UNIVERSITY OF CAMBRIDGE

MAMA/218, NST3AS/218, MAAS/218

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 11 June 2026 2:00pm to 5:00pm

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 (a) The *inverse Gaussian distribution* $\mathrm{IG}(\mu, \lambda)$ with parameters $\mu > 0$ and $\lambda > 0$ has probability density function

$$f(x; \mu, \lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left(-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right)$$

for $x > 0$. Show that $\{\mathrm{IG}(\mu, \lambda) : \mu > 0, \lambda > 0\}$ forms an exponential dispersion family with natural parameter $\theta = -1/(2\mu^2)$, and give the forms of the the cumulant function $K(\theta)$ and the dispersion parameter ϕ. If $Y \sim \mathrm{IG}(\mu, \lambda)$, show that $\mathbb{E}(Y) = \mu$ and find $\mathrm{Var}(Y)$ and the variance function $V(\mu)$. Give the form of the canonical link function $g(\mu)$ for an inverse Gaussian generalised linear model (GLM). *[Hint: You may use without proof the fact that* $\int_0^\infty u^{-1/2} e^{-u}\,\mathrm{d}u = \sqrt{\pi}$*.]*

(b) A data set contains information on the response times (in seconds) of 30 subjects across four hand-eye coordination tasks with different difficulty levels. Each subject attempted each task twice.

```
> head(data, n=10)

    response difficulty subject
1  0.1751991          1       1
2  0.4760675          2       1
3  0.6515696          3       1
4  0.5488463          4       1
5  0.1733278          1       1
6  0.3373843          2       1
7  0.6545860          3       1
8  0.5465658          4       1
9  0.1920279          1       2
10 0.6435905          2       2
```

A researcher fits an inverse Gaussian GLM in order to understand the dependence of the `response` time on the `difficulty` of each task. The default link function for fitting inverse Gaussian GLMs in R is $g(\mu) = 1/\mu^2$, and the model is summarised as follows.

```
> model1 <- glm(response~difficulty, family=inverse.gaussian, data=data)
> summary(model1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.0827     1.3577   6.690 1.58e-10 ***
difficulty   -1.4323     0.4254  -3.367 0.000886 ***

    Null deviance: 258.03  on 239  degrees of freedom
Residual deviance: 245.00  on 238  degrees of freedom
AIC: -42.874
```

[QUESTION CONTINUES ON THE NEXT PAGE]

Write down a mathematical description of `model1`. Interpret the estimate of the coefficient associated with `difficulty`. Is this surprising? Explain your answer.

(c) Provide a formula for the Pearson residuals associated with an inverse Gaussian GLM in terms of the responses Y_i, the fitted values $\hat{\mu}_i$ and the dispersion estimate $\hat{\phi}$. Give a formula for $\hat{\phi}$. Calculate an approximate value for $\hat{\phi}$ in `model1`, assuming a condition which you should state.

(d) The researcher then fits another model to the data:

```
> model2 <- glmer(response~difficulty + (1|subject),
                  family=inverse.gaussian, data=data)
> summary(model2)

      AIC       BIC    logLik -2*log(L)  df.resid
  -45.366   -31.443       ???       ???       ???

Random effects:
 Groups   Name        Variance Std.Dev.
 subject  (Intercept) 1.7533   1.3241
 Residual             0.7591   0.8712

Fixed effects:
            Estimate Std. Error t value Pr(>|z|)
(Intercept)   8.9966     1.1987   7.505 6.13e-14 ***
difficulty   -1.2317     0.3746  -3.288  0.00101 **
```

Write down a mathematical description of `model2`. Explain why the researcher may have fitted this model. *[Hint: `glmer` fits a generalised linear mixed effects model (GLMM) with normal random effects, and the default link function for `glmer` is $g(\mu) = 1/\mu^2$.]*

(e) Formulate a hypothesis test of H_0: GLM against H_1: GLMM in terms of the parameters of the two models. Propose a test statistic and calculate its value using the summary outputs for `model1` and `model2`. What is the p-value of this test? What do you conclude? You may find the following code and outputs useful.

```
> pchisq(3.366 - 0.874, df = 1, lower.tail = FALSE)
0.1144262
> pchisq(5.366 - 0.874, df = 1, lower.tail = FALSE)
0.03405382
> pchisq(3.366 - 0.874, df = 2, lower.tail = FALSE)
0.2876531
> pchisq(5.366 - 0.874, df = 2, lower.tail = FALSE)
0.1058217
```

2 (a) Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent and identically distributed random variables taking values in $[0,1]^d \times \{1,2\}$, where $d \in \mathbb{N}$. Define the L-nearest neighbour classifier that uses the norm $\|x\|_\infty = \max_{1 \leqslant j \leqslant d} |x_j|$ at a point $x \in [0,1]^d$. Give the form of the corresponding estimated conditional class probabilities $\hat{p}_k(x)$ for $k \in \{1,2\}$.

