HOMEWORK-4

EE559

1. $X_1, X_2 \cdots X_n \rightarrow$ iid Poisson sample

$\lambda \rightarrow \Gamma(2,1)$ prior distribution. $\Rightarrow P(\lambda) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda} = \lambda e^{-\lambda}$

To find MAP estimate of λ

Sol:

$$= \arg\max_{\lambda} f(\lambda | x_1, x_2 \cdots x_n)$$

$$= \arg\max_{\lambda} \underbrace{f(x|\lambda) f(\lambda)}$$

$$= \prod_{i=1}^{n} f(x_i|\lambda) f(\lambda)$$

$$= f(\lambda) \prod_{i=1}^{n} f(x_i|\lambda)$$

$$= \lambda e^{-\lambda} \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = \lambda^{1+\sum x_i} e^{-\lambda} e^{-n\lambda} \prod_{i=1}^{n} \frac{1}{x_i!}$$

taking log

$$\mathcal{L} = \arg\max_{\lambda} \left[1 + \sum x_i \log(\lambda) \underbrace{- \lambda - n\lambda}_{-(n+1)\lambda} + \log\left(\prod_{i=1}^{n} \frac{1}{x_i!} \right) \right]$$

taking derivative $\mathcal{L} = 0$

$$\frac{\sum_{i=1}^{n} x_i + 1}{\lambda} - (n+1) = 0$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i + 1}{n+1}$$

To ensure this is maximum, we take $\frac{d^2}{d\lambda^2} = -\frac{\sum_{i=1}^{n} x_i + 1}{\lambda^2}$

$\frac{\sum x_i + 1}{\lambda^2} > 0$ for $\lambda > 0$ $\therefore \frac{d^2}{d\lambda^2}$ is negative $\Rightarrow$ it is maximum.

$\therefore \hat{\lambda}_{MAP} = \frac{\sum x_i + 1}{n+1}$ maximizes posterior distribution.

2.

$$P_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} \qquad \lambda > 0 \qquad x = 0, 1 \text{ ---}$$

Likelihood:

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

} Given $x_1\ x_2 \text{ ---} x_n \Rightarrow$ iid poisson.

∴ log likelihood is

$$\ell(\lambda) = \sum_{i=1}^{n} (x_i \log\lambda - \lambda - \log x_i!)$$

TO find MLE: $\frac{d}{d\lambda}\ell(\lambda) = 0 \Rightarrow \frac{d}{d\lambda}\left[x_i \log\lambda - \lambda - \log x! \right]$ → goes to 0

$$\Rightarrow \sum_{i=1}^{n} \left(\frac{x_i}{\lambda} - 1\right) = 0 \qquad \text{—(i)}$$

$$\Rightarrow \sum_{i=1}^{n} \frac{x_i}{\lambda} - n = 0 \Rightarrow \left[\frac{1}{\lambda} \sum_{i=1}^{n} x_i\right] - n = 0$$

$$\Rightarrow \hat{\lambda}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Fisher Information

$$I(\lambda) = -E\left[\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2}\right]$$

● from (i)

$$\frac{\partial^2 \ell(\lambda)}{\partial \lambda^2} = \sum_{i=1}^{n} \left(\frac{-x_i}{\lambda^2}\right)$$

$$\therefore I(\lambda) = -E\left[\sum_{i=1}^{n} \left(\frac{-x_i}{\lambda^2}\right)\right]$$

$$I(\lambda) = \frac{1}{\lambda^2} + E\left[\sum_{i=1}^{n} x_i\right] = \frac{n\lambda}{\lambda^2} = \frac{n}{\lambda}$$

∴ Asymptotic normality:

$$\hat{\lambda}_{MLE} = \bar{X} \approx N\left(\lambda, (n/\lambda)^{-1}\right) \Rightarrow \hat{\lambda}_{MLE} \simeq N\left(\lambda, \frac{\lambda}{n}\right)$$

Comparison with Central Limit Theorem

if $X_1 \dots X_n$ iid with $E[X] = \mu$
$Var(X) = \sigma^2$

then, according to CLT,

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

for current problem,

$X_i \sim$ poisson
$E[X_i] = \lambda$ & $Var(X_i) = \lambda$

$$\therefore \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} \xrightarrow{d} N(0,1)$$

$$\Rightarrow \bar{X} - \lambda \simeq N(0, \lambda/n)$$

$$\boxed{\bar{X} \simeq N(\lambda, \lambda/n)} \quad \text{for large } n$$

3. $Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2)$

to show: MLE = least squares of β vector

MLE:

$$L(\beta_0, \beta_1 \dots \beta_p, \sigma^2) = \prod_{i=1}^{n} \frac{1}{2\pi\sigma^2} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_1 - \dots - \beta_p X_p)^2}{2\sigma^2}\right)$$

$\because \epsilon = Y - (\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)$ & $N(0, \sigma^2)$
& $P(\epsilon) = P(Y | X_1 \dots X_p)$

taking log,

$$l(\beta_0, \beta_1 \dots \beta_p, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_p X_{ip})^2 \quad \text{---(i)}$$

To find MLE, we need to maximise log-likelihood wrt $\beta_0, \beta_1, \dots$ & σ^2.

