x}))$

(ii) $\dfrac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}$

(iii) $\exp(-(y-\mathbf{w}^\top\mathbf{x})^2)$

(iv) $\mathbf{w}^\top\mathbf{w}\exp(-\mathbf{w}^\top\mathbf{x})$

ANSWER:

(ii) $\dfrac{1}{1+\exp(-\mathbf{w}^\top\mathbf{x})}$

b) NoseyParker.com decides to perform stochastic gradient descent to optimize the logistic regressor. If $\pi(\mathbf{x})$ is the conditional probability for class one being correct, and η is the *learning rate*, and the observed class is positive, which of the following gives the correct update for the weights $\mathbf{w}$? [3%]

(i) $\mathbf{w} \leftarrow \mathbf{w} - \eta(1 - \pi(\mathbf{x}))\mathbf{x}$

(ii) $\mathbf{w} \leftarrow \mathbf{w} + \eta\pi(\mathbf{x})\mathbf{x}$

(iii) $\mathbf{w} \leftarrow \mathbf{w} + \eta(1 - \pi(\mathbf{x}))\mathbf{x}$

(iv) $\mathbf{w} \leftarrow \mathbf{w} - \eta\pi(\mathbf{x})\mathbf{x}$

ANSWER:

(i) $\mathbf{w} \leftarrow \mathbf{w} - \eta(1 - \pi(\mathbf{x}))\mathbf{x}$

c) The logistic regression system is fully trained. User Simon3477 is online and has a historical purchasing history which is represented by the following vector

$$\mathbf{x} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}.$$

Which of the following training results would cause NoseyParker.com to predict with probability greater than 0.5 that Simon3477 will download the game: [3%]

(i) $\mathbf{w} = \begin{bmatrix} -1 & -1 & 1 & 1 & 1 & -4 \end{bmatrix}$

(ii) $\mathbf{w} = \begin{bmatrix} 2 & -1 & -2 & 1 & 1 & -1 \end{bmatrix}$

(iii) $\mathbf{w} = \begin{bmatrix} 2 & -1 & 4 & 1 & 1 & 1 \end{bmatrix}$

(iv) $\mathbf{w} = \begin{bmatrix} 4 & 3 & -3 & 1 & 1 & -2 \end{bmatrix}$

ANSWER:

(iii) $\mathbf{w} = \begin{bmatrix} 2 & -1 & 4 & 1 & 1 & 1 \end{bmatrix}$

d) The inner product between the feature vector and the weight matrix aims to approximate: [3%]

(i) The logarithm of the odds ratio between the positive and negative classes.

(ii) The probability of the positive class being correct.

(iii) The mean of the Gaussian random variable which is most likely to generate the given class conditional density.

(iv) The logarithm of the probability of the positive class.

ANSWER:

(i) The logarithm of the odds ratio between the positive and negative classes.

e) NoseyParker.com's data scientist suggests *naïve Bayes* as an alternative approach to modelling. Which one of the following is *not* true of the naïve Bayes classifier? [3%]

(i) The features are conditionally independent given the class.

(ii) The features are distributed according to a multivariate Gaussian.

(iii) The data is conditionally independent given the model parameters.

(iv) The class labels are discrete values.

ANSWER:

(ii) The features are distributed according to a multivariate Gaussian.

2. In the answer booklet, state whether each of the following mathematical equalities is True or False.

a) $p(x|y) = p(y|x)p(y)$ [2%]

ANSWER:

False

b) $p(y) = \dfrac{p(y|x)p(x)}{p(x|y)}$ [2%]

ANSWER:

True

c) $p(y|x) = \dfrac{p(x,y)}{p(x|y)p(x)}$ [2%]

ANSWER:

False

d) $\dfrac{p(y)}{p(x)} = \dfrac{p(y|x)}{p(x|y)}$ [2%]

ANSWER:

True

e) $\dfrac{p(y,x)}{p(x)} = \dfrac{p(y|x)}{p(y)}$ [2%]

ANSWER:

False

3. Consider the following five equalities:

(1) $y = \mathbf{x}^\top \text{diag}(\mathbf{x})\mathbf{w}$

(2) $y = \mathbf{x}^\top \mathbf{w}$

(3) $\mathbf{y} = \text{diag}(\mathbf{x})\mathbf{w}$

(4) $\mathbf{Y} = \mathbf{x}\mathbf{w}^\top$

(5) $y = \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w}$

Find the correct equality that matches each of the following pieces of python code. Assume that the mathematical operator diag$(\mathbf{z})$ forms a diagonal matrix with diagonal elements given by elements of $\mathbf{z}$ and that in python the `numpy` library as been imported as `np` and we are given `x` and `w` as one dimensional numpy arrays.

a) `y = np.sum(x*w)` [3%]

ANSWER:

(2) $y = \mathbf{x}^\top \mathbf{w}$

b) `y = x*w` [3%]

ANSWER:

(3) $\mathbf{y} = \text{diag}(\mathbf{x})\mathbf{w}$

c) `y = np.outer(x,w)` [3%]

ANSWER:

(4) $\mathbf{Y} = \mathbf{x}\mathbf{w}^\top$

d) `y = np.sum(w*x**2)` [3%]

ANSWER:

(1) $y = \mathbf{x}^\top \text{diag}(\mathbf{x})\mathbf{w}$

e) `y = np.sum(x*w)**2` [3%]

ANSWER:

(5) $y = \mathbf{w}^\top \mathbf{x}\mathbf{x}^\top \mathbf{w}$

SECTION B

Answer **EITHER** Question 4 **OR** Question 5 in this section.

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

a) What is the difference between *epistemic* and *aleatoric* uncertainty? [10%]

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) In a regression problem we are given a vector of real-valued targets, $\mathbf{y}$, consisting of n observations $y_1 \ldots y_n$ which are associated with multidimensional inputs $\mathbf{x}_1 \ldots \mathbf{x}_n$. We assume a linear relationship between y_i and $\mathbf{x}_i$ where the data are corrupted by independent Gaussian noise giving a likelihood function 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) 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%]

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

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

[15%]

ANSWER:

(i)

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

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

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

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

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

(ii)

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

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

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

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

Give a disadvantage of the polynomial basis. Suggest a potential solution 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.

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

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

Assuming that each data point is independent and identically distributed, derive a suitable *objective function* that should be minimized to recover $\mathbf{w}$ and σ^2. Explain your reasoning at each step. [15%]

ANSWER:

The independence assumption means that

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

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

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

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:

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

c) Find the stationary point of this objective function where it is minimized with respect to the vector $\mathbf{w}$. [25%]

ANSWER:

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

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

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

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

this can be rewritten in matrix form as

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

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

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

as required.

END OF QUESTION PAPER