UNIVERSITY OF CAMBRIDGE

MATHEMATICAL TRIPOS Part III

Wednesday, 16 June, 2021 12:00 pm to 3:00 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
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

Consider $n \geqslant 2$ observations $(x_1, Y_1), \ldots, (x_n, Y_n) \in \mathbb{R}^2$ satisfying $\sum_{i=1}^n x_i = 0$ and $\sum_{i=1}^n x_i^2 > 0$. For $\underline{x} := (x_1, \ldots, x_n)^\top$, let $X := (\underline{x}, \underline{x})$ and $Y := (Y_1, \ldots, Y_n)^\top$. In the following we will consider the two problems of regressing Y on $\underline{x}$ and Y on X.

(a) Consider the penalised-loss function version of ridge regression. State the ridge solution to the problem of regressing Y on $\underline{x}$. Using the same regularisation parameter as when regressing Y on $\underline{x}$, find the ridge solutions to the problem of regressing Y on X.

(b) Consider the two problems of part (a) but in their constrained-optimisation form with a common constraint parameter t. Graphically illustrate their solutions as t varies.

(c) Consider the constrained-optimisation version of Lasso regression. State the two problems of Lasso regression of Y on $\underline{x}$ and Y on X with common constraint parameter. Derive analytically the set of solutions to the second problem in terms of the solution to the first problem.

(d) Graphically illustrate your result for part (c) as the common constraint parameter t varies.

(e) Assume that $Y_i \overset{ind.}{\sim} N(x_i\beta_0, \sigma^2)$ for some $\beta_0 \in \mathbb{R}$ and $\sigma^2 > 0$ unknown. Using your results from parts (a) and (c), compute the bias and variance of the ridge and Lasso predictions when regressing Y on X and Y on $\underline{x}$ with identical regularisation and optimisation constraint parameters, and compare the effect on each of them of including the second copy of $\underline{x}$ in the design matrix.

[*Hint: recall that in both ridge and Lasso regression the estimate for the intercept parameter* α *is* $\hat{\alpha} := n^{-1} \sum_{i=1}^n Y_i$.]

2

An app's analyst has data for the number of times 80 users click on a given service repeatedly offered to them. For each user they collected their age and fitness level (numerical variables) and geographic region and sex (categorical variables with R and 2 levels, respectively). The analyst stores all the data in data frame `data1` and runs the following commands in `R`.

```
> head(data1)
click user age fitness region sex
    0    1  28       2      1   F
    1    1  28       2      1   F
    0    1  28       2      1   F
    0    1  28       2      1   F
    0    1  28       2      1   F
    1    2  33       0      1   M
> mod1 <- glm(click~age+fitness+region+sex, family = binomial,
              data = data1)
```

The analyst decides to aggregate the data for each user so that they can analyse the proportion of clicks of each user relative to the number of times the service was offered to them. These two new variables are called `prop` and `noffers`, respectively, and the new data is stored in data frame `data2` (where with some abuse of notation the variables `age`, `fitness`, `region` and `sex` only contain one representative for each user). They decide to run the following commands.

```
> data2[1:2,]
prop noffers user age fitness region sex
 0.2       5    1  28       2      1   F
 0.5       2    2  33       0      1   M
> mod2 <- glm(prop~age+fitness+region+sex, family = binomial,
              data = data2, weights = noffers)
```

(a) Write down the algebraic form of the two fitted models. Argue why they give rise to the same fitted coefficients for the model parameters.

(b) What assumption on the dependence between data points in `data1` does `mod2` make? Explain whether you think this is a reasonable assumption or not.

The analyst decides to change the first model to incorporate a possible correlation $\rho \in [-1, 1]$ between any two individual responses of a given user.

(c) Compute the resulting mean and variance of the proportion of clicks of each user relative to the number of times the service was offered to them. Would a standard quasi-binomial model be an appropriate model to fit to the proportions? Justify your answer.

