UNIVERSITY OF CAMBRIDGE

MAT3, MAMA, MGM3

MATHEMATICAL TRIPOS Part III

Wednesday, 5 June, 2019 1:30 pm to 4:30 pm

PAPER 218

STATISTICAL LEARNING IN PRACTICE

Attempt no more than ***FOUR*** *questions.*

There are ***SIX*** *questions in total.*

The questions carry equal weight.

STATIONERY REQUIREMENTS
Cover sheet
Treasury Tag
Script paper
Rough paper

SPECIAL REQUIREMENTS
None

You may not start to read the questions printed on the subsequent pages until instructed to do so by the Invigilator.

1

We run a clinical trial to investigate the effect of a treatment on 100 ill patients. We split patients into 4 groups of equal size according to their sex (`F` or `M`) and whether they underwent the treatment or not (`1` and `0`); each response variable is the proportion of healthy patients of the corresponding group at the end of the trial. The data is stored in data frame `treatments`. Consider the following (shortened) `R` code:

```
> treatments
  sex treatment healthypatients numpatients prop
1   F         0              19          25 0.76
2   F         1              14          25 0.56
3   M         0              17          25 0.68
4   M         1              18          25 0.72
> treatments.glm <- glm(prop ~sex+treatment, family=binomial,
weights = numpatients, data=treatments)
> summary(treatments.glm)
...
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.8537     0.3770   2.265   0.0235 *
sexM          0.1854     0.4310   0.430   0.6671
treatment1   -0.3698     0.4317  -0.857   0.3917
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 2.5323  on 3  degrees of freedom
Residual deviance: 1.6102  on 1  degrees of freedom
...
```

(a) Write down the algebraic form of the fitted model. Show that the distribution of each observation belongs to an exponential dispersion family. What are its natural and dispersion parameters and its variance function?

(b) For each of the fitted coefficients, interpret the meaning of their sign in terms of the fitted values. Conclude that the fitted model does not respect the order `prop[2]` $<$ `prop[3]` $<$ `prop[4]` $<$ `prop[1]` of the responses.

(c) Propose an extension of fitted model [within logistic regression models] with an extra parameter and state how the code above should be changed to fit it. Does it recover all the order relationships between entries of `prop`? Explain.

(d) State the deviance of model `treatments.glm` in terms of its fitted proportions and the observed proportions. Assume that these fitted proportions are close enough to the observed proportions. Show that the deviance of model `treatments.glm` is approximately the generalised Pearson statistic. Hence, approximate the distribution of the generalised Pearson statistic, and conclude whether there is evidence of overdispersion in a test with significance level 5% approximately. [You may use any result for deviance without proof.]

2

(a) Define the *Akaike Information Criterion* (AIC) and the *Bayesian Information Criterion* (BIC), introducing all necessary notation.

(b) For what main purpose are AIC and BIC used? For each of the two criteria, give a practical motivation to use it over the other. Mathematically, how do their expressions differ? Briefly comment on the practical effect of this difference.

A practitioner has data $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^n$, $n, p \geqslant 2$, X full rank, and would like to relate Y to X by a linear regression model without intercept and with Gaussian noise with mean 0 and variance 1. They consider $K \geqslant 2$ models with vectors of regression coefficients $\beta^{k)}, k = 1, \dots, K$. Let $1 \leqslant p^* < p$ be fixed throughout and assume the models satisfy the following: for any $k \in \{1, \dots, K\}$, $\beta^{k)}$ has exactly $p - p_k$ coordinates that are always identically 0 for some $p_k \in [p^*, p]$; and, $p_k = p^*$ for one and only one $k \in \{1, \dots, K\}$.

(c) State the log-likelihood and maximum likelihood estimator of a generic model.

(d) Assume $p \leqslant n$ and that the data is generated by the model with $p_k = p^*$. Find the distribution of the AIC and BIC of a generic model. Hence, show the following statements:

(i) if $n > e$, the practitioner selects the true model when they choose the model with either minimal expected AIC or minimal expected BIC;

(ii) assume $p^* = 1$ and $p = 2$; then, if $n > e^2$, the probability that the practitioner chooses the wrong model is fixed with n if they use standard AIC whilst, if they use standard BIC, it is strictly smaller and vanishes as $n \to \infty$.