From (i) we notice, that we maximise log-likelihood by minimising

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2 \quad \text{——(ii)}$$

From least squares, we know,

$\hat{\beta}_0, \hat{\beta}_1 \cdots \hat{\beta}_p$ minimise the RSS

i.e $\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$ —(iii)

which gives us normal equation, $\hat{\beta} = (X^T X)^{-1} X^T y$

∴ From (ii) & (iii) we observe, log-likelihood function is maximized when RSS is minimized, which is same as minimising $\beta_0, \beta_1 \cdots \beta_p$ which minimize the RSS.

∴ MLEs for $\beta_0, \beta_1 \cdots \beta_p$ are same as the least square estimates. ∴ MLEs are also BLUE.

---

4. Given:

↳ $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$

$\varepsilon \sim N(0, \sigma^2)$

↳ prior distribution of $\beta_i = \sim N(0, \sigma^2/\lambda)$

MAP estimate:

likelihood function for given β,

$$P(Y|X, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2\right)$$

log likelihood:

$$l(\beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$$

note:
multivariate normal distribution for linear regression:

$$P(Y) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(Y-\mu)^T \Sigma^{-1} (Y-\mu)\right)$$

$\Sigma = \sigma^2 I \Rightarrow |\Sigma| = (\sigma^2)^n = \sigma^{2n}$

Prior distribution:

$$P(\beta) = \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi\sigma^2/\lambda}} \exp\left(-\frac{\lambda \beta_i^2}{2\sigma^2}\right)$$

log-prior:

$$= -\frac{p}{2} \log\left(\frac{2\pi\sigma^2}{\lambda}\right) - \frac{\lambda}{2\sigma^2} \sum_{i=1}^{p} \beta_i^2$$

Posterior distribution:

$$\log p(\beta | Y, X) = l(\beta) + \log p(\beta)$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 - \frac{p}{2}$$

$\therefore$

$$\beta_{MAP} = \arg\max_{\beta} \left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 - \frac{\lambda}{2\sigma^2} \sum_{i=1}^{p} \beta_i^2 \right)$$

should be min

i.e. when $\sum_{i=1}^{}$

$$\therefore \beta_{MAP} = \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2 + \lambda \sum_{i=1}^{p} \beta_i^2 \right)$$

Interpretation:

This is equivalent to the Ridge Regression problem

↳ $\lambda \sum_{i=1}^{p} \beta_i^2$ is added to ordinary least square to shrink co-efficients & prevent over-fitting.

↳ β_{MAP} : estimate that maximizes posterior distribution taking into account both liklihood & prior beliefs.

5. Given:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\epsilon \sim N(0, \sigma^2)$$

prior distribution for β_i :

$$\beta_i \sim \text{Laplace}\left(0, \frac{\sigma^2}{\lambda}\right)$$

MAP Estimate :

likelihood function :

$$P(Y \mid X, \beta) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2\right)$$

taking log

$$l(\beta) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2$$

prior distribution :

$$P(\beta_i) = \frac{\lambda}{2\sigma^2} \exp\left(-\frac{\lambda(|\beta_i|)}{\sigma^2}\right)$$

for joint,

$$\Rightarrow P(\beta) = \prod_{i=1}^{p} \frac{\lambda}{2\sigma^2} \exp\left(\frac{-\lambda|\beta_i|}{\sigma^2}\right)$$

log :

$$\log P(\beta) = \sum_{i=1}^{p} \left(\log \frac{\lambda}{2\sigma^2} - \frac{\lambda|\beta_i|}{\sigma^2}\right)$$

$$= p \log \frac{\lambda}{2\sigma^2} - \frac{\lambda}{\sigma^2} \sum_{i=1}^{p} |\beta_i|$$

Posterior distribution :

$$= \log p(\beta \mid Y, X) = l(\beta) + \log P(\beta)$$

$$= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2 + \phi$$

Maximising $= \arg\min_{\beta} \left(\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_p X_{ip})^2 + \frac{2\lambda}{\sigma^2} \sum_{i=1}^{p} |\beta_i| \right)$

$$= \beta_{MAP}$$

Interpretation:

↳ penalty term, $\frac{\alpha\lambda}{\sigma^2}\sum_{i=1}^{p}|\beta_i|$ is added to ordinary least squares to shrink some co-efficients to 0.

