Indian Institute of Technology Madras

MS4610 – Introduction to Data Analytics

Tutorial session - 1

Probability Review

16th October 2020

Questions

1. Suppose that in a class of 60 students, everyone places their pen in a single box, and then each pick up one pen at random. What is the expected value of X, the number of people that get back their own pen?

2. Customers arrive at a point of sales counter in a store at the rate of 10 per hour. Find:
   i. The probability that exactly 3 customers arrive at the counter in an hour.
   ii. The probability that exactly 2 customers arrive at the counter during a 30-minute period.

Questions

3. Let X be exponentially distributed with mean 1. Once we observe the experimental value x of X, we generate a Normal random variable Y with zero mean and variance x + 1. What is the joint PDF of X and Y?

4. Let X and Y be two jointly continuous random variables with joint PDF

$$f_{XY}(x,y) = \left\{ \begin{matrix} 6xy, \ \ 0 \le x \le 1, \ 0 \le y \le \sqrt{x} \\ 0, \ otherwise \end{matrix} \right\}$$

   i. Find $f_X(x)$ and $f_Y(y)$
   ii. Are X and Y independent?
   iii. Find the conditional PDF of X given Y=y
   iv. Find E[X|Y=y] for $0 \le y \le 1$

Regression – Some formulae

Studentised residuals:

Hat matrix $H = X(X^TX)^{-1}X^T$

Leverage h_{ii} is the i^{th} diagonal element of H.

Studentised residual is

$$t_i = \frac{e_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

$\hat{\sigma}$ is appropriate estimate of σ (standard deviation of the errors)
e_i is the i^{th} residual

Regression – Some formulae

Confidence intervals:

Expected value at response x*:

$$\hat{\mu} \pm \ t_{1-\alpha/2,n-2} . s\sqrt{\frac{1}{n} + \ \frac{(x^{*}-\bar{x})^{2}}{\sum(x_{i}-\bar{x})^{2}}}$$

Prediction intervals:

$$\hat{y}(x^{*}) \pm \ t_{1-\alpha/2,n-2} . s\sqrt{1 + \frac{1}{n} + \frac{(x^{*} - \bar{x})^{2}}{\sum(x_{i} - \bar{x})^{2}}}$$

Regression – Some formulae

Variance inflation factor

- VIF for each variable:

$$VIF(\beta_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}}$$

Multicollinearity is high if VIF is greater than 5 or 10.

- Estimated variance of coefficient estimate:

$$\widehat{Var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\widehat{var}(X_j)} \cdot \frac{1}{1 - R_j^2}$$

S^2 is estimate of the variance of error term
X_j – j^{th} independent variable