CS7DS3 Assignment 1

February 26, 2020

To be submitted by **12 noon** on **Wednesday 6th May, 2020**. Submit your solutions on Blackboard. Please remember to put your **name** and **student number** on the front of your script.

Please show your work. Where code has been used, an outline description, along with relevant results, is sufficient.

If you have any questions about the assignment, email me: arwhite@tcd.ie or post a question to the message board on Blackboard.

Question 1

Recall Question 2 for Assignment 1, where a company wished to model online traffic for a website using a Poisson distribution. Recall that a Poisson distribution with rate parameter θ has pdf

$$f(y|\theta) = \theta^y \frac{\exp\{-\theta\}}{y!}.$$

The company records covariates $x = x_1, \ldots, x_P$, that they would like to use to predict the number of unique visits that their website home page receives. The predictor variables could include, e.g., website layout, time of day, social media activity, etc., but their specific form is not important.

a) Express the pdf $f(y|\theta)$ in exponential family form, that is, in the form:

$$f(y|\theta) = h(y)g(\theta)\exp\{\phi(\theta)s(y).\}$$

Explicitly identify h, g, ϕ and s. [5 marks]

b) Explain how the natural parameter ϕ can be used as a link function to incorporate a single predictor variables x into a Poisson regression model. Express the rate parameter θ as function of x , i.e., $\phi^{-1}(x)$. [5 marks]

c) Use the link function $\phi(x)$ to specify the log-likelihood of the Poisson regression model. Compute the first derivative of the model with respect to β_1, the regression coefficient for x. [5 marks]

d) The company fit a Poisson regression model to data using a single covariate x. This variable recorded which of two advertising strategies (Strategy 0 or Strategy 1) were used on the preceding day. Traffic to the website between 12 and 12:15pm was then recorded for the following day.

The following output from the regression was produced:

```
Call:
glm(formula = y ~ strategy, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.64317  -0.81635   0.09564   0.59717   1.49792

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.3001     0.1924   1.559    0.119
strategy1     1.6094     0.2108   7.634 2.27e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 120.941  on 39  degrees of freedom
Residual deviance:  42.344  on 38  degrees of freedom
AIC: 154.09

Number of Fisher Scoring iterations: 5
```

Interpret this output. Which strategy is better, and by how much? [5 marks]

e) The company have another data set that they would like to analyse using Poisson regression. For GDPR reasons, they are unwilling to release individual responses $y_1, \ldots, y_n$.

(The same restrictions do not apply to predictor variables x in this case.) Data summaries for y are allowed, however, since these do not reveal sensitive information to the same extent. Using your answer from part c), explain what data summaries are required to perform a Poisson regression. [5 marks]

Question 2

Two teams of epidemiologists are trying to model a disease outbreak. The goals of each team, are similar, but different:

- Team A wish to predict the number of cases as accurately as possible, with a view to ensuring that enough public resources (e.g., hospital beds, protective equipment, etc.) are allocated to manage the disease as effectively as possible.
- Team B would like to identify the variables that appear to be most influential at driving the disease; their goal is to inform policy changes (e.g., closing businesses, restricting travel, administering medicines, etc.) that could lead to a reduction in the number of disease cases that are being observed.

Both teams use a similar disease model, which can take a large number of covariates as input to the model. Neither team expects that all inputs are necessary for their models, but they are unsure about which models to include.

Using the different paradigms for model selection discussed in class (prediction, understanding and regularisation) and their associated methods (AIC, cross-validation, BIC, Lasso, etc.), discuss which approaches might be most suitable for Team A and B respectively. (Suggested word limit: 100-150 words.)

[15 marks]