(b) Let $e_1(x) = \mathbb{E}\{\hat{p}_1(x) \mid X_1, \ldots, X_n\}$. Show that

$$\mathbb{E}\Big\{\big(\hat{p}_1(x) - e_1(x)\big)^2\Big\} \leqslant \frac{1}{L}.$$

(c) Show that if X_1 has density function $f(x)$ with $\inf_{x \in [0,1]^d} f(x) \geqslant M > 0$, then for $\{L/(nM)\}^{1/d} \leqslant \delta \leqslant 1$, we have

$$\mathbb{P}\Big(\|X_{(L)} - x\|_\infty > \delta\Big) \leqslant \min\left\{1, \frac{nM\delta^d}{(nM\delta^d - L)^2}\right\},$$

where $X_{(L)}$ is the Lth closest data point to x under $\|\cdot\|_\infty$. You may wish to use *Chebyshev's inequality*, which states that if Z is a real-valued random variable with $\mathbb{E}(Z) = \mu$ and $\mathrm{Var}(Z) \leqslant \sigma^2$, then $\mathbb{P}(|Z - \mu| > t) \leqslant \sigma^2/t^2$.

(d) Using the result from part (c), or otherwise, deduce that

$$\mathbb{E}\Big(\|X_{(L)} - x\|_\infty\Big) \leqslant \left(\frac{2L}{nM}\right)^{1/d} + \frac{2}{L}.$$

(e) Suppose that $|p_1(x) - p_1(x')| \leqslant C\|x - x'\|_\infty$ for $x, x' \in [0,1]^d$, where $C > 0$. Use part (d) to provide an upper bound for $\mathbb{E}\{|e_1(x) - p_1(x)|\}$.

(f) Derive an upper bound for $\mathbb{E}\{|\hat{p}_1(x) - p_1(x)|\}$. Hence, provide conditions on L which ensure that the risk (under misclassification loss) of the L-nearest neighbour classifier converges to the Bayes risk as $n \to \infty$. You may quote any result you need from lectures, without proof.

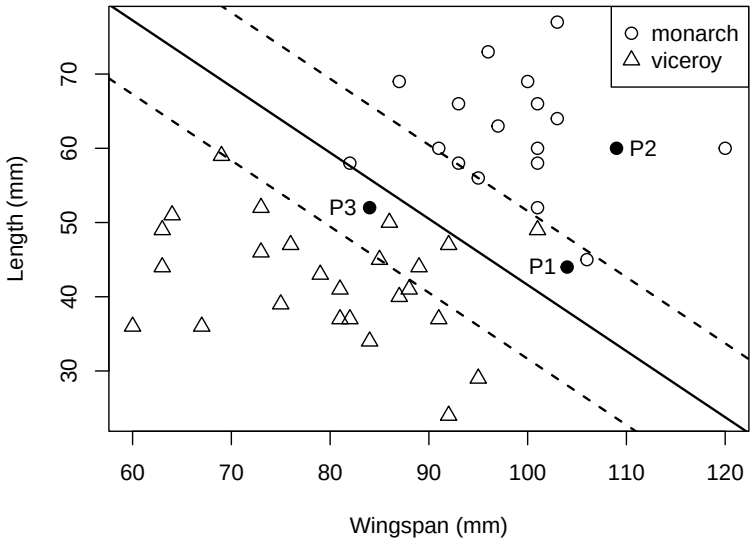
3 A researcher collects data on $n = 45$ butterfly specimens, with the aim of building a classifier to predict the `species` (either `monarch` or `viceroy`) based on measurements of the `wingspan` and body `length` (both measured in millimetres). A linear support vector machine (SVM) classifier is fitted to the data.

```
> head(data, n=5)

  wingspan  length  species
1       93      59  monarch
2      106      66  monarch
3       92      37  viceroy
4       79      34  viceroy
5      112      40  monarch

> model1 <- svm(species~wingspan + length, data=data,
                kernel="linear", cost=0.01, scale=FALSE)
```

