Data Mining Mathematics Cheat Sheet

1. Descriptive Statistics

Central Tendency

Mean (Average)

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Median

- Middle value when ordered
- If n is even, average of two middle values

Mode

- Most frequent value(s)

Spread

Variance

$$\sigma^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

Standard Deviation

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

2. Data Normalization

Min-Max Scaling

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Scales data to [0,1] range

Z-Score Standardization

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

Transforms to mean=0, std=1

3. Classification Metrics

Basic Metrics

- **TP** (True Positives): Correct positive predictions
- **TN** (True Negatives): Correct negative predictions
- **FP** (False Positives): Incorrect positive predictions
- **FN** (False Negatives): Incorrect negative predictions

Evaluation Formulas

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity)

$$\text{Recall} = \frac{TP}{TP + FN}$$

Specificity

$$\text{Specificity} = \frac{TN}{TN + FP}$$

F1 Score

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

4. Information Theory

Entropy

$$\text{Entropy} = -\sum_{i=1}^{n} p_i \log_2(p_i)$$

Weighted Entropy

$$\text{Weighted Entropy} = \sum_{j=1}^{m} \frac{n_j}{n} \cdot \text{Entropy}(j)$$

where n_j is size of subset j

Information Gain

$$\text{IG} = \text{Entropy(parent)} - \text{Weighted Entropy(children)}$$

5. Distance Measures

Euclidean Distance

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

6. Averages

Weighted Average

$$\text{Weighted Avg} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

where w_i are weights

7. Clustering

Sum of Squared Error (SSE)

$$\text{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2$$

where μ_i is centroid of cluster C_i

8. Association Rules

Support

$$\text{Support}(A) = \frac{\text{count}(A)}{\text{total transactions}}$$

Confidence

$$\text{Confidence}(A \to B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}$$

Lift

$$\text{Lift}(A \to B) = \frac{\text{Confidence}(A \to B)}{\text{Support}(B)}$$

Key Interpretations

AUC-ROC Values

- `0.5`: Random classifier
- `> 0.7`: Good classifier
- `>0.8`: Strong classifier
- `1.0`: Perfect classifier

Lift Values

- `>1`: Positive association
- `=1`: Independent
- `<1`: Negative association

Information Gain

- Higher value = Better split
- Used for decision tree feature selection

Distance Measures

- Euclidean: Straight-line distance
- Manhattan: Grid-based distance
- Used in KNN and clustering