↳ equivalent to 'Lasso Regression Problem'

↳ β_{MAP} : point estimate that maximises posterior distribution, considering both data liklihood & prior beliefs.

6.

SVD of $X = U\Sigma V^T$

$U = n \times n$ orthogonal

let $\Sigma = n \times p$ diagonal with singular values $\sigma_1, \sigma_2, \cdots \sigma_p$

$V = p \times p$ orthogonal matrix.

Ridge regression estimator is given by

$$\hat{\beta}_i^{Ridge} = (X^TX + \lambda I)^{-1} X^T y \quad -(i)$$

Substituting $X = U\Sigma V^T$ into (i) $= \left[\underbrace{(V\Sigma^T U^T)(U\Sigma V^T)}_{I}\right]^{-1} + \lambda I \; V\Sigma^T U^T y$

$$\hat{\beta}_{Ridge} = (V\Sigma^T\Sigma V^T + \lambda I)^{-1} V\Sigma^T U^T y$$

Let $I = VV^T$

$$= (V(\Sigma^T\Sigma + \lambda I)V^T)^{-1} V\Sigma^T U^T y$$

$\therefore$ Since V is orthogonal,

$$\hat{\beta}_{Ridge} = V(\Sigma^T\Sigma + \lambda I)^{-1} \underbrace{V^T V}_{I} \Sigma^T U^T y$$

$$= V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T U^T y \quad \text{——(ii)}$$

$\therefore$ predicted values are:

$$\hat{y} = X\hat{\beta}_{Ridge} = U\Sigma \underbrace{V^T V}_{I}(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T U^T y$$

$$\Sigma^T\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2 \cdots \sigma_p^2)$$

$$\therefore (\Sigma^T\Sigma + \lambda I)^{-1} = \text{diag}\left(\frac{1}{\sigma_1^2+\lambda}, \cdots \frac{1}{\sigma_p^2+\lambda}\right)$$

$$\therefore \hat{y} = U \cdot \text{diag}\left(\frac{\sigma_1^2}{\sigma_1^2+\lambda} \cdots \frac{\sigma_p^2}{\sigma_p^2+\lambda}\right) U^T y \quad \text{——(iii)}$$

$$= \sum_{j=1}^{p} u_j \frac{\sigma_j^2}{\sigma_j^2+\lambda} u_j^T y \qquad ; u_j = \text{columns of } U$$

Where, $\frac{\sigma_j^2}{\sigma_j^2+\lambda}$ acts as a shrinkage factor for corresponding singular vector u_j. $\because$ for fixed λ, shrinkage factor decreases as σ_j decreases i.e if σ_j is smaller, components associated will have greater shrinkage, & less effect on prediction helping reduce overfitting.

(b) from part (a) with SVD of $X = U\Sigma V^T$

we have,

$$(X^TX + \lambda I)^{-1} = V(\Sigma^T\Sigma + \lambda I)^{-1} V^T$$

$$\therefore (X^TX + \lambda I)^{-1} X^TX = V(\Sigma^T\Sigma + \lambda I)^{-1} \Sigma^T\Sigma V^T \qquad \because X^TX = V\Sigma^T\underbrace{U^TU}_{I}\Sigma V^T = V\Sigma^T\Sigma V^T$$

$$\therefore \text{tr}[(X^TX + \lambda I)^{-1} X^TX] = \text{tr}[V(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T\Sigma V^T]$$

$$\text{tr}(AB) = \text{tr}(BA) \Rightarrow VV^T = I$$

$$\therefore \text{trace} = \text{tr}[(\Sigma^T\Sigma + \lambda I)^{-1}\Sigma^T\Sigma] \qquad \text{from (iii) part (a)}$$

$$= \text{diag}\left(\frac{\sigma_j^2}{\sigma_j^2+\lambda}\right)$$

$\therefore$ Trace = Summing of diag elements.

$$\Rightarrow \text{tr}[(X^TX + \lambda I)^{-1} X^TX] = \sum_{j=1}^{p} \frac{\sigma_j^2}{\sigma_j^2+\lambda}$$

Conclusion:
when $\lambda = 0$, there is no regularisation & effective degrees of freedom = p. As $\lambda \uparrow$, degrees of freedom $\downarrow$ reflecting increased shrinkage of co-efficients.