This SVM classifier is determined by the solution $\left(\hat{\alpha}, \hat{\beta}, \hat{\xi}\right)$ to the optimisation problem

$$\begin{aligned} \underset{\alpha\in\mathbb{R},\, \beta\in\mathbb{R}^2,\, \xi\in\mathbb{R}^n}{\text{minimise}} \quad & \left(\|\beta\|_2^2 + C\sum_{i=1}^{n} \xi_i \right), \\ \text{subject to} \quad & Y_i\left(\alpha + X_i^{\mathsf{T}}\beta\right) \geqslant 1 - \xi_i \text{ and } \xi_i \geqslant 0, \text{ for all } i \in [n], \end{aligned}$$

where $Y_i = +1$ if $\texttt{species}_i = \texttt{monarch}$ and $Y_i = -1$ if $\texttt{species}_i = \texttt{viceroy}$, with $X_i = (\texttt{wingspan}_i, \texttt{length}_i)^{\mathsf{T}}$, and where $C \geqslant 0$. The fitted SVM is plotted below.



[QUESTION CONTINUES ON THE NEXT PAGE]

(a) What are the quantities ξ_i called? Give names and formulae (in terms of $\hat{\alpha}$ and $\hat{\beta}$) for the solid and dashed diagonal lines on the plot.

(b) For each of the points `P1`, `P2` and `P3` (indicated on the plot, with label `monarch`), give a plausible value for the corresponding quantity $\hat{\xi}_i$. Which of these points are support vectors?

(c) Suppose that the researcher now increases the value of C and fits another linear SVM. Explain what will happen to the number of misclassified points, the width of the margin, and the number of support vectors.

(d) Show that the training error $\widehat{\mathrm{Err}}_{\mathrm{tr}}$ of the SVM classifier satisfies

$$\widehat{\mathrm{Err}}_{\mathrm{tr}} \leqslant \frac{1}{n}\sum_{i=1}^{n} \hat{\xi}_i.$$

(e) The researcher runs the following code.

```
> n <- nrow(data)
> errs <- rep(0, n)
> for (i in 1:n) {
    model_i <- svm(species~wingspan + length, data=data[-i,],
                   kernel="linear", cost=0.01, scale=FALSE)
    errs[i] <- predict(model_i, data[i,]) != data$species[i]
}
> err <- sum(errs) / n
```

What is the name of this procedure? Writing $\widehat{\mathrm{Err}}$ for the `err` value computed by the researcher, show that

$$\widehat{\mathrm{Err}} \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\big\{Y_i(\hat{\alpha} + X_i^{\mathsf{T}}\hat{\beta}) \leqslant 1\big\}.$$

Interpret this result. Suggest a procedure for calculating $\widehat{\mathrm{Err}}$ that is more computationally efficient than the one used by the researcher.

4 (a) Let $\mathcal{X}$ be a set. What does it mean for a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to be a *kernel*? State the Moore–Aronszajn theorem in terms of objects ϕ, $\mathcal{H}$ and $\langle \cdot, \cdot \rangle$ which you should define.

(b) Consider the 'alphabet' $\mathcal{A} = \{\mathtt{a}, \mathtt{b}\}$ and let $\mathcal{A}_L$ be the set of length-L strings of characters from $\mathcal{A}$. For example, if $L = 4$ then $\mathtt{abaa} \in \mathcal{A}_L$. Define $g : \mathcal{A}_L \to \mathbb{R}$ by

$$g(x, y) = \sum_{r=1}^{L} \mathbf{1}\{x_r = y_r\}$$

so that, for example, $g(\mathtt{abaa}, \mathtt{bbab}) = 2$. Show that g is a kernel.

(c) Let $x_1, \ldots, x_n \in \mathcal{X}$. Define a linear function $\Sigma : \mathcal{H} \to \mathcal{H}$ by

$$\Sigma v = \frac{1}{n} \sum_{i=1}^{n} \big\langle \phi(x_i), v \big\rangle \phi(x_i),$$

for $v \in \mathcal{H}$. Show that if $v \in \operatorname{argmax}_{u \in \mathcal{H}, \langle u, u \rangle = 1} \big\langle u, \Sigma u \big\rangle$, then $\Sigma v = \lambda v$ for some $\lambda \geqslant 0$. *[Hint: You may wish to use the method of Lagrange multipliers.]*

(d) Define $K \in \mathbb{R}^{n \times n}$ by $K_{ij} = k(x_i, x_j)$, and assume that K is positive definite. Show that $\lambda > 0$ and deduce that there exists $\alpha \in \mathbb{R}^n$ such that

$$v = \sum_{i=1}^{n} \alpha_i \phi(x_i).$$

(e) Show that $K\alpha = n\lambda\alpha$ and $\alpha^{\mathsf{T}} K \alpha = 1$.