3

(a) In the context of generalised linear models (GLMs), define *overdispersion* and give a reason that may cause it. Give a generalisation of a GLM that may account for overdispersion when such reason is present in the original GLM.

Suppose a football manager is interested in knowing how the number of goals their top 10 players scored varies across competitions and top two rivals. The data set goals contains the following variables: player (factor with levels 1, . . . , 10); numgoals (numerical valued), competition (factor with levels A, B and C); and, rival (factor with levels R and S). They run the following (shortened) code:

```
> head(goals)
  player numgoals competition rival
1      1        4           A     R
2      1        2           A     S
3      1        0           B     R
4      1        0           B     S
5      1       20           C     R
6      1       10           C     S
> goals.glmer <- glmer(numgoals ~ competition + rival + (1|player),
                       data = goals, family = "poisson")
> summary(goals.glmer)
...
Random effects:
Groups Name        Variance Std.Dev.
player (Intercept) 0.1168   0.3418
Number of obs: 60, groups:  player, 10

Fixed effects:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.7287     0.1856   3.926 8.65e-05 ***
competitionB  -0.8899     0.2465  -3.610 0.000306 ***
competitionC   1.4961     0.1472  10.167  < 2e-16 ***
rivalS         0.4385     0.1124   3.901 9.57e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Correlation of Fixed Effects:
            (Intr) cmpttB cmpttC
competitinB -0.387
competitinC -0.648  0.488
rivalS      -0.368  0.000  0.000
```

(b) Write down the algebraic form of the fitted model and estimated coefficients. How does it differ from a negative binomial model?

(c) What is the meaning (from a modelling perspective) of including covariate player in the way it appears in the call to glmer and why you think the manager did so?

How do you interpret the estimated fixed coefficient `rivalS` and what rival does it suggest to be the strongest?

(d) Can we expect to be able to use the usual deviance-based test for the model above to test for overdispersion in our data? Briefly comment why yes or no. Describe in detail the construction of an approximate 95% confidence level test for overdispersion using parametric bootstrap and the output above.

(e) Consider the following code:

```
> goals.glm <- glm(numgoals ~competition + rival, data = goals,
family = "poisson")
> customcorr(goals.glm)
              (Intercept)  competitionB  competitionC
competitionB       -0.480
competitionC       -0.804         0.488
rivalS             -0.457     -1.75e-16     -2.25e-16
```

The function `customcorr` returns the correlation matrix of the estimated coefficients of the fitted model `goals.glm`. From a modelling perspective, what can you conclude by comparing its output to that before part (b)?

4

(a) Cite the defining assumptions of Linear Discriminant Analysis (LDA). Under these and for a given prior distribution on the space of labels, find the Bayes classifier in terms of a discriminant function justifying every step in your calculations. Find the classification boundaries, identifying the parameters that define them, and conclude the Bayes classifier is a linear classifier.

Assume we have i.i.d. data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \{1, \ldots, L\}, n > L$, and a new observation $x \in \mathbb{R}^p$ to classify.

(b) State the form of the LDA classifier. Argue whether it may be more or less robust to outliers than the (soft margin and linear) support vector machine classifier when $L = 2$ [in your argument you may quote without proof any result from the course].

(c) Assume $p > n$. Is the LDA classifier well defined? Justify your answer. Is it well defined if the sample covariance matrix is regularised by adding to it a strictly positive (constant) multiple of the identity matrix? Justify your answer. Name a practical procedure to find a value of this constant for prediction purposes.

(d) Show that the LDA classifier can be modified to be nonlinear by using a positive definite kernel function. Remember to justify that the resulting modification is well defined and nonlinear. [You may quote any theorem from the course without proof.]

