Version MJFG/4

EGT3
ENGINEERING TRIPOS PART IIB

Tuesday 23 April 2019   2 to 3.40

Module 4F10

DEEP LEARNING AND STRUCTURED DATA

*Answer not more than **three** questions.*

All questions carry the same number of marks.

*The **approximate** percentage of marks allocated to each part of a question is indicated in the right margin.*

*Write your candidate number **not** your name on the cover sheet.*

STATIONERY REQUIREMENTS
Single-sided script paper

SPECIAL REQUIREMENTS TO BE SUPPLIED FOR THIS EXAM
CUED approved calculator allowed
Engineering Data Book

10 minutes reading time is allowed for this paper at the start of the exam.

You may not start to read the questions printed on the subsequent pages of this question paper until instructed to do so.

1 A classifier is to be built for a two-class problem. There are n, d-dimensional, training vectors, $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with class labels, $y_1, \ldots, y_n$. If observation $\mathbf{x}_i$ belongs to class ω_1 then $y_i = 1$, and if it belongs to class ω_2 then $y_i = 0$. The classifier has the form

$$P(\omega_1|\mathbf{x}, \mathbf{b}) = \frac{1}{1 + \exp(-\mathbf{b}^\mathsf{T}\mathbf{x})}$$

The parameters of the model are to be trained to maximise the log-probability that can be written as

$$\mathscr{L}(\mathbf{b}) = \sum_{i=1}^{n} \left(y_i \log(P(\omega_1|\mathbf{x}_i, \mathbf{b})) + (1 - y_i)\log(1 - P(\omega_1|\mathbf{x}_i, \mathbf{b})\right)$$

(a) What form of decision boundary can be obtained with this type of classifier? [10%]

(b) The parameters of the classifier, $\mathbf{b}$, are to be trained using gradient ascent based approaches.

(i) Derive an expression for the derivative of $\mathscr{L}(\mathbf{b})$ with respect to $\mathbf{b}$. How can this expression be used to train the model parameters? [25%]

(ii) Show that the Hessian, $\mathbf{H}$, that may be used to train this classifier can be expressed as

$$\mathbf{H} = -\mathbf{S}^\mathsf{T}\mathbf{R}\mathbf{S}$$

Find expressions for the matrices $\mathbf{S}$ and $\mathbf{R}$. [25%]

(iii) Give an appropriate formula for finding the model parameters that involves the Hessian. What are the advantages and disadvantages of using this form of expression compared to the one in (b)(i)? [15%]

(iv) Comment on the form of the Hessian matrix in (b)(ii) and what it implies for this optimisation problem. [10%]

(c) A regularisation term is now added to the training criterion. The new training criterion, $\tilde{\mathscr{L}}(\mathbf{b})$, has the form

$$\tilde{\mathscr{L}}(\mathbf{b}) = \mathscr{L}(\mathbf{b}) + \lambda\mathbf{b}^\mathsf{T}\mathbf{b}$$

How does this regularisation term change the expressions for the derivative and Hessian in (b)? Why might this form of regularisation be useful? [15%]

2 A neural network based sentiment classification system is to be trained on a corpus of film reviews. There are N film reviews and each review is labelled with one of K sentiment labels. Each of the words in a review is mapped to a d-dimensional vector representation, so the ith, L length, review will be mapped to the L length sequence $\mathbf{X}^{(i)} = \{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_L^{(i)}\}$. Thus the available training data, $\mathscr{D}$, are

$$\mathscr{D} = \{\{\mathbf{X}^{(1)}, y^{(1)}\}, \dots, \{\mathbf{X}^{(N)}, y^{(N)}\}\}$$

where $y^{(i)} \in \{\omega_1, \dots, \omega_K\}$ and ω_j indicates the sentiment class. It is decided to train the network in the following configuration. The classification of an unknown mapped review, $\mathbf{X}^\star$, is based on

$$\mathbf{h} = \mathbf{f}(\mathbf{X}^\star); \quad \mathbf{t}^\star = \mathbf{g}(\mathbf{h})$$

where $\mathbf{h}$ is an l-dimensional, fixed-length, vector, and the estimate of the sentiment, $\hat{\omega}^\star$, is obtained from the K-dimensional vector $\mathbf{t}^\star$.

(a) Describe an appropriate form of neural network for the function $\mathbf{g}()$. For this network, what is a suitable training criterion to estimate the model parameters based on $\mathscr{D}$? You should clearly describe all terms in the training criterion and why you think the form is appropriate. [25%]

(b) Initially it is proposed to use a recurrent neural network for the function $\mathbf{f}()$.

