Sessional-1 AML5102 Applied Machine Learning

Formulas

1. Mean = Average = Expected value of a random variable X is

$$\mathbb{E}[X] = \sum_{x} x\, p(x)$$

2. PDF for univariate Gaussian distribution

$$\frac{1}{\sqrt{\sigma^2 2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

3. PDF for multivariate Gaussian distribution

$$\frac{1}{\sqrt{det(\Sigma)(2\pi)^D}} e^{\frac{-1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}$$

   where Σ is the covariance matrix, $|\Sigma|$ is the determinant of the covariance matrix, and μ is the mean vector.

4. Euclidean distance between two data points a and b is $\sqrt{\sum_{i=1}^{d}(a_i - b_i)^2}$ where a_i and b_i are the values of the individual features and d is the number of features

5. Manhattan distance between two data points a and b is $\sum_{i=1}^{d} |a_i - b_i|$ where a_i and b_i are the values of the individual features and d is the number of features

6. Both Manhattan distance and Euclidean distance are special cases of Minkowski distance for p=1 and 2 respectively. Minkowski distance formula is given by $\left(\sum_{i=1}^{d} |a_i - b_i|^p\right)^{\frac{1}{p}}$ where a_i and b_i are the values of the individual features and d is the number of features

7. Mahalanobis distance between two points in a multivariate Gaussian distribution is given by

$$\sqrt{(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)}$$

   where Σ is the covariance matrix, and μ is the mean vector.

8. A modified version of Euclidean distance, called weighted Euclidean distance is sometimes used in Nearest Centroid whose formula is as follows:

$$d_W(\mathbf{a}, \mathbf{b}) = \sqrt{\Sigma_{i=1}^{n} w_i (a_i - b_i)^2}$$

   where a and b are two data points and wi is the corresponding weight. This is also written as :

$$\sqrt{(a-b)^T W (a-b)}$$

where W is given by

$$W = \begin{bmatrix} w_1 & 0 & ... & 0 \\ 0 & w_2 & ... & 0 \\ 0 & 0 & .. & 0 \\ 0 & 0 & w_{n-1} & 0 \\ 0 & 0 & 0 & w_n \end{bmatrix}$$

9. Inertia (WCSS) $= \sum_{i=1}^{k} \sum_{x \in C_i} (x - \mu_i)^2$ where μ_i is the centroid of the cluster C_i and k is the total number of clusters

10. z transform in StandardScaler

$$z = \frac{x - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation for the feature x

11. Silhouette Score for a cluster with n data points is

$$\frac{1}{n} \sum_{i=1}^{n} s_i$$

where

$$s_i = \frac{b_i - a_i}{max(b_i, a_i)}$$

where a_i is the average distance between ith data point and other data points in the same cluster and b_i is the average distance between ith data point and data points in other clusters

12. Standard deviation of a random variable X and its realization with vector x has the following formula

$$\sigma = \sqrt{\frac{1}{n} \Sigma_{i=1}^{n} (x_i - \mu)^2} = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

13. Covariance between two random variables X and Y is given by $\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$

14. Covariance of x and y (where x and y are realizations of random variables X and Y respectively) is given by

$$Cov(x, y) = \frac{\sum_{i=1}^{n} (x - \bar{x})(y - \bar{y})}{n}$$

Correlation coefficient ρ is given by

$$\rho = Correl(x, y) = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

Using both formulas above, correlation coefficient can be written as

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

where $\bar{x}$ and $\bar{y}$ are the mean of x and y respectively. σ_x and σ_y are the standard deviation of x and y respectively.