

Formula Sheet

1. **Sup(X)** = fraction of transactions that contain the itemset X
2. **Sup(X → Y)** = fraction of transactions that contain both X and Y
3. **Confidence(X → Y)** = Sup(X → Y) / Sup(X)

4. **Lift Measure:**

$$\begin{aligned}
 \text{lift}(X \rightarrow Y) &= \frac{\text{Confidence of a rule}}{\text{Expected confidence of a rule}} \\
 &= \frac{\text{conf}(X \rightarrow Y)}{\frac{\text{sup}(Y)}{\text{sup}(X)}} \\
 &= \frac{\text{sup}(X \rightarrow Y)}{\text{sup}(X) * \text{sup}(Y)}
 \end{aligned}$$

5. **Conviction Measure:**

$$\begin{aligned}
 \text{Conviction}(A \rightarrow B) &= \frac{P(A)P(\bar{B})}{P(AB)} \\
 &= \frac{P(A) * (1 - P(B))}{P(A) - P(AB)} \\
 &= \frac{1 - \text{sup}(B)}{1 - \text{con } f(A \rightarrow B)}
 \end{aligned}$$

6. **Manhattan (or city block) distance:**

$$L_1(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_d - y_d| = \sum_{i=1}^d |x_i - y_i|$$

7. **Euclidean distance:** $d(x, y) = \sqrt{(|x_1 - y_1|^2 + |x_2 - y_2|^2 + \dots + |x_d - y_d|^2)}$

8. **Chebyshev distance:** $d(x, y) = \max\{|x_1 - y_1|, |x_2 - y_2| \dots |x_d - y_d|\}$

9. **Weighted distance:**

$$\begin{aligned}
 d(x, y) &= \sqrt{(w_1|x_1 - y_1|^2 + w_2|x_2 - y_2|^2 + \dots + w_d|x_d - y_d|^2)} \\
 d(x, y) &= w_1|x_1 - y_1| + w_2|x_2 - y_2| + \dots + w_d|x_d - y_d|
 \end{aligned}$$

10. **Jaccard similarity:** $\text{JSim}(X, Y) = \frac{X \cap Y}{X \cup Y}$

11. **Jaccard distance:** $Jdist(X, Y) = 1 - JSim(X, Y)$

12. **Weighted formula (variable to mixed type):** $d(i, j) = \frac{\sum_{f=1}^p \delta_{ij}^{(f)} d_{ij}^{(f)}}{\sum_{f=1}^p \delta_{ij}^{(f)}}$

f – feature or variable; $\delta_{ij}^{(f)}$ - weight for feature f; i, j – observations i, j

13. **K-means (Cost Function):**
$$L(\Delta) = \sum_{i=1}^n \|x_i - \mu_{k(i)}\|^2 = \sum_{k=1}^K \sum_{i \in C_k} \|x_i - \mu_k\|^2$$
$$= \sum_{k=1}^K \sum_{i, j \in C_k} \|x_i - x_j\|^2$$

14. **Expected information,** $I(s_1, s_2 \dots s_m) = - \sum_{i=1}^m \frac{|s_i|}{|S|} \log_2 \frac{|s_i|}{|S|} = - \sum_{i=1}^m p_i \log p_i$

S is the training dataset consisting of s data samples

S contains s_i samples of class C_i for $i = \{1, 2, \dots, m\}$, m is the total number of classes

15. **Entropy:** $E(A) = \sum_{j=1}^v \frac{|s_{1j}| + |s_{2j}| + \dots + |s_{mj}|}{|S|} I(s_{1j} \dots s_{mj})$

16. **Information Gain:** $Gain(A) = I(s_1, s_2 \dots s_m) - E(A)$

17. **Intrinsic Info:** $IntrinsicInfo(A) = - \sum_{j=1}^v p_j \log p_j$

v – number of values of attribute A

18. **Gain Ratio:** $GainRatio(A) = \frac{Gain(A)}{IntrinsicInfo(A)}$

19. **Gini(S):** $\sum_i p_i (1 - p_i)$

20. **Gain(S, A):** $\sum_i \frac{|s_i|}{|S|} Gini(s_i)$

21. **Cost complexity of tree, $CC(T) = \text{error}(T) + \alpha * L(T)$**

$\text{error}(T)$ is the fraction of training records that are misclassified by tree T

$L(T)$ is the number of terminal nodes

α is a penalty factor for tree size

22. **Accuracy:** $(TP + TN) / (TP + TN + FP + FN)$

TP – True positives; FP – False positives

TN – True negatives; FN – False negatives

23. **Precision:** $TP / (TP + FP)$

24. **Recall:** $TP / (TP + FN)$