UNIVERSITY OF
CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Thursday, 9 June, 2022 9:00 am to 12:00 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 **SIX** questions in total.
The questions carry equal weight.

STATIONERY REQUIREMENTS	**SPECIAL REQUIREMENTS**
Cover sheet	None
Treasury tag	
Script paper	
Rough paper	

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

1 The times to failure of 61 components on a ship were recorded, along with the type of component (labelled `type1`, `type2`, `type3`) and the position of the component (`posout` being 1 for outside and 0 for inside) in the ship. In the (shortened) `R` code below, `mod1` fits an exponential generalised linear model (GLM), and `mod2` fits a gamma GLM to these data.

```
> summary(mod1,dispersion=1)
Call:
glm(formula = times ~ type + pos, family = Gamma, data = ship)
...
 Coefficients:
            Estimate Std. Error z value Pr(>|t|)
(Intercept)  0.15032    0.03607   4.168  3.08e-05  ***
type2       -0.01228    0.04776  -0.257  0.797
type3        0.08323    0.06417   1.297  0.195
posout       0.02368    0.04556   0.520  0.603
...
(Dispersion parameter for Gamma family taken to be 1)
Null deviance: 21.061 on 60 degrees of freedom
Residual deviance: 18.185 on 57 degrees of freedom
AIC: 304.48
...
 > summary(mod2)
Call:
glm(formula = times ~ type + pos, family = Gamma, data = ship)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15032    0.02009   7.481  5.02e-10 ***
type2       -0.01228    0.02661  -0.461  0.6463
type3        0.08323    0.03575   2.328  0.0235 *
posout       0.02368    ?         0.933  0.3548
...
(Dispersion parameter for Gamma family taken to be 0.3103711)
Null deviance: 21.061 on 60 degrees of freedom
Residual deviance: 18.185 on 57 degrees of freedom
AIC: 304.48
```

(a) Show that the gamma distribution with probability density function parametrised as

$$f(y;\alpha,\gamma) = \left(\frac{\gamma}{\alpha}\right)^{\gamma} \frac{1}{\Gamma(\gamma)} y^{\gamma-1} e^{-\frac{\gamma}{\alpha}y}, \quad y > 0, \alpha, \gamma > 0,$$

is an exponential dispersion family in the unknown parameters α, γ. Identify the mean and variance functions and the canonical link function.

[QUESTION CONTINUES ON THE NEXT PAGE]

(b) State the log-likelihood of the response variables $Y \in \mathbb{R}^{61}$ in the gamma GLM with the canonical link function and compute the Fisher information matrix with respect to the coefficient $\beta \in \mathbb{R}^4$ of the predictors.

(c) Show that we obtain the same maximum likelihood estimators $\hat{\beta}$ in model `mod1` and model `mod2`. State a large sample statement for the asymptotic distribution of $\hat{\beta}$ in the Gamma GLM using your results in (b). Use this to explain how the missing standard error for the variable `posout` in model `mod2` can be obtained from the one in model `mod1` (you do not have to compute its numerical value).

(d) Explain the test that is carried out by the following `R` code, specifying the null and alternative hypotheses and the expression for the test statistic. Is the test valid? What can we conclude from this test?

```
> 1-pchisq((21.061-18.185)/0.31,df=3, lower.tail=TRUE)
[1] 0.02582103
```

2 A data analyst has been asked to build a classifier from training data $X_1, \dots, X_n \in \mathbb{R}^p$ with given associated labels $Y_1, \dots, Y_n \in \{0,1\}$. An (old) book on statistical learning suggests to implement a support-vector machine solving the constrained optimisation problem

$$\min_{\beta \in \mathbb{R}^p} \|\beta\| \quad \text{subject to} \quad Y_i(X_i^\top \beta) \geqslant 1, \quad i = 1 \dots, n.$$

However, the analyst could not find a solution of this optimisation problem for his data set.

(a) What would you suggest to the analyst in terms of how the input data and the optimisation problem should be modified to fit a general hard-SVM?

(b) Define the hard-SVM classifier. Explain the main ideas of this classification approach.

The data analyst followed your suggestions in (a). Still, there was no solution found.

