# Machine Learning: Exercise Sheet 2

Linxi Wang
Lusiné Nazaretyan
Ya Qian
Thomas Herold
Karen Nazaretyan

## Exercise 1

(a) $$p(x)=\int_0^{+\infty}\lambda\eta e^{-\lambda x-\eta y}dy=\lambda\eta e^{-\lambda x}\left(-\frac{1}{\eta}e^{-\eta y}\right)\Big|_0^{+\infty}$$
$$=\lambda e^{-\lambda x}$$

$$p(y)=\int_0^{+\infty}\lambda\eta e^{-\lambda x-\eta y}dx=\eta e^{-\eta y}$$

$\because$ $p(x)p(y)=\lambda\eta e^{-\lambda x-\eta y}=p(x,y)$

$\therefore$ $x$ and $y$ are independent.

(b) $$l(\lambda)=\ln p(D|\lambda)=\ln\prod_{i=1}^{N}p(x_i,y_i|\lambda)=\sum_{i=1}^{N}\ln p(x_i,y_i|\lambda)$$
$$=\sum_{i=1}^{N}\ln(\lambda\eta e^{-\lambda x_i-\eta y_i})=N\ln(\lambda\eta)-\sum_{i=1}^{N}(\lambda x_i+\eta y_i)$$
$$=N\ln(\lambda\eta)-N\lambda\bar{x}-N\eta\bar{y}$$

$$\frac{\partial l(\lambda)}{\partial\lambda}=N\frac{1}{\lambda\eta}\cdot\eta-N\bar{x}=0\quad\Rightarrow\quad\hat{\lambda}=\frac{1}{\bar{x}}$$

(c) when $\eta=1/\lambda$, we have $p(x,y)=\lambda\cdot\frac{1}{\lambda}e^{-\lambda x-\frac{1}{\lambda}y}=e^{-\lambda x-\frac{y}{\lambda}}$

$$l(\lambda)=\ln\prod_{i=1}^{N}e^{-\lambda x_i-\frac{y_i}{\lambda}}=\sum_{i=1}^{N}\left(-\lambda x_i-\frac{y_i}{\lambda}\right)=-\lambda N\bar{x}-\frac{N\bar{y}}{\lambda}$$

$$\frac{\partial l(\lambda)}{\partial\lambda}=-N\bar{x}+\frac{N\bar{y}}{\lambda^2}=0\quad\Rightarrow\quad\hat{\lambda}=\sqrt{\frac{\bar{y}}{\bar{x}}}$$

(d) when $y = 1-\lambda$, we have $p(x,y) = \lambda(1-\lambda)e^{-\lambda x-(1-\lambda)y}$

$$l(\lambda) = \ln \prod_{i=1}^{N} \lambda(1-\lambda) e^{-\lambda x_i-(1-\lambda)y_i} = \sum_{i=1}^{N} \ln[\lambda(1-\lambda)] - \sum_{i=1}^{N}[\lambda x_i + (1-\lambda)y_i]$$

$$= N \ln[\lambda(1-\lambda)] - N\lambda\bar{x} - N(1-\lambda)\bar{y}$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{N}{\lambda} - \frac{N}{1-\lambda} - N\bar{x} + N\bar{y} = 0$$

$$\Rightarrow (\bar{x}-\bar{y})\lambda^2 - (\bar{x}-\bar{y}+2)\lambda + 1 = 0$$

$$\hat{\lambda} = \frac{(\bar{x}-\bar{y}+2) + \sqrt{(\bar{x}-\bar{y})^2+4}}{2(\bar{x}-\bar{y})}$$

Exercise 2:

(a). $\hat{\beta} = (X^TX)^{-1}X^T\mathbf{y}$ where $\mathbf{y} = \mathbf{X}\beta + \varepsilon$

$$\therefore \hat{\beta} = (X^TX)^{-1}X^T(X\beta+\varepsilon) = \beta + (X^TX)^{-1}X^T\varepsilon$$

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

$\hat{\beta}$ also follows a Gaussian distribution, with

$$E(\hat{\beta}) = E(\beta + (X^TX)^{-1}X^T\varepsilon) = \beta + (X^TX)^{-1}X^T E(\varepsilon)$$

$$= \beta + 0 = \beta$$

$$Var(\hat{\beta}) = E[\hat{\beta} - E(\hat{\beta})][\hat{\beta} - E(\hat{\beta})]^T$$

$$= E[\hat{\beta} - \beta][\hat{\beta} - \beta]^T$$

$$= E[(X^TX)^{-1}X^T\varepsilon\varepsilon^T X(X^TX)^{-1}]$$

$$= (X^TX)^{-1}X^T E(\varepsilon\varepsilon^T) X(X^TX)^{-1}$$

$$= (X^TX)^{-1}X^T \sigma^2 X(X^TX)^{-1}$$

$$= \sigma^2 (X^TX)^{-1}X^TX(X^TX)^{-1}$$

$$= \sigma^2 (X^TX)^{-1}$$

(b) In most of the practical applications, not all the $\beta$s are important in modeling $y$, if we know the distribution instead of just the estimate itself of $\beta$, then we can delete some $\beta$s that are not significantly important under the significance level. Which is very useful in variable selection in a regression model.

(c). $\because \hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$

$\hat{y}_* = x_*^T \hat{\beta}$ also follows a Gaussian distribution.

$E(\hat{y}_*) = x_*^T \beta$

$Var(\hat{y}_*) = Var(x_*^T \hat{\beta}) = x_*^T [\sigma^2 (X^TX)^{-1}] x_*$

Therefore $\hat{y}_* \sim N(x_*^T \beta, \ \sigma^2 x_*^T (X^TX)^{-1} x_*$

(d). If we know the distribution of $\hat{y}_*$, we can compare it with the real data, if the two differs quite a lot, then this model would not be appropriate for modeling my data. Therefore knowing the distribution would be helpful to decide whether the model is justifiable to your real data.