UNIVERSITY OF CAMBRIDGE

MAMA/218, NST3AS/218, MAAS/218

MAT3

MATHEMATICAL TRIPOS Part III

Thursday 6 June 2024 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 **ALL** questions.
There are **FOUR** 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 scientist is investigating the effect of six different doses of a pesticide on salmonella bacteria. They applied the same dose to three replicate plates containing salmonella and recorded the number of salmonella colonies eradicated on each plate by the pesticide. The following abbreviated R-code was used to analyse the data.

```
> head(salmonella, 9)

    colonies dose
 1        15    0
 2        21    0
 3        29    0
 4        16   10
 5        18   10
 6        21   10
 7        16   33
 8        26   33
 9        33   33

> model1 <- glm(colonies ~ dose, family="poisson", data = salmonella)
> summary(model1)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.3219950  0.0540292  61.485   <2e-16 ***
dose        0.0001901  0.0001172   1.622    0.105
...
Null deviance: 78.358  on 17  degrees of freedom
Residual deviance: 75.806  on 16  degrees of freedom
AIC: 172.34
...
> df1 <- model1$df.residual
> Xsq <- sum(residuals(model1,type="pearson")^2)
> phi.hat <- Xsq/df1
> 1-pchisq(Xsq,df1)
[1] 4.908651e-11

> model2 <- glm.nb(colonies ~ dose, data = salmonella)
> summary(model2)
...
```

[QUESTION CONTINUES ON THE NEXT PAGE]

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) 3.3169342  0.1087599  30.498   <2e-16 ***
dose        0.0002097  0.0002474   0.848    0.397
...
    Null deviance: 18.672  on 17  degrees of freedom
Residual deviance: 18.011  on 16  degrees of freedom
AIC: 141.66
...
              Theta:  9.23
          Std. Err.:  3.99
...
```

(a) Write down the model mathematically that has been fitted in `model1`. State the log-likelihood and the deviance. Give an interpretation of the fitted intercept and dose coefficients on the expected number of eradicated colonies.

(b) Write down mathematical expressions for the terms `Xsq` and `phi.hat` computed in the `R`-code and name the quantities they are estimating. Find an approximation for `phi.hat` from the output of `model1`.

(c) State the null and alternative hypotheses for the hypothesis test performed in the `R`-code. Using the output of the test, explain why the scientist has fitted `model2`. Does `model2` improve on `model1` according to the model outputs?

(d) Recall that a negative binomial GLM uses the log-link function and has variance function $V(\mu) = \mu + \theta^{-1}\mu^2$ for a parameter $\theta > 0$, which can be estimated by maximum likelihood.

For which value of θ does `model2` fit a Poisson GLM? To test whether the parameter θ is necessary, the scientist carried out a likelihood ratio test using the following `R` commands.

```
> test_stat <- 2*(logLik(model2) - logLik(model1))
> pval <- 1 - pchisq(test_stat, df = 1)
```

Explain why this test is not valid, and state in detail how a valid test can be carried out using the parametric bootstrap. You do not need to state the likelihood of the negative binomial distribution.

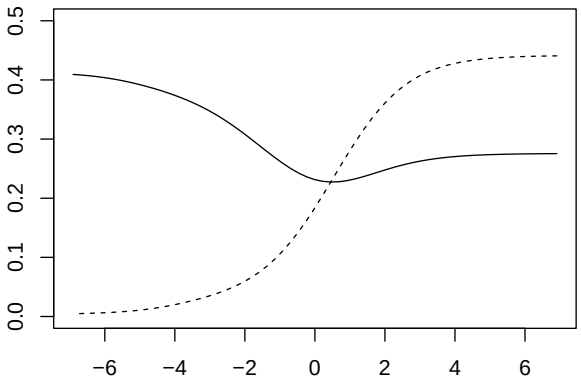
2 Suppose we have a set of training data $(x_1, Y_1), \ldots, (x_n, Y_n)$ and an independent set of test data $(x_1, Y_1^*), \ldots, (x_n, Y_n^*)$, where for $i = 1, \ldots, n$

$$Y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,1),$$
$$Y_i^* = f(x_i) + \varepsilon_i^*, \quad \varepsilon_i^* \overset{\text{i.i.d.}}{\sim} N(0,1),$$

with $x_i \in \mathbb{R}^p$ and an unknown regression function f. Suppose that $\sum_{i=1}^n x_i = 0$. Let Y, $Y^* \in \mathbb{R}^n$ be the vectors with components Y_i, Y_i^*, $i = 1, \ldots, n$, respectively, and let X be the matrix with the x_i as rows.

A data analyst has fitted a ridge regression model to the data using the code shown below (recall that `glmnet` fits ridge regression for a grid of penalisation parameters).