(c) Explain why the hard-SVM does not have a solution in this case. Give a relaxed formulation of the constrained optimisation problem in (a) to address this issue. Discuss how the tuning parameter of this relaxation influences the margin and how it is related to the hard-SVM.

After learning about the kernel trick, the analyst applied a feature map ϕ to the data, yielding $Z_i = \phi(X_i) \in \mathbb{R}^q$ for $q \in \mathbb{N}$. It turned out that for the transformed data the hard-SVM classifier now had a solution.

(d) Explain the kernel trick (also called kernel method) for SVMs and state the optimisation problem solved by a hard-SVM with a kernel, but without penalisation.

The data analyst then applied a logistic classifier to the transformed data, obtaining the following (shortened) R output.

```
> model2 <- glm(Y~Z, data = dat, family="binomial")
> summary(model2)
...
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    -9.004  12038.209  -0.001    0.999
X1              9.447  16708.962   0.001    1.000
X2             39.494  25259.925   0.002    0.999
...
```

(e) Why are the standard errors so large? What modification to the logistic regression will produce smaller standard errors and more stable estimates?

3 A researcher wants to build an email spam classifier based on a training set of $n = 500$ emails. They have hand-picked 10 words/symbols that they believe to have the highest discriminating power: `dollar`, `winner`, `password`, `edu`, `credit`, `discount`, `as`, `I`, `fun`, `trial`, and performed a logistic regression in `R`. Each row in the dataset represents one email. The first column (`spam`) encodes whether an email is spam (class 1) or not (class 0). The remaining 10 columns count the number of times a particular word/symbol appears in the email. Part of the `R` output is shown below.

```
> model.logit <- glm(spam ~ dollar + winner + password + edu + credit
           + discount + as + I  + fun + trial, family = binomial)
> summary(model.logit)
...
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.391451   1.447857  -3.724   0.0002 ***
dollar             1.859318   1.333052   1.395   0.1631
winner             5.680691   1.955451   2.905   0.0037 **
password           0.923072   1.268399   0.728   0.4668
edu               -6.890095   1.857351  -3.710   0.0002 ***
credit             2.269523   1.007462   2.253   0.0243 *
discount           1.198028   1.785944   0.671   0.5023
as                -3.176676   1.832839  -1.733   0.0831 .
I                 -1.866328   0.846315  -2.205   0.0274 *
fun                4.347929   1.104687   2.066   0.0388 *
trial              0.864456   1.774681   0.487   0.6262
...
```

(a) Write down the algebraic form of the fitted model `model.logit`. How would you interpret the coefficient `dollar`? Write down also the algebraic form of the logistic classifier $\hat{C}^{\text{logit}}$ and how it is fitted.

(b) Describe algebraically the decision boundary of this classifier. In practice, one may want to only classify an email as spam if the predicted spam probability is larger than a given threshold $q \in (0, 1)$. Let $\hat{C}_q^{\text{logit}}$ be the corresponding classifier. Show that $\hat{C}^{\text{logit}} = \hat{C}_{0.5}^{\text{logit}}$. What effect does varying q have on the decision boundary?

Instead of hand-picking significant words, the researcher now wants to include 5000 common English words into the classification. He realised that the logistic regression fit did not converge.

(c) State the log-likelihood of the model fitted in (a) and explain why logistic regression does not have a unique solution in this case.

(d) How can principal component analysis (PCA) help in fitting a well-defined logistic classifier? Explain the main ideas of PCA. Discuss why the principal components are different from hand-picking significant words and how the researcher can decide how many principle components should be considered.

[QUESTION CONTINUES ON THE NEXT PAGE]

(e) For data $X_1, \ldots, X_n \in \mathbb{R}^p$ and $d \leqslant p$ consider the optimisation problem

$$\underset{\mu, z_i, A}{\operatorname{argmin}} \sum_{i=1}^{n} \|X_i - \mu - Az_i\|^2$$

over $\mu \in \mathbb{R}^p$, $z_1, \ldots, z_n \in \mathbb{R}^d$ with $\sum_{i=1}^n z_i = 0$ and $A = (u_1, \ldots, u_d) \in \mathbb{R}^{p \times d}$, where $u_1, \ldots, u_d$ are non-vanishing orthogonal vectors in $\mathbb{R}^p$. Find first the solutions $\hat{\mu}$, $\hat{z}_i$ of this optimisation problem for fixed A and then determine $\hat{A}$.

