UNIVERSITY OF CAMBRIDGE

MAMA/218, NST3AS/218, MAAS/218, MAAMM2/SLP

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 12 June 2025 1:30 pm to 4:30 pm

PAPER 218

STATISTICAL LEARNING IN PRACTICE

Before you begin please read these instructions carefully

Candidates have THREE HOURS to complete the written examination.

Attempt no more than **FOUR** questions.
There are **FIVE** 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 An analytics company wishes to analyse the effect of money spent on two different advertising services `advert1` and `advert2` on the number of visitors to a website. The company collects for each of 40 websites the total amount (in hundreds of pounds) spent on each advertising service, and the total number of visitors to the website (over a fixed time period), collected in the dataset `websites`.

```
> head(websites, 6)
  visitors  advert1  advert2
        60     0.89     1.24
        74     1.13     0.62
        53     0.84     0.57
        50     1.21     0.55
        60     0.57     0.74
        72     1.13     0.96
```

Throughout this question assume that the values of `advert1` and `advert2` are deterministic (fixed design). Three models were fit to the dataset as given by the `R` code below.

```
> model1 <- glm(visitors~advert1+advert2, family = poisson, data = websites)
> summary(model1)$coefficients
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    3.8879     0.1020  38.1226    <2e-16
advert1        0.1923     0.0886   2.1708    0.0299
advert2        0.1430     0.0714   2.0018    0.0453

> model2 <- glm(visitors~advert1+advert2, family = quasipoisson, data = websites)
> summary(model2)$coefficients
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    3.8879     0.1305  29.7965    <2e-16
advert1        0.1923     0.1134   1.6967    0.0982
advert2        0.1430     0.0914   1.5646    0.1262

> model3 <- glm.nb(visitors~advert1+advert2, data = websites)
> summary(model3)$coefficients
             Estimate Std. Error  z value  Pr(>|z|)
(Intercept)    3.8879     0.1259  30.8860    <2e-16
advert1        0.1917     0.1094   1.7516    0.0798
advert2        0.1437     0.0887   1.6198    0.1053
```

(a) State the statistical model being fitted in `model1`. Give an interpretation of the fitted coefficients of `(Intercept)` and `advert1` in `model1` (assuming the model is correct).

(b) State the statistical model being fitted in `model2`.

(c) What is the value outputted by the following line of `R` code?

```
> sum(residuals(model1, type="pearson")^2)
```

[You may leave your answer as an unsimplified numerical expression.]

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) State the name given to the model fit by `glm.nb`. Construct a test of the null of `model1` against the alternative of `model3`, giving the p-value of your test. You may find the following `R` outputs helpful.

```
> 1-pchisq(AIC(model1)-AIC(model3), df=1)
0.1595522
> 1-pchisq(AIC(model1)-AIC(model3)+1, df=1)
0.08437928
> 1-pchisq(AIC(model1)-AIC(model3)+2, df=1)
0.04608555
> 1-pchisq(AIC(model1)-AIC(model3)+4, df=1)
0.0144816
```

(e) Suppose the analyst knows that there are unobserved independent random variables λ_i with mean $\alpha_i\nu_i$ and variance $\alpha_i^2\nu_i$, for constants $\alpha_i, \nu_i > 0$, and that the number of visitors to the ith website conditional on λ_i, are independently Poisson(λ_i) distributed. Suppose the analyst also knows that the common mean function across all the three models `model1`, `model2` and `model3` is correct. Suggest, with reason, which of the three p-values for the coefficient `advert1` provided in the `summary` outputs is most appropriate for testing the null of the coefficient of `model1` being zero in the following two cases: (i) $\nu_i = \nu$ does not depend on i; (ii) $\alpha_i = \alpha$ does not depend on i. [You may assume that large sample approximations may be validly applied in both cases, and may use any results from the course.]

(f) The analyst now decides that they are unhappy to model the mean function as in `model1`, `model2` and `model3`, and so instead decides to fit a CART (regression) decision tree on the data. They perform the following `R` code on a smaller dataset, using only the first three datapoints in the `websites` dataset.

```
> train <- websites[1:3,]
> fit.dt <- rpart(visitors~advert1+advert2, data=train, cp=0, maxdepth=1, minbucket=1)
> single.test.datapoint <- data.frame(advert1=0, advert2=0)
> predict(fit.dt, newdata=single.test.datapoint)
```

Find the numerical output given by the `predict` command in the final line of code, and explain why this estimator is likely to be a heavily biased (in comparison to its variance) estimate for the expected number of visitors to a website with no money spent on both of the two advertising services.

2 Suppose we have $Y \in \mathbb{R}^n$ and $X \in \mathbb{R}^{n \times p}$ as a vector `Y` and matrix `X` in `R`. The function `glmnet` can be used to fit elastic nets (without an intercept term) of the form

$$\hat{\beta}_{\alpha,\lambda} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^p} \left(\frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \left\{ \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 \right\} \right),$$

for $\lambda \geqslant 0$, $\alpha \in [0,1]$.

