Ashish Kannur(kannura@tcd.ie)
ID: 19300875

Ashish Kannur
ID: 19300875

CS7DS3 - Assignment-2

Question 1:

a) Given $f(y|\theta) = \frac{\theta^y e^{-\theta}}{y!}$

Exponential family form is given as

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

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

$$= \exp\left[\log\left(\frac{\theta^y e^{-\theta}}{y!}\right)\right]$$

$$= \exp\{\log \theta^y + \log(e^{-\theta}/y!)\} \quad \ldots (\log(ab) = \log a + \log b)$$

$$= \exp\{y \log\theta + \log e^{-\theta} - \log y!\} \quad \ldots (\log a^b = b \log a)$$

$$= \exp[y \log\theta] \cdot \exp[\log(e^{-\theta}/y!)]$$

$$= e^{-\theta}/y! \cdot \exp[y \log\theta]$$

$$= \frac{1}{y!} e^{-\theta} \exp\{\log\theta\, y\}$$

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

we get, $h(y) = \frac{1}{y!}$, $g(\theta) = e^{-\theta}$, $\phi(\theta) = \log\theta$

and $s(y) = y$

Ashish Kannur(kannura@tcd.ie)
ID: 19300875

b) Generating canonical form of equation in part a) we have,

$$f(y/\theta) = \frac{1}{y!} \exp\{\eta \cdot y - A(\eta)\}$$

where $\eta = \phi(\theta) = \log\theta$

$$A(\eta) = e^{\eta} = \lambda$$

$$E(S(y)) = E(y) = \frac{d}{d\eta} A(\eta) = \frac{d}{d\eta} e^{\eta}$$

$$= e^{\log\theta}$$

$$= \theta \quad \text{---} \quad (1)$$

Defining a link function to (1)

$$\phi(\theta_i) = \eta_i$$

$$= \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$$

$$= x_i'\beta$$

$$\boxed{\theta_i = \phi^{-1}(x_i'\beta)} \quad \text{---} \quad (2)$$

As variance and mean of poisson distribution is λ, therefore for observations $y_1, y_2 \dots y_n$

$$E(y_i) = \lambda$$

$$\therefore y_i = E(y_i) + \varepsilon_i \quad i = 1, 2 \dots n$$

From eq^n (2), identity link function is

$$\theta_i = \phi^{-1}(x_i'\beta)$$

$$\therefore \theta_i = \exp(x_i'\beta) \dots \left[\text{From part (a) } \phi(\theta) = \log\theta\right]$$

Ashish Kannur(kannura@tcd.ie)
ID: 19300875

1 c) $$f(y_i|\theta_i) = \frac{\theta_i^{y_i} e^{-\theta_i}}{y_i!} \quad \text{— (1)}$$

From part (b) using the link function,
$\theta_i = e^{x_i \beta}$

Calculating joint probability from (1)
as y_i & θ_i are independent random variable

$$P(y|\theta) = \prod_{i=1}^{n} \left(P(y_i|\theta_i)\right)$$
$$= \prod_{i=1}^{n} \frac{e^{-\theta_i} \theta_i^{y_i}}{y_i!} = L(\beta) \quad \text{— (2)}$$

Taking log likelihood of eq^n (2), we get
$$\log\left(L(\beta)\right) = \sum_{i=1}^{n} y_i x_i \beta - e^{x_i \beta} - \ln y_i! \quad \text{— (3)}$$

Now differenciating eq^n (3),
$$\frac{d}{d\beta} \log\left(L(\beta)\right) = \sum_{i=1}^{n} y_i x_i - e^{x_i \beta} \cdot x_i$$
$$= \sum_{i=1}^{n} \left(y_i - e^{x_i \beta}\right) x_i$$

For MLE, equate above equation to zero

$$\therefore \boxed{\sum_{i=1}^{n} \left(y_i - e^{x_i \beta}\right) = 0}$$

Ashish Kannur(kannura@tcd.ie)
ID: 19300875

1 d) From the given output, following interpretation is derived

→ Number of fisher score iterations imply that 5 iterations were used to fit the model and weight the model parameters.

→ Residual deviance of 42.344 is less than null deviance of 120.941. This signifies that model performance improves with inclusion of 1 extra parameter which is inferred from 39 degree of freedom

→ Since Residual deviance is less than null deviance, it signifies that model is a good fit and is is appropriate

→ p-value for strategy 1 is significantly small and use of quasi poisson will also not make the predictor insignificant.

→ Strategy 1 is selected over strategy 0 due to following reason:
→ Poisson regression model by default uses log-link function which leads to exponential increase in traffic of website between 12:00 pm to 12:15 pm by $e^{1.6094}$

Ashish Kannur(kannura@tcd.ie)
ID: 19300875

1e) From the solution in part (c), it is evident that maximum log likelihood for coef β is derived by equating first derivative equal to 0, that is,

$$\sum_{i=1}^{n} (y_i x_i - e^{x_i \beta} x_i) = 0$$

In order to calculate coefficient β of the model, only matrix $(y^T x)$ is required from the data.

2 Solution:

➔ Team A is implementing predictive modelling to predict the number of cases with the current number of cases and related features, to ensure proper availability and management of resources.

Preferred Methods for this scenario are AIC or cross-validation.

AIC technique is used when there is no data available which is out of given sample data. It is an in-sample fit technique which estimates the likelihood of the model to predict future values. The actual unknown likelihood of the data is measured with the likelihood fitted on the data along with a constant. AIC comes handy when it is difficult to predict false negatives than the false positives. Thus AIC is suitable for this scenario where, if the cases predicted is less than the expected then there may be shortage of resources.

Cross-validation is used to predict how accurately a predicted model will perform actually. By this method, overfitting or selection bias problem can be avoided. This is done by verifying by the amount the MSE exceeds the predicted error value.

➔ Team B is performing a statistical exploratory data analysis to determine the main causes in the spread of the disease.

Preferred method/model: BIC.

BIC offers penalization for complex models by forming large number of clusters. This helps to eliminate unnecessary features. BIC helps us estimate the posterior probability assuming that one of its models is correct. BIC suits this scenario as tasks here is to find the actual features causing the disease.