4 A research team analysed the efficacy of educating students about the health impact of smoking. Four groups of students were compared according to their exposure to a television-based smoking prevention program (TV $\in \{0,1\}$) and/or to a school-based curriculum (SC $\in \{0,1\}$). Students belonging to the same school were assigned to the same group. There were 28 schools (recorded in a categorical variable named school), and each school was assigned randomly to a group. The tobacco and health knowledge for each student was measured before (PTHK) and after the study (THK). The research team fitted three different models to the data, producing the following (shortened) R output.

```
> lm1 <- lm(THK ~ PTHK + TV + SC,  data=smoking)
> summary(lm)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.73734    0.07866  22.088  < 2e-16 ***
PTHK         0.32525    0.02589  12.561  < 2e-16 ***
TV           0.04534    0.06518   0.696    0.487
SC           0.47987    0.06529   7.350 3.15e-13 ***
...
> lme1 <- lmer(THK ~ PTHK + TV + SC + (1 |school), data=smoking, REML=FALSE)
> summary(lme1)
...
    AIC      BIC   logLik deviance df.resid
  5381.3   5413.5  -2684.6   5369.3     1594
...
Random effects:
 Groups   Name        Variance Std.Dev.
 school   (Intercept) 0.0437   0.209
 Residual             1.6531   1.286
...
Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.78880    0.10696  16.724
PTHK         0.30973    0.02595  11.933
TV           0.02175    0.10530   0.207
SC           0.47023    0.10532   4.465
...
```

[QUESTION CONTINUES ON THE NEXT PAGE]

```
> lme2 <- lmer(THK ~ PTHK + TV + SC + (1 + PTHK|school), data=smoking, REML=FALSE)
> summary(lme2)
...
   AIC      BIC   logLik deviance df.resid
  5376.8   5419.9  -2680.4   5360.8     1592
...
Random effects:
 Groups   Name        Variance  Std.Dev. Corr
 school   (Intercept) 0.0001866 0.01366
          PTHK        0.0094648 0.09729  1.00
 Residual             1.6384188 1.28001
...
Fixed effects:
            Estimate Std. Error t value
(Intercept)  1.73214    0.09262  18.701
PTHK         0.30165    0.03208   9.403
TV           0.08805    0.09345   0.942
SC           0.51516    0.09340   5.516
...
```

(a) Write down algebraically the model fitted by `lme1` and state the estimated values of all parameters. How do you interpret the random and fixed effect intercepts in the output of `lme1` in the context of the study?

(b) Explain the modelling decision behind using `school` as a random effect in the models `lme1` and `lme2`. How does it affect the model fit over model `lm1`? Justify your answer using the `R` output.

(c) What is the estimated variance for the response variables in `lme1` according to the `R` output (you do not have to compute its numerical value)? Why is there no estimate column for the random effect in model `lme1`?

(d) The research team wants to use a statistical test to see if `lme1` or `lme2` fits the data better. Describe what changes algebraically in model `lme2` compared to `lme1`. Suggest a valid test and discuss in detail how the corresponding p-value can be computed.

5 (a) For a set of possible parameters $\theta_1, \ldots, \theta_p \in \mathbb{R}$ and family of statistical models M_S, where each M_S depends on the parameters θ_i with i in $S \subset \{1, \ldots, p\}$, define the *Akaike Information criterion* (AIC) and state the backward-selection rule for model selection by AIC. Which variable is selected below in the first step after calling the `step` function?

```
> step(lm(medv~., data=Boston))

Start:  AIC=1599.85
medv ~ crim + zn + indus + chas + nox + rm + age + dis + rad +
    tax + ptratio + lstat

          Df Sum of Sq   RSS    AIC
- indus    1      1.08 11350 1597.8
- age      1      1.69 11351 1597.9
<none>                 11349 1599.8
- chas     1    245.31 11595 1608.7
- tax      1    256.28 11606 1609.2
- zn       1    263.59 11613 1609.5
- crim     1    311.49 11661 1611.6
- rad      1    430.71 11780 1616.7
- nox      1    546.10 11896 1621.6
- ptratio  1   1157.70 12507 1647.0
- dis      1   1258.52 12608 1651.1
- rm       1   1744.36 13094 1670.2
- lstat    1   2733.54 14083 1707.0
...
```

(b) Explain how AIC is related to the bias-variance trade-off.