(i) Describe how a recurrent neural network can be used for the function $\mathbf{f}()$. You should clearly describe the equations required, using the L-length sequence $\mathbf{X}^{(i)}$ as an example, and the resulting form for the fixed-length vector $\mathbf{h}$. [20%]

(ii) A bi-directional recurrent neural network is proposed as an alternative approach. Briefly describe this form of network and how it can be used for $\mathbf{f}()$. [15%]

(c) The network is now modified so that an attention mechanism is used for the function $\mathbf{f}()$. The key that is used to determine the relevance of a mapped word to sentiment classification, for word j of the review $\mathbf{X}^{(i)}$, is the mapped word vector $\mathbf{x}_j^{(i)}$. Describe, including appropriate equations, how this attention mechanism can be used for the function $\mathbf{f}()$ again using review $\mathbf{X}^{(i)}$ as an example. [20%]

(d) Briefly state the advantages and disadvantages of the possible forms of network for $\mathbf{f}()$ described in (b) and (c). [20%]

3 The exponential family of probability distributions for d-dimensional data may be described by the following equation

$$p(\mathbf{x}|\alpha) = \frac{1}{Z}\exp\left(\alpha^{\mathsf{T}}\mathbf{f}(\mathbf{x})\right)$$

where α is the vector of parameters associated with the distribution and $\mathbf{f}(\mathbf{x})$ is a function of the data point $\mathbf{x}$ that returns a vector of the same dimension as α.

(a) What expression must be satisfied by Z for $p(\mathbf{x}|\alpha)$ to be a valid probability density function? [10%]

(b) Show that the distribution with parameters $\mu = [\ \mu_1 \ \ \ldots \ \ \mu_d\ ]^{\mathsf{T}}$, $\mu_i > 0$,

$$p(\mathbf{x}|\mu) = \frac{1}{Z}\prod_{i=1}^{d}\mu_i^{x_i} \quad \text{where } \mathbf{x} = [\ x_1 \ \ \ldots \ \ x_d\ ]^{\mathsf{T}} \text{ and } x_i \geq 0$$

is a member of the exponential family. Find expressions for α, $\mathbf{f}(\mathbf{x})$ and Z. It may be useful to consider the univariate exponential distribution: $p(x) = \lambda\exp(-\lambda x)$. [30%]

(c) Rather than a single distribution, a mixture of distributions from this family is to be used. This has the form

$$p(\mathbf{x}|\alpha) = \sum_{m=1}^{M} c_m \frac{1}{Z_m}\exp\left(\alpha_m^{\mathsf{T}}\mathbf{f}(\mathbf{x})\right)$$

The parameters of this distribution, $\alpha_1, \ldots, \alpha_M$, are to be trained on n independent samples of data, $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The priors, c_1 to c_M, are known and not re-estimated. Maximum Likelihood (ML) training is used to estimate the model parameters.

(i) Write down an expression for the log-likelihood of the training data using this mixture distribution. [15%]

(ii) The parameters of the model, $\alpha_1, \ldots, \alpha_M$, are to be estimated using Expectation Maximisation (EM). The following form of auxiliary function is to be used

$$Q(\alpha, \hat{\alpha}) = \sum_{m=1}^{M}\sum_{i=1}^{n} P(m|\mathbf{x}_i, \alpha)\log(p(\mathbf{x}_i|m, \hat{\alpha}_m))$$

What statistics must be extracted from the training data to allow the model parameters to be estimated? [25%]

(iii) Discuss, including appropriate equations, how the auxiliary function can be maximised. [20%]

4 (a) Explain how kernels can be used to turn linear machine learning algorithms into non-linear ones. [15%]

(b) A hard-margin SVM classifier of the form $y(\mathbf{x}) = \mathbf{w}^{\mathsf{T}}\mathbf{x}$ is to be trained on a linearly separable dataset $\{(\mathbf{x}_n, t_n)\}_{n=1}^{N}$, where $\mathbf{x}_n \in \mathbb{R}^d$ and $t_n \in \{-1, 1\}$. The classifier parameters $\mathbf{w}$ are obtained by minimising the objective criterion $\frac{1}{2}||\mathbf{w}||^2$ subject to constraints.

(i) What are the constraints of the optimisation problem for $\mathbf{w}$? [10%]

(ii) By using Lagrange multipliers, $a_1 \geq 0, \ldots, a_N \geq 0$, write the Lagrangian function for the resulting constrained optimisation problem. [10%]

(iii) Derive an expression for $\mathbf{w}$ in terms of Lagrange multipliers and the training data, that minimises the Lagrangian obtained in (b)(ii). [15%]

(c) The perceptron algorithm, shown below, is an alternative way of training a classifier given a linearly separable dataset:

Algorithm 1 Perceptron

```
**Input**: Dataset $\{(\mathbf{x}_n, t_n)\}_{n=1}^{N}$ **Output**: $\mathbf{w}$
$\mathbf{w} \leftarrow \mathbf{0}$
**repeat**
    Errors $\leftarrow$ False
    **for** $n = 1, \ldots, N$ **do**
        **if** $t_n \mathbf{w}^{\mathsf{T}} \mathbf{x}_n < 0$ **then**
            $\mathbf{w} \leftarrow \mathbf{w} + t_n \mathbf{x}_n$
            Errors $\leftarrow$ True
**until** Not Errors
```

(i) What are the constraints that $\mathbf{w}$ must satisfy for the perceptron algorithm to terminate? Contrast these to the ones obtained in (b)(i) and discuss why SVMs are typically preferred. [20%]

(ii) Show that the value of $\mathbf{w}$ returned by Algorithm 1 can be written in the same form as the expression obtained in (b)(iii). You should clearly describe all terms in the expression. [15%]

(iii) By using your answer to (c)(ii), or otherwise, write the pseudo-code for a kernelised perceptron algorithm. [15%]

END OF PAPER

THIS PAGE IS BLANK