(a) State how the ridge and Lasso estimators are related to the elastic net.

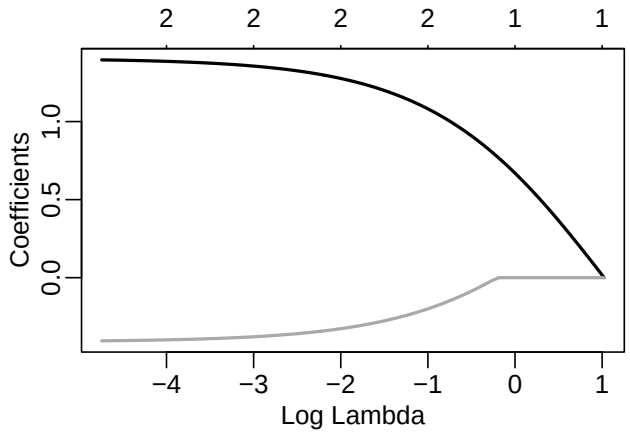
(b) Suppose that $X^\top X = n\, I_p$. Compute an explicit formula for $\hat{\beta}_{\alpha,\lambda}$ in terms of the OLS estimator $\hat{\beta}_{\mathrm{OLS}}$ (and α, λ). Hence show that

$$\min\left(|(\hat{\beta}_{0,\lambda})_j|,\, |(\hat{\beta}_{1,\lambda})_j|\right) \leqslant |(\hat{\beta}_{\alpha,\lambda})_j| \leqslant \max\left(|(\hat{\beta}_{0,\lambda})_j|,\, |(\hat{\beta}_{1,\lambda})_j|\right),$$

for all $\lambda \geqslant 0$, $\alpha \in [0,1]$, $j = 1, \ldots, p$. For each $\alpha \in (0,1]$, define $(\lambda^*_\alpha)_j$ to be the minimum value of $\lambda \geqslant 0$ such that $(\hat{\beta}_{\alpha,\lambda})_j = 0$. Show that if $(\hat{\beta}_{\mathrm{OLS}})_j \neq 0$ then the function $\alpha \mapsto (\lambda^*_\alpha)_j$ is strictly decreasing. Briefly explain the behaviour of $(\lambda^*_\alpha)_j$ as $\alpha \downarrow 0$.

(c) Suppose $p = 2$, $n = 100$, and $X^\top X = n\, I_p$. Consider the following `R` output.

```
> unregularised.fit <- glmnet(X, Y, family="gaussian", alpha=0.5, lambda=0, intercept=FALSE)
> coef(unregularised.fit)[,]
(Intercept)          V1          V2
   0.000000        1.41       -0.41
> en.fit <- glmnet(X, Y, family="gaussian", alpha=0.5, intercept=FALSE)
> plot(en.fit, xvar="lambda")
```



Find the explicit equations for the two curves in the plot.

(d) Suppose now $p = 2$, $X^\top X = n \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$ for some $\rho \in (-1, 1)$, and the OLS estimator $(\hat{\beta}_{\mathrm{OLS}})_j > 0$ for $j = 1, 2$. Define $\lambda^\dagger$ to be the minimum value of $\lambda \geqslant 0$ for which $(\hat{\beta}_{1,\lambda})_j = 0$ for *some* $j \in \{1, 2\}$, and $\lambda^\ddagger$ to be the minimum value of $\lambda \geqslant 0$ for which $(\hat{\beta}_{1,\lambda})_j = 0$ for *all* $j \in \{1, 2\}$. Find explicit expressions for $\lambda^\dagger$ and $\lambda^\ddagger$ in terms of $\hat{\beta}_{\mathrm{OLS}}$ (and ρ), and show that the difference $\lambda^\ddagger - \lambda^\dagger$ depends only on $\hat{\beta}_{\mathrm{OLS}}$.

3 Consider the K class classification problem for data $(X, Y) \in \mathbb{R}^p \times \{1, \ldots, K\}$. Suppose $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. copies drawn from the same distribution as (and independent of) (X, Y). Given a classifier $h_n : \mathbb{R}^p \to \{1, \ldots, K\}$ trained on the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ let $R(h_n) := \mathbb{E}[\ell(h_n(X), Y)]$ where ℓ denotes the misclassification loss.

(a) Define the conditional class probabilities $p_k(x)$ for $k = 1, \ldots, K$. Define the Bayes classifier h^* and Bayes risk R_{Bayes}. State what it means for a classifier h_n (trained on the n observations above) to be consistent.