(c) Let M_1 be a normal linear model with $p+1$ parameters (for the unknown coefficient vector $\beta \in \mathbb{R}^p$ and the unknown error variance $\sigma^2 > 0$), and let M_2 be a normal linear model with additional q parameters. Let $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, respectively, be the maximum likelihood estimators for σ^2 in the two models. Consider now the following:

(i) State the log-likelihood in model M_1.

(ii) Show that the AIC value in model M_1 is given by

$$\mathrm{AIC}(M_1) = n(\log(2\pi\hat{\sigma}_1^2) + 1) + 2(p+1).$$

(iii) Conclude that M_2 has a smaller AIC value if $\hat{\sigma}_2^2/\hat{\sigma}_1^2 < e^{-2q/n}$.

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) Explain how the AIC value can be obtained from the following (shortened) R output (you do not have to compute its numerical value).

```
> lm1 <- lm(y ~ x1 + x2, data=dat)
> summary(lm1)
...
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.4681     0.1917 -18.089  < 2e-16 ***
x1            1.4902     0.1525   9.772 4.17e-16 ***
x2            5.1046     0.1626  31.401  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.359 on 97 degrees of freedom
Multiple R-squared:  0.9126,Adjusted R-squared:  0.9108
F-statistic: 506.5 on 2 and 97 DF,  p-value: < 2.2e-16
```

6 A data scientist has been given data $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$ for $p = 2$. He notices that the two predictors (columns of X) have suspiciously similar values. He decides to fit two different models to the data, a linear regression and a Bayesian regression. He obtained the following (shortened) R-output:

```
> lm1 <- lm(y ~ x1 + x2 - 1, data=dat)
> lm2 <- brm(y ~ x1 + x2 - 1, data=dat, family = gaussian,
+                  prior = c(prior(normal(0, 1))))
> summary(lm1)
...
Coefficients:
   Estimate Std. Error t value Pr(>|t|)
x1   -228.7      235.3  -0.972    0.360
x2    230.6      235.3   0.980    0.356

Residual standard error: 0.08145 on 8 degrees of freedom
Multiple R-squared:  0.9986,Adjusted R-squared:  0.9983
F-statistic:  2953 on 2 and 8 DF,  p-value: 3.35e-12
> summary(lm2)
..
   Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
x1     1.00      0.68    -0.33     2.37 1.00     1183     1476
x2     0.98      0.68    -0.38     2.30 1.00     1187     1481
..
```

(a) State the prior distribution in the Bayesian model. How are the estimates for the coefficients $x1$ and $x2$ obtained in the Bayesian model?

(b) Suppose now that $p \geqslant 1$ and that X is deterministic with full rank. Consider for β a prior distribution on $\mathbb{R}^p$ with Lebesgue-density $\pi(\beta) = \prod_{j=1}^p \phi(\beta_j)$ for a function $\phi : \mathbb{R} \to [0, \infty)$. State the log-likelihood of the observations and show that the density of the posterior distribution of β given the observations in Y is

$$\pi(\beta|Y) = C \exp\left(-\frac{1}{2}\langle X^\top X(\beta - \hat{\beta}), \beta - \hat{\beta}\rangle + \sum_{j=1}^p \log\phi(\beta_j)\right)$$

for the least-squares estimator $\hat{\beta}$ and a suitable constant $C > 0$.

(c) Show that Lasso and ridge regression estimators can be obtained as maximum-a-posteriori estimators from the posterior density in (b) with respect to different ϕ functions. Argue that ridge regression is also the posterior mean for the same ϕ function.

(d) Use (c) to explain why the coefficient estimates for $x1$ and $x2$ are different in the output of the linear and the Bayesian models.

(e) Give a 95% credible interval for $x1$ using the R output. Discuss the interpretation of this interval relative to a frequentist 95% confidence interval.

END OF PAPER