```
> fit <- glmnet(X, Y, family="gaussian", intercept=TRUE, alpha=0)
> Y.hat <- predict(fit, X, type="response")
> err.train <- colMeans((Y - Y.hat)^2)
> err.test <- colMeans((Y.star - Y.hat)^2)
> plot(log(fit$lambda), err.test, type="l", ylim=c(0,0.5), ylab="", xlab="")
> points(log(fit$lambda), err.train, type="l", lty="dashed")
```



(a) State the penalised optimisation problem of ridge regression, and give a closed form expression for its solution when the penalisation parameter is non-vanishing. State a formula for the fitted values obtained by ridge regression.

(b) Suggest a criterion for choosing the penalisation parameter in (a). Explain shortly how bias and variance of the regression model are affected by the penalisation parameter.

[QUESTION CONTINUES ON THE NEXT PAGE]

(c) Formulate the coordinate descent algorithm for finding a solution of the optimisation problem in (a). State explicit solutions for the coordinate updates.

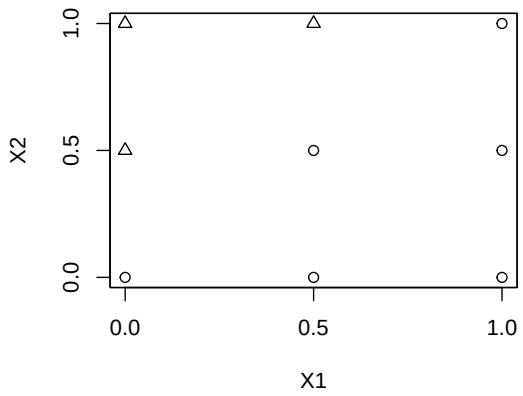
(d) What do the solid and dashed curves in the plot represent according to the code? State mathematical expressions for the values of the solid and dashed curves as the horizontal axis of the figure approaches $+\infty$. Using the plot, state a mathematical expression for the value of the solid curve as the horizontal axis of the figure approaches $-\infty$.

(e) Suppose a regression algorithm produces fitted values of the form $HY \in \mathbb{R}^n$ for some $H \in \mathbb{R}^{n\times n}$ depending on the x_i, $i = 1, \ldots, n$. Show that

$$\mathbb{E}\left[\|Y - HY\|_2^2 + 2\text{trace}(H)\right] = \mathbb{E}\left[\|Y^* - HY\|_2^2\right].$$

Find the matrix H for the ridge regression model. Together with the plot, argue that the training error is not useful for judging how well the ridge regression model generalises on unseen data.

3 Suppose we are given i.i.d. training data (X_i, Y_i), $i = 1, \ldots, 9$, with $X_i \in [0,1]\times[0,1]$ and $Y_i \in \{$'triangle','circle'$\}$ as given in the plot below (`X1` and `X2` are the coordinates of the X_i). Let X be the matrix with the X_i as rows and let Y be the vector with components Y_i.



(a) Give a detailed description of the CART algorithm for computing a classifier for this data, and define its prediction and training errors. Carefully introduce all necessary notation. *Hint: The cost function for classification is*

$$Q = \frac{N(U)}{N(R)}G(U) + \frac{N(V)}{N(R)}G(V) - G(R),$$

for regions U, V and R as well as functions N and G, which you should specify.

(b) Show that the cost function Q in (a) is non-positive for all regions considered in (a). *Hint: The function $p \mapsto p(1-p)$ is concave on $[0,1]$.*

(c) Explain shortly how the CART algorithm is used to compute the output of the following abbreviated `R`-code:

