Formula Sheet

Data Mining

1. Equal-width binning

$$\texttt{width}\ w = \frac{\max - \min}{n}$$

2. Normalization Formulas

$$x_i' = \frac{(x_i - \min)}{\max - \min}(\max\ _{new} - \min\ _{new}) + \min\ _{new}$$

$$x_i' = \frac{x_i}{10^j}$$

3. Linear Regression

- Residuel:

$$e_i = |y_i - h_\theta(y_i)|$$

- Cost function:

$$J(\theta_0, \theta_1) = \frac{1}{2m}\Sigma_{i=1}^m (h_\theta(x_i) - y_i)^2$$

- Linear Regression on One-Dimensional Data:

$$\theta_1 = \frac{\Sigma_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^m (x_i - \bar{x})^2}$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

4. Perceptron Algorithm

$$f(x) = \Sigma_{j=0}^n w_j x_j > 0, y = 1$$

$$f(x) = \Sigma_{j=0}^n w_j x_j < 0, y = -1$$

5. KNN

- Distance:

$$d(x, x') = \sqrt{\Sigma_{i=1}^n (x_i - x_i')^2}$$

6. KMean

- New Mean:

$$c'_k = \frac{1}{|C_k|}\Sigma_{x_i \in C_k} x_i$$

7. Hierarchical Clustering

- MAX, MIN, AVERAGE, CENTROID

Statistical

Basic Statistics

- Mean

$$\bar{x} = \frac{1}{n}\Sigma_{i=1}^{n} x_i$$

- Median
- Range

$$R = \max - \min$$

- Population variance

$$\sigma^2 = \frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$$

- Sample variance

$$s^2 = \frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2$$

- Population standard deviation

$$\sigma = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

- Sample standard deviation

$$s = \sqrt{\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \bar{x})^2}$$

- IQR

$$\texttt{IQR} = \texttt{Q3} - \texttt{Q1}$$

- Outliner

$$x_i < Q1 - 1.5 \cdot \texttt{IQR}$$

$$x_i > Q3 + 1.5 \cdot \texttt{IQR}$$

Inferential Statistics

- Standard Error

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Confidence Interval (95%)

$$[\bar{x} - 2\sigma_{\bar{x}}, \bar{x} + 2\sigma_{\bar{x}}]$$

$$[\bar{x} - 2\frac{\sigma}{\sqrt{n}}, \bar{x} + 2\frac{\sigma}{\sqrt{n}}]$$

- Single T-Test, Paired T-test

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

- Indpendent Samples T-Test

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{S^2_{\mathtt{Pooled}}(\frac{1}{m} + \frac{1}{n})}}$$

$$\begin{aligned} S^2_{\mathtt{Pooled}} &= \frac{\mathtt{df}_{\mathtt{x}}}{\mathtt{df}_{\mathtt{total}}} s_x^2 + \frac{\mathtt{df}_{\mathtt{y}}}{\mathtt{df}_{\mathtt{total}}} s_y^2 \\ &= \frac{(m-1)s_x^2 + (n-1)s_y^2}{m+n-2} \end{aligned}$$

- One-Way ANOVA

$$\mathtt{MST} = \frac{\mathtt{SST}}{p-1} = \frac{\Sigma_{i=1}^{p} n_i(\bar{x_i} - \bar{x})^2}{p-1}$$

$$\mathtt{MSE} = \frac{\mathtt{SSE}}{n-p} = \frac{\Sigma_{i=1}^{p}\Sigma_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2}{n-p}$$

$$\begin{gathered} = \frac{1}{n-p}((Y_{11} - \bar{Y_1})^2 + (Y_{21} - \bar{Y_1})^2 + \cdots + (Y_{n_1 1} - \bar{Y_1})^2 + \\ (Y_{12} - \bar{Y_2})^2 + (Y_{22} - \bar{Y_2})^2 + \cdots + (Y_{n_2 2} - \bar{Y_2})^2 + \\ \cdots + \\ (Y_{1p} - \bar{Y_p})^2 + (Y_{2p} - \bar{Y_p})^2 + \cdots + (Y_{n_p p} - \bar{Y_p})^2) \end{gathered}$$

$$F = \frac{\mathtt{MST}}{\mathtt{MSE}}$$

- Post-Hoc Test

$$\mathtt{Tukey's\ HSD} = \frac{\bar{Y_i} - \bar{Y_j}}{\sqrt{\frac{\mathtt{MSE}}{n}}}$$

Security and Privacy

- Prevalence by UCT

$$\texttt{Prevalence} = \texttt{Average of GroupA} - \texttt{Average of GroupB}$$

- Prevalence by NST

$$\texttt{Prevalence} = \frac{\Sigma \texttt{s}_\texttt{i}}{\Sigma \texttt{k}_\texttt{i}}$$

  , where s_i is the number of people he/she knows are engagingin sensitive activity and k_i is total number of people he/she knows.

- Prevalence by NRRT

$$\texttt{Prevalence}\ s = (P - ct)/(1 - c)$$

  , where P is the proportion of people answer "Yes", c is probability of answer "Yes" in first question (Coffee), t is the number of people who answer "Yes" in alternative non-sensitive question (Taxi).

- Prevalence by RRT

$$\texttt{Prevalence}\ s = (1 - \theta - P)/(1 - 2\theta)$$

  , where P is the proportion of people answer "Yes", θ is the ratio of the positive question.