[QUESTION CONTINUES ON THE NEXT PAGE]

(d) The analyst assumes the proportions to be independent and fits a (rescaled) beta-binomial model to them (with all covariates). Write down the algebraic form of the model. Is it an appropriate model when only considering the means and variances derived in part (c)? Justify your answer. [*Hint: if X follows a binomial distribution with parameters m, p, and p follows a beta distribution with mean and variance parameters μ, θ, respectively, the marginal distribution of X when averaging over the distribution of p is a beta-binomial distribution with parameters m, μ, θ; such a beta-binomial distribution has mean μ and variance $(1 + (m-1)\theta/(1+\theta))\mu(1-\mu)/m$.*]

(e) Briefly compare the hierarchical forms of the fitted beta-binomial regression model and the following fitted model.

```
> glmer(prop~ age+fitness+region+sex + (1|user),  family = "binomial",
        weights = noffers)
```

3 An online content creator records how many times each of their 151 videos was viewed, the amount of time (in days) each video has been public and the investment in each video (in £100s). They store all the data into data frame `data` and run the following commands in `R`.

```
> head(data)
video nviews time investment
    1   6301  449        3.1
    2   8933  151       2.22
    3  77912  288         10
    4  11981  145      12.92
    5  21121  301       7.59
    6   8320   73       15.2
> attach(data)
> mod <- glm(nviews/time~ investment, family = "poisson")
> sum(residuals(mod, type="pearson")^2)/mod$df.residual
[1] 1.3119
> pnorm(1.644854)
[1] 0.95
```

(a) Explain in words why it is sensible that the creator input `nviews/time~ investment` rather than `nviews~ time+investment` to the first argument of the `glm` function.

(b) Define the concept of *overdispersion*. Construct a statistical test for whether `mod` suffers from overdispersion at 5% significance level, justifying its use. Is the null hypothesis rejected for this data and model? What is the implication on the confidence interval executed by the following commands and how would you resolve it?

```
> cs <-  summary(mod)$coefficients
> c(cs[1,1] - 1.96 * cs[1,2], cs[1,1] + 1.96 * cs[1,2])
```

The creator decides to increase the complexity of the model and includes categorical variable `video` as one of the covariates.

(c) Prove that the resulting model is non-identifiable.

Realising this, the creator creates a new categorial variable `genre` with 5 levels. They now wish to predict the success of a new video in three months time [we assume a month to be 30 days long]. For this, they categorise success in the following way: a video is of low, medium and high success if the number of times it is viewed belongs to $[0, 9999]$, $[10000, 49999]$ and $[50000, \infty)$, respectively.

(d) Write down the algebraic form of a classifier of success in three months taking `genre` and `investment` as the inputs and that is fitted from the data at hand. For a fixed genre, is the classifier linear or non-linear? Justify your answer.

4

Assume $Y \sim \text{Multinomial}(1, p_1, \ldots, p_L)$, where $p_l = p_l(x), l = 1, \ldots, L$, for some $x \in \mathbb{R}^p$ and $p, L \in \mathbb{N}$. Suppose we model this relationship via a fully-connected neural network; take the output function of any such model to be sofmax throughout the question.

(a) Write down the algebraic form of this model.

(b) Describe the logistic classification model as a fully-connected neural network.

(c) In a fully connected deep neural network, name a phenomenon that implies that the ReLU activation function is preferred to the sigmoid one.

Assume we have independent data $(x_1, Y_1), \ldots, (x_n, Y_n)$ and that we fit the parameters of a neural network from part (a) via maximisation of the log-likelihood minus an elastic-net penalty.

(d) Write down the optimisation problem and a numerical algorithm to solve it [it should address how to select the regularisation parameters, but need not include details of the passes through the network and you may take $\frac{d}{dx}|x|_{x=0} = 0$ for simplicity].

(e) An issue with the ReLU activation function is that some units may "die". Define the concept of a *dead ReLU unit* and modify the ReLU activation function to avoid this issue whilst preserving the superiority over the sigmoid function alluded to in part (c). Justify your chosen modification.

5

Let $(X, Y) \in \mathbb{R}^p \times \{0, 1\}, p \geqslant 3$, be a generic data point with X being sampled from a continuous distribution.

(a) Define the risk $R(\psi)$ [via the usual 0–1 loss] of a classifier ψ and find the expression for the Bayes classifier ψ^{Bayes}. Introduce any notation you need.

Assume we have n i.i.d. copies $(X_1, Y_1), \ldots, (X_n, Y_n)$ of (X, Y).

(b) Define the k-nearest neighbours classifier ψ^{kNN} and its risk $R_n(\psi^{\text{kNN}})$. Conditioning on $X = X_1$, prove that for any $n \geqslant 1$

$$R_n(\psi^{\text{1NN}}) \leqslant 2R(\psi^{\text{Bayes}}).$$

Assume the data is stored in `R` via objects `X` and `Y`. Consider the following block of code in `R`.