(f) Given a new test point $x \in \mathcal{X}$, the *first kernel principal component score* of x is defined as $s(x) = \langle v, \phi(x) \rangle$. Explain why computing $s(x)$ by first finding v and $\phi(x)$ is *not* generally feasible. Propose a more practical method for computing $s(x)$.

(g) Give two reasons why a researcher might prefer to use kernel principal components analysis (PCA) rather than standard PCA.

5 Consider the following data set consisting of medical records for $n = 768$ adult women with Akimel O'odham native American (also known as Pima Indian) heritage. Each row in the data set consists of eight covariates (`pregnant`, `glucose`, ..., `age`) and one label (`diabetes`). A researcher wants to use this data to classify `diabetes` status as `positive` or `negative` based on the covariates.

```
> head(data, n=5)

  pregnant glucose pressure triceps insulin mass pedigree age  diabetes
1        6     148       72      35       0 33.6    0.627  50  positive
2        1      85       66      29       0 26.6    0.351  31  negative
3        8     183       64       0       0 23.3    0.672  32  positive
4        1      89       66      23      94 28.1    0.167  21  negative
5        0     137       40      35     168 43.1    2.288  33  positive
```

The data set is divided randomly into equally sized training and testing sets, yielding matrices `x_train`, `y_train`, `x_test` and `y_test`. The covariates in `x_train` and `x_test` are then standardised to have zero mean and unit variance. The researcher fits a model on the training data using the following R code.

```
> model1 <- keras_model_sequential() %>%
    layer_dense(units=24, activation="relu", input_shape=c(8)) %>%
    layer_dense(units=16, activation="relu") %>%
    layer_dense(units=2, activation="softmax")

> model1 %>% compile(loss="categorical_crossentropy",
    optimizer=optimizer_sgd(learning_rate=0.04))

> model1 <- model1 %>% fit(x_train, y_train, epochs=100,
    batch_size=16, validation_data=list(x_test, y_test))
```

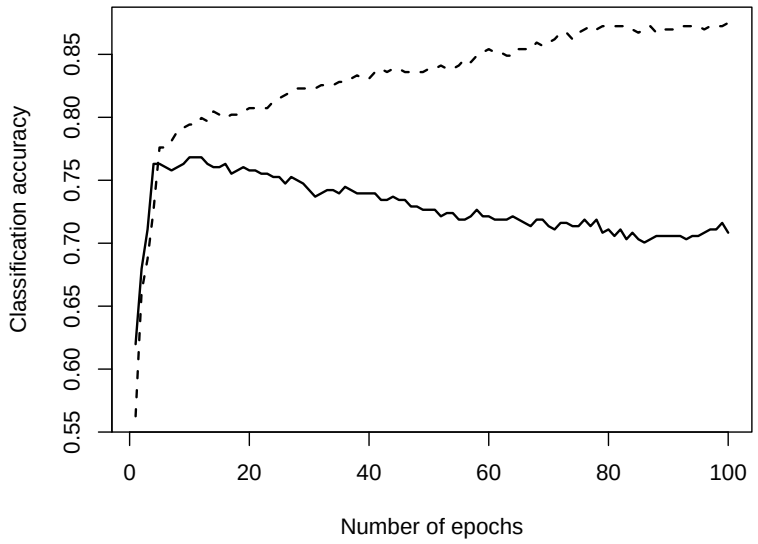
(a) Give a mathematical description of `model1`. How many parameters does it have? Write down its loss function and interpret it in terms of the likelihood of a distribution. You may code the `diabetes` labels in whichever manner is most convenient.

(b) Define what is meant by *stochastic gradient descent*. How many mini-batches are there in each epoch when training `model1`? What is the total number of times each parameter is updated during training of `model1`?

(c) Give three reasons why parameter estimates obtained by mini-batch stochastic gradient descent might not be maximum likelihood estimators.

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) The following figure displays accuracy on the training and testing sets as a function of the number of completed epochs during training. Identify which line (dashed or solid) is the training accuracy and which is the testing accuracy, and describe the phenomenon depicted in the plot when the number of epochs is large. Suggest a sensible value for the number of epochs the researcher should use when training the classifier, and interpret this approach in terms of regularisation.



END OF PAPER