(b) Define the L-nearest neighbour classifier $h_n^{L\text{NN}}$ (constructed using the training data $(X_1, Y_1), \ldots, (X_n, Y_n)$). Briefly state how the choice of L dictates a bias–variance tradeoff. [You need not introduce or state any formal results here.]

(c) In the case $K = 2$, and if there exists some $c > 0$ such that $\frac{1}{2} + c \leqslant p_{h^*(x)}(x) \leqslant 1 - c$ for all $x \in \mathbb{R}^p$, show that $h_n^{1\text{NN}}$ is not consistent.

(d) Show that

$$\lim_{n\to\infty} R(h_n^{1\text{NN}}) \leqslant 2R_{\text{Bayes}} - \frac{K}{K-1}\{R_{\text{Bayes}}\}^2.$$

*[**Note:** Throughout this question you may assume that p_k satisfies $\lim_{n\to\infty} \mathbb{E}[p_k(X_{(1)})f(X)] = \mathbb{E}[p_k(X)f(X)]$ for any bounded (measurable) function $f : \mathbb{R}^p \to \mathbb{R}$ and any $k \in \{1, \ldots, K\}$, where $X_{(1)}$ denotes the nearest neighbour of $(X_1, \ldots, X_n)$ to X. You may also ignore any potential measurability issues throughout.]*

*[**Note:** Throughout this question you may ignore ties.]*

4 Each row of a data frame `text` contains 40 different numeric attributes of an online review that is either positive or negative, together with the positive (`P`) or negative (`N`) label itself, and the date of the review.

```
> dim(text)
 10000 42
> text[1:4, ]
     date  review  att1  att2  att3  att4  att5  ...
  Jan2012       P   0.0   0.7   0.1   0.9  -0.9  ...
  Jan2012       P  -0.2   0.4  -0.3   0.1   0.3  ...
  Jan2012       N   0.9  -0.9   0.7   0.8   0.2  ...
  Feb2012       N   0.1   0.7  -0.2   0.1  -0.4  ...
> text[9999:10000, ]
     date  review  att1  att2  att3  att4  att5  ...
  May2025       N   0.0  -0.1   0.9  -0.3  -0.8  ...
  May2025       P   0.4   0.7  -0.8   0.0  -0.2  ...
```

The goal is to build a classifier using the dataset `text` to predict the state (positive/negative) of reviews on an independent test dataset `new.text` consisting of new recent reviews (of which we have access to the same attributes and labels).

```
> dim(new.text)
 200 40
> new.text[1:4, ]
     date  review  att1  att2  att3  att4  att5  ...
 June2025       N   0.4  -0.8   0.4  -0.8  -0.7  ...
 June2025       P   0.6   0.4  -0.7   0.9   0.9  ...
 June2025       P   0.2  -0.1   0.4   1.0   0.2  ...
 June2025       N  -0.3  -0.5  -0.9  -0.1   0.2  ...
```

A classifier is built using the code below.

```
> x_train <- as.matrix(text[,3:42])
> y_train <- model.matrix(~ review-1, data=text)
> x_test <- as.matrix(new.text[,3:42])
> y_test <- model.matrix(~ review-1, data=new.text)
> nn.model <- keras_model_sequential() %>%
       layer_dense(units = 40, activation = "relu", input_shape = c(40)) %>%
       layer_dense(units = 20, activation = "relu") %>%
       layer_dense(units = 2, activation = "softmax")
> compile(nn.model, optimizer="sgd", loss="categorical_crossentropy", metrics="accuracy")
> nn.fit <- fit(nn.model, x_train, y_train, epochs=5, batch_size=2,
                validation_data = list(x_test, y_test))
> metric1 <- tail(1-nn.fit$metrics$val_accuracy,1)
> # Comment: above line calculates one minus "accuracy" on "validation_data"
> metric1
 0.1923
```

(a) Write out the mathematical model being fit in `nn.fit`, clearly defining all necessary quantities and functions. [You may specify the numeric class labelling of `review` however you wish.] How many parameters are there in the model in total?

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) State the mathematical form of the loss function being minimised to train `nn.fit`. Briefly explain how overfitting is avoided here. State what is being calculated in the 'forward pass' step of the training algorithm, and how many times in fitting `nn.fit` above this 'forward pass' step is performed.

(c) Suggest two reasons why AIC may not be appropriate to compare `nn.fit` with the logistic classifier.

(d) Another model is proposed that is identical to `nn.fit` except for that all instances of the ReLU activation function are replaced with the activation function $\sigma(\eta) = \eta\,\Phi(\eta)$, where Φ denotes the cumulative distribution function of the standard normal distribution. Give one reason why this alternative activation function may be more preferable over ReLU, and one reason why it may be less preferable.

