University of Glasgow

Thursday 20 May 2021
Available from 09:30 BST
Expected Duration: 2 hours
Time Allowed: 4 hours
Timed exam within 24 hours

DEGREES OF MSc, MSci, MEng, BEng, BSc,MA and MA (Social Sciences)

Machine Learning H
COMPSCI 4061

(Answer all of the 3 questions)

This examination paper is an open book, online assessment and is worth a total of 60 marks

1. Linear regression question

(a) Considering fitting a linear regression model $t_n = \mathbf{w}^T\mathbf{x}_n$, a general formulation of loss function is defined as follows:

$$L = |t_n - \mathbf{w}^T\mathbf{x}_n|^q$$

Assuming q can only be 0.5, 2.5, or 4.5, which value of q is most likely to have been used to produce the result in figure 1, and why?

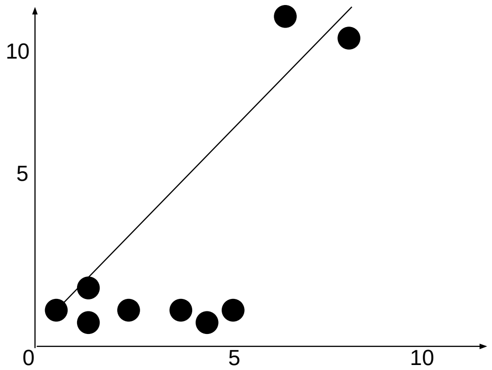

Figure 1: Data for linear regression.

[5 marks]

(b) Let's consider accounting for the noise in linear regression, the model changes to $t_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n$, where ϵ_n is a random variable following a Gaussian distribution. What is the consequence if the mean of the Gaussian distribution is not zero? How would you detect if this has happened in fitted models?

[4 marks]

(c) The following matrix contains estimated parameters values from three types of 10th order polynomial regression models. The model type is indicated by the columns. The parameter of each polynomial order is placed in a corresponding row. Give your best estimate of what each model is and explain why.

Polynomial order 0 (top) to 10 (bottom)

```
        Model 1          Model 2          Model 3
[[ 1.19648635e+01  1.11285803e+01  1.08381535e+01]
 [-1.29443055e+01 -3.29359603e-01 -0.00000000e+00]
 [ 5.79522897e+01 -1.94725736e-01 -0.00000000e+00]
 [-1.09582035e+02 -9.64898104e-02 -1.16126582e-01]
 [ 6.23248849e+01 -1.87327081e-02 -1.59001968e-02]
 [ 7.48519704e+01  3.32402164e-02 -0.00000000e+00]
 [-1.46955431e+02  4.50182751e-02  1.38119952e-03]
 [ 1.04735797e+02  9.53751777e-03  3.22128802e-03]
 [-3.88035781e+01 -3.60588365e-02  1.61616847e-04]
 [ 7.43343695e+00  1.40369595e-02 -8.65262203e-05]
 [-5.82870289e-01 -1.62830483e-03 -7.74413350e-05]]
```

[6 marks]

(d) In addition to polynomial regression, linear regression can be generalized using other basis functions. One of most widely used examples is the Fourier analysis, let's consider the following linear regression model:

$$t_n = \sum_{j}^{m} A_j \cos(jx_n + \theta_j)$$

What is the basis function of choice here? How would you deal with the unknown parameters A_j and θ_j? (Hint: you might find the following trigonometry identity useful, $\cos(a+b) = \cos(a)\cos(b) + \sin(a)\sin(b)$).

[5 marks]

2. Probabilistic modelling and Bayesian inference question

(a) Consider the linear regression model with noise $t_n = \mathbf{w}^T\mathbf{x}_n + \epsilon_n$, the maximum likelihood estimator for σ^2, the variance of ϵ_n, is:

$$\widehat{\sigma^2} = \frac{1}{N}(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})^T(\mathbf{t} - \mathbf{X}\widehat{\mathbf{w}})$$

where $\widehat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$. This estimator is biased, explain what the statistical implication of this is.

[4 marks]

(b) The marginal likelihood of the linear regression model within the Bayesian framework is following:

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

What is $p(\mathbf{w})$ in the above formula? Can you use the marginal likelihood to select polynomial order? Why? What is the impact on selecting optimal polynomial order if you replace $p(\mathbf{w})$ with $p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \sigma^2)$ in the right-hand side of the equation?

[6 marks]

(c) The likelihood of logistic regression is defined by:

$$p(t_n|\mathbf{w}, \mathbf{x}_n) = g(\mathbf{w}^T\mathbf{x}_n)^{t_n}(1 - g(\mathbf{w}^T\mathbf{x}_n))^{1-t_n}$$

where $g(a) = \frac{1}{1+\exp(-a)}$. Use an example of a few data points to explain how the likelihood function tells how well (and bad) the parameter $\mathbf{w}$ fits the data.

[4 marks]

(d) The Metropolis-Hastings algorithm is a general method to draw samples from $p(\mathbf{w})$. A key part of the algorithm is a proposal distribution $q(\mathbf{w}^*|\mathbf{w})$. To accept or reject a proposed sample ($\mathbf{w}^*$), we compute the following ratio:

$$r = \frac{p(\mathbf{w}^*)}{p(\mathbf{w})}\frac{q(\mathbf{w}|\mathbf{w}^*)}{q(\mathbf{w}^*|\mathbf{w})}$$

Give an example of a good choice for $q(\mathbf{w}|\mathbf{w}^*)$. Explain why it is a good choice and how are you going to draw sample $\mathbf{w}^*$ from it.

[6 marks]

3. *AUC and Clustering*

(a) Let's consider a binary classifier trained on a falsely labeled dataset. The issue is all positive (*1*) and negative (*0*) labels are swapped during training. The classifier outputs in the table below:

Correct label	0	0	0	1	1	1
False label during training	1	1	1	0	0	0
Probability of being the *positive* class	0.65	0.57	0.72	?	0.35	0.23

(i) What would be the AUC (computed with the correct labels) when the classifier is perfectly trained on the false data? And why?

[2 marks]

(ii) Provide the range of possible values for the missing output (labeled '?') that would be produced by the classifier in (i). Explain why.

[2 marks]

(iii) What would be the AUC (computed with the correct labels) of a random classifier trained on the falsely labeled data? Why?

[2 marks]

(b) Considering performing clustering on the following data.

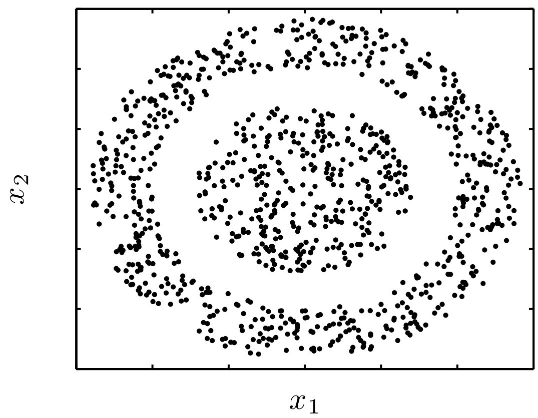

Figure 2: Data for clustering.

Outline what would happen if we fit a *Gaussian mixture model* to this data. How will it split/group the data?

[4 marks]

(c) Outline a strategy to apply the kernel trick to Gaussian mixture model.

[4 marks]

(d) Outline how and why cross-validation can be used to select the number of clusters for a Gaussian mixture model.

[6 marks]