5

Each row of the data frame `letter` contains 16 different attributes of a pixel image of one of the 26 capital letters in the English alphabet, together with the letter label itself. A practitioner builds a classifier using the following (shortened) `R` code:

```
> head(letter)
  lettr xbox ybox width    ...
1     L    3    8     4    ...
2     N    2    2     3    ...
3     I    1    3     2    ...
4     X    5    7     7    ...
5     Q    2    2     2    ...
6     G    5   10     6    ...

> x <- as.matrix(letter[,-1])
> for(col in 1:16) x[,col] <- x[,col]/max(x[,col])
> y <- model.matrix(~lettr-1, data=letter)
> layer1 <- layer_dense(units = 100, activation = 'relu',
input_shape = c(16))
> layer2 <- layer_dense(units = 75, activation = 'relu')
> layer3 <- layer_dense(units = 50, activation = 'relu')
> layer4 <- layer_dense(units = 26, activation = 'softmax')
> letter.n <- keras_model_sequential(list(layer1, layer2, layer3, layer4))
> compile(letter.n, optimizer='sgd', loss='categorical_crossentropy',
metrics='acc')
> fit(letter.n, x, y, batch_size=1, epochs=5)
```

(a) Write down algebraically the model fitted in `letter.n`. How many parameters are in the model [you may leave the answer as a sum]?

(b) What does the `for` loop do and why do you think the practitioner included it? The practitioner would like to use a sigmoid activation function in one and only one of the layers `layer1`, `layer2`, `layer3`. In which would you suggest them to use it? Justify your answer.

(c) Define the *cross-entropy loss function* algebraically. Write down the steps of the numerical method instructed by the arguments `optimizer='sgd', loss='categorical_crossentropy'` and `batch_size=1, epochs=5` in the last two lines of code, introducing any necessary notation. At what step of the numerical method is back-propagation used?

(d) Consider a feed-forward neural network with the following characteristics: at least two hidden layers; at least two neurons in every layer including the output layer; no bias nodes in any layer; fully connected; and, identity activation functions in all layers except in the output layer, which has softmax output function. Assume that the two nodes in the output layer have fixed and distinct vectors of parameters, and that you are given a data set of size n. Show that for any such network there are infinitely many minimisers of the cross-entropy loss function. Hence, between the arguments `batch_size=1` and `batch_size=`n, which would you choose to fit the model and why? [You may quote any theorem from the course without proof.]

6

In this question you may assume that the mean of any stochastic process is 0 and that this is known. For a moving average model of order 1 (MA(1)), we denote its MA parameter by θ and its white noise parameter by σ^2.

(a) Define a *moving average model of order q* (MA(q)). Compute the autocovariance function of an MA(1) process in terms of θ and σ^2. Thus, show that if an MA(1) process is Gaussian, θ and σ^2 are not identifiable without further conditions.

(b) Define an *invertible* MA(q) process. Using the defining MA(1) identity, prove that an MA(1) process is invertible if $|\theta| < 1$.

(c) Assume we have $n \geqslant 2$ observations $X_1, \ldots, X_n$ of a Gaussian MA(1) process.

(i) What is the distribution of X_1? Identify its parameters in terms of θ and σ^2. Show that the covariance matrix of the observations is positive definite; you should use the defining MA(1) identity along the way.

(ii) Propose two ways to find an estimator of θ. Given any of these two estimators, find an estimator of σ^2. Justify every step in your calculations. State the asymptotic properties of all these estimators, citing one assumption.

(d) The data frame `ts` contains a time series. Consider the following (shortened) `R` code:

```
> arima(ts, order=c(1,0,0))
...
Coefficients:
         ...  intercept
      0.4773     0.0079
s.e.  0.0278     0.0728

sigma^2 estimated as 1.452:  log likelihood = -1605.55,  aic = 3217.1
> arima(ts, order=c(0,0,1))
...
Coefficients:
         ...  intercept
      0.9147     0.0084
s.e.  0.0130     0.0613

sigma^2 estimated as 1.026:  log likelihood = -1432.9,  aic = 2871.79
```

Based on the output, select a model. Give its name and the reason for selecting it. Based on the output, construct a candidate for an approximate confidence interval for the non-intercept parameter of your selected model. Which of the following three statements about your interval do you expect to be true? (i) It roughly has the correct significance level; (ii) it is too wide to have the right significance level; (iii) it is too narrow to have the right significance level. Justify your answer.

END OF PAPER