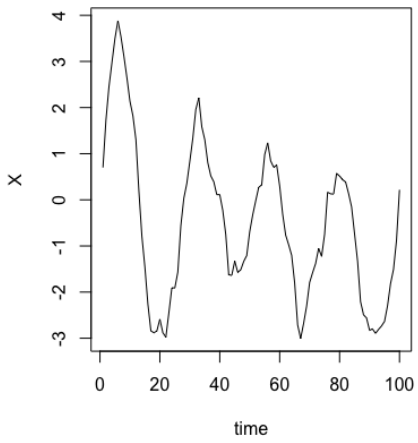
```
> ks <- seq(1, 100, by=1)
> ls <- rep(0, length(ks))
> for (k in ks)   ls[k] <- mean(knn.cv(X,Y,k) != Y)
```

(c) Describe in words, and with any necessary references to the previous lines, what the thrid line executes [you need not write down any algorithms]. How would you use `ls` to construct a classifier?

(d) Write down the algebraic form of an extension to the k-nearest neighbours classifier whose risk improves upon ψ^{kNN} for any $k \in \mathbb{N}$ under sufficient assumptions. Quote any result from the course that justifies your answer.

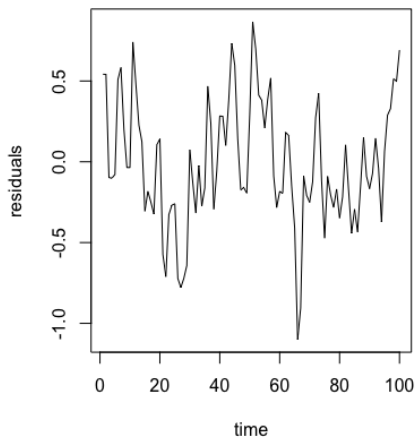
6

(a) Define (*weak*) *stationarity* for a time series $X := (X_t)_{t \in T}$. An analyst has the time series with 100 data points plotted below in `R`. They ask you to assess whether it is stationary. Does it seem to be so? Justify your answer.



(b) They wish to remove any possible trend and seasonality. Write down algebraically a way to do so [in doing so, you should identify the period of the season, for which you may assume the time window under consideration includes only full seasons]. Under what assumption on the data is your procedure sensible for removing seasonality after de-trending?

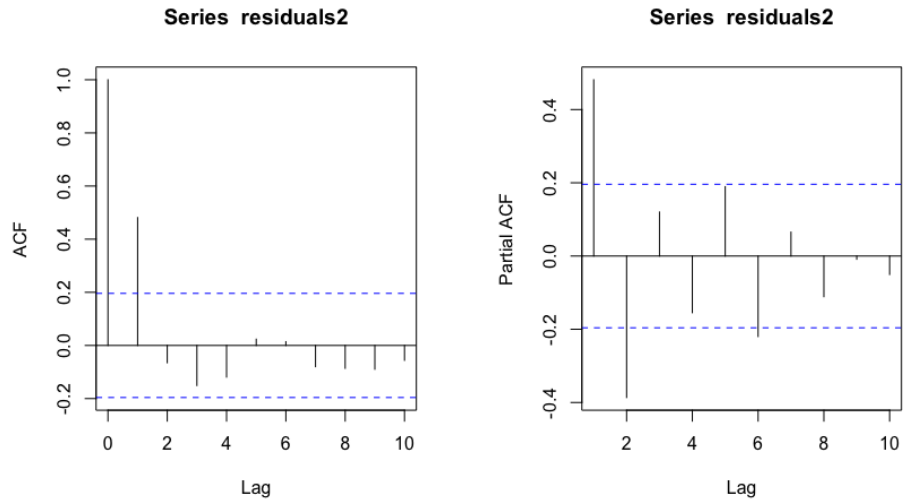
The following figure shows the remaining residuals stored in object `residuals`.



[QUESTION CONTINUES ON THE NEXT PAGE]

(c) Thinking about the business cycles, the analyst realises that there must be another cycle whose period is 50. Does this mean a violation of the assumption on the original data referred to in part (b)? The analyst decides to de-season `residuals` by applying the transformation $1 - B^{50}$ to the data, where B is the backshift operator. Find two drawbacks in applying this transformation.

The following figures show the sample autocorrelation function (ACF) and the partial ACF of the remaining de-seasoned residuals using the procedure in part (b) stored in object `residuals2`.



(d) Which ARMA model provides a good fit to `residuals2` according to the figures above?

The analyst then runs the following code in `R`.

```
model<-auto.arima(residuals2, ic="aic")
fc<-forecast(model$fitted, h = 1, level = 95)
```

(e) Briefly explain in words what the two lines of code execute.

The analyst does not trust the nominal coverage of the confidence interval shown when executing `plot(fc, shaded=F)`.

(f) Why would they not trust it? Explain a way to address part of the underlying issues.

END OF PAPER