The following `R` code is then performed.

```
> errors <- rep(0, 10000)
> for (i in 1:10000) {
    nn.model.i <- nn.model
        nn.fit.i <- fit(nn.model.i, x_train[-i,], y_train[-i,], epochs=5, batch_size=2,
                        validation_data = list(x_train[i,], y_train[i,]))
        errors[i] <- tail(1-nn.fit.i$metrics$val_accuracy,1)
  }
> metric2 <- mean(errors)
> metric2
 0.0405
```

(e) State the name given to the quantity `metric2`, and give its algebraic form. [You may make reference to previous parts of the question, and do not need to specify how relevant weights are estimated.] State a reason why `metric2` may not be reasonable to compute in practice.

(f) Suggest a reason why `metric2` does not appear to act as a good approximation of `metric1`.

5 A dataset `finances` is collected by following 10 people over the course of 10 years, recording each year the amount of money they earned, and the amount of money they spent (both in units of tens of thousands of pounds). For the ith person on the jth year, let X_{ij} be the total money earned and Y_{ij} be the total money spent ($i = 1, \ldots, 10$, $j = 1, \ldots, 10$).

```
> head(finances,12)
   person year earned spent
1       1    1   3.74  1.97
2       1    2   3.86  2.45
3       1    3   3.88  2.19
4       1    4   4.06  2.42
5       1    5   4.06  2.46
6       1    6   4.13  2.38
7       1    7   4.34  2.43
8       1    8   4.58  2.78
9       1    9   4.60  2.85
10      1   10   4.68  2.82
11      2    1   5.81  6.62
12      2    2   5.78  6.72
> levels(finances$person)
 [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

A data analyst fits the following two models.

```
> model1 <- glm(spent ~ earned, family = gaussian, data = finances)
> model2 <- lmer(spent ~ earned + (1|person), data = finances, REML = FALSE)
```

[Note that fitting a glm with `family = gaussian` fits an identical model to that of `lm`, but outputs the null and residual deviances, and the AIC.]

(a) Write down the mathematical models being fitted in `model1` and `model2`. [You need not specify any estimates of the parameters in the models.]

(b) Suggest one reason why `model1` may be unreasonable, and how `model2` addresses your stated issue.

A reduced output of the two fitted models is given below (with some estimates removed, either replaced with `???` or `V1`, `V2`, `V3`, `V4`).

```
> summary(model1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.68782    0.56115   3.008  0.00334 **
earned      0.54218    0.09722   5.577 2.17e-07 ***

    Null deviance: 203.27 on 99 degrees of freedom
Residual deviance: 154.30 on 98 degrees of freedom
AIC: 333.16

> summary(model2)

     AIC      BIC  logLik deviance df.resid
   -82.4    -72.0    45.2      ???      ???
```

[QUESTION CONTINUES ON THE NEXT PAGE]

```
Random effects:
 Groups  Name        Variance Std.Dev.
 person  (Intercept)    ???       V3
 Residual               ???       V4
Number of obs: 100, groups: person, 10

Fixed effects:
           Estimate Std. Error t value
(Intercept)      V1        ???     ???
earned           V2        ???     ???

Correlation of Fixed Effects:
      (Intr)
earned   ???
```

(c) State the training error (with respect to the squared error loss) of `model1`.

The data analyst also performs the following `R` code:

```
> a <- model.matrix( ~ earned, data = finances)
> A <- t(a) %*% a
> A
100.0 562.600
562.6 3331.778
> y_on_person <- lm(spent ~ person - 1, data = finances)
> x_on_person <- lm(earned ~ person - 1, data = finances)
> gy <- summary(y_on_person)$coefficients[,1]
> gx <- summary(x_on_person)$coefficients[,1]
> length(gy)
> 10
> sum(gy)
47.381
> sum(gx)
56.26
> t(gy) %*% gy
244.1738
> t(gx) %*% gx
332.2809
> t(gx) %*% gy
274.898
```

(d) Explain why for the matrix $\texttt{A} = A \in \mathbb{R}^{2\times 2}$ in the outputs above the following two results hold: (i) $A_{11} = 100$ and (ii) $0.1 \times A_{12} = \texttt{sum(gx)}$. *[Hint: Note that, given a fitted* `model`*, the call* `summary(model)$coefficients[,1]` *extracts its fitted coefficients.]*

(e) Show that $(\texttt{V1}, \texttt{V2}, \texttt{V3}, \texttt{V4})$ (each given in the output of `summary(model2)` above) satisfies

$$(\texttt{V1}, \texttt{V2}, \texttt{V3}, \texttt{V4}) \in \operatorname*{argmin}_{(\alpha,\beta,\sigma,\tau)\in\mathbb{R}\times\mathbb{R}\times(0,\infty)\times(0,\infty)} \omega(\alpha, \beta, \sigma, \tau),$$

for a function ω that you should specify. *[Note: The function ω may be written in terms of numerical and matrix expressions that need not be simplified, but must not contain expressions explicitly in terms of the original data (Y_{ij}, X_{ij}).]*

END OF PAPER