```
> randomForest(X,Y,mtry=2)

Type of random forest: classification
      Number of trees: 500
No. of variables tried at each split: 2
...
Confusion matrix:
  1 2 class.error
1 3 3   0.5000000
2 2 1   0.6666667
```

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) Name the procedure that is performed by the following R-code:

```
> err <- NULL
> for(i in 1:9) {err[i] <- predict(randomForest(X[-i,],Y[-i]),X[i,])!=Y[i]}
> mean(err)
```

Name an alternative method that approximates the result of `mean(err)` without having to fit the 9 random forests.

(e) Sketch the decision boundaries for the data above for

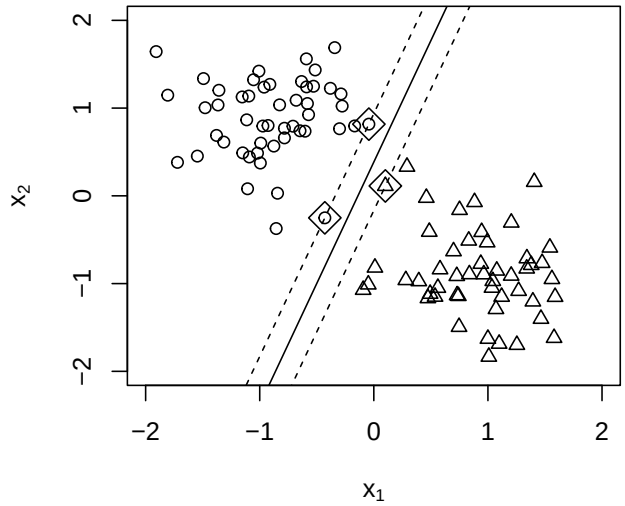
(i) a nearest neighbour classifier,

(ii) an unpruned maximal tree classifier,

(iii) a random forest classifier with default parameters.

Explain what you expect as training errors for the three classifiers on this dataset.

4 A statistician has trained a support vector classifier on data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with $X_i \in \mathbb{R}^p$ and $Y_i \in \{-1, 1\}$. Using the fitted classifier, they produced the following plot.



(a) Write down a mathematical expression for the support vector classifier and state the associated penalised optimisation problem with a penalisation parameter $\lambda > 0$.

In the following, let $\hat{\gamma} \in \mathbb{R}^{1+p}$ denote a solution of the optimisation problem in (a).

(b) Define *separating hyperplane*. How many support vectors are there according to the plot, and how are they related to $\hat{\gamma}$? How are the solid and dashed lines in the plot related to $\hat{\gamma}$? Discuss if we obtain the same solid and dashed lines as seen in the plot when the penalisation parameter in (a) vanishes.

(c) Let $X_i^* = (1, X_i) \in \mathbb{R}^{1+p}$ and $Z_i = \|X_i^*\|_2^{-1} X_i^*$, and suppose that $Y_i Z_i^T \hat{\gamma} \geqslant 1$ for all $i = 1, \ldots, n$. Consider the following algorithm for computing $\hat{\gamma}$:

(i) Initialise at $\gamma^{(0)} \in \mathbb{R}^{1+p}$.

(ii) For $m = 0, 1, 2, \ldots$: If there exists Z_i such that $Y_i Z_i^T \gamma^{(m)} \leqslant 0$, then set $\gamma^{(m+1)} = \gamma^{(m)} + Y_i Z_i$. Otherwise, return $\gamma^{(m)}$.

Show that $\|\gamma^{(m+1)} - \hat{\gamma}\|_2^2 \leqslant \|\gamma^{(m)} - \hat{\gamma}\|_2^2 - 1$. Conclude that the above algorithm takes no more than $\|\gamma^{(0)} - \hat{\gamma}\|_2^2$ many steps to converge to $\hat{\gamma}$.

(d) Define the logistic classifier for a two-class classification problem, and state the optimisation problem that needs to be solved to find its parameters. Prove that there is no solution for the data in the plot above. Discuss how we can modify the fitting procedure to find an approximate solution.

END OF PAPER