Data Mining Algorithms and Formulas

1. Linear Regression

Goal

Predict a continuous variable (Y) from input variables (X).

Regression Equation

$$ Y = b_0 + b_1 X + \epsilon $$

☑ Slope and Intercept (Standard Formula)

$$ b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} $$ $$ b_0 = \bar{Y} - b_1 \bar{X} $$

☑ Slope using Correlation

$$ b_1 = r \cdot \frac{s_Y}{s_X} $$

Where:

- (r) = Pearson correlation
- (s_X, s_Y) = standard deviations of X and Y

☑ Intercept from Slope

$$ b_0 = \bar{Y} - b_1 \bar{X} $$

Error Metrics

- **Mean Squared Error (MSE)** $$ MSE = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 $$
- **R^2 (Coefficient of Determination)** $$ R^2 = 1 - \frac{\sum (Y_i - \hat{Y}_i)^2}{\sum (Y_i - \bar{Y})^2} $$

---

2. Naive Bayes Classifier

Bayes' Theorem

$$ P(C|X) = \frac{P(X|C) , P(C)}{P(X)} $$

Categorical Naive Bayes

$$ P(X|C) = \prod_{i=1}^{n} P(x_i|C) $$

Gaussian Naive Bayes (Numerical Features)

$$ P(x_i|C) = \frac{1}{\sqrt{2\pi\sigma_{C,i}^2}} \exp\left(-\frac{(x_i - \mu_{C,i})^2}{2\sigma_{C,i}^2}\right) $$

---

3. Pearson Correlation

Measures strength of linear relationship between two variables.

☑ Standard Formula

$$ r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})} {\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} $$

☑ Alternate Formula (Summation Form)

$$ r = \frac{n(\sum xy) - (\sum x)(\sum y)} {\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}} $$

Correlation Interpretation

Value of r	Meaning
1	Perfect positive correlation
-1	Perfect negative correlation
0	No correlation

---

4. K-Means Clustering

Objective Function

$$ J = \sum_{i=1}^{K} \sum_{x_j \in C_i} ||x_j - \mu_i||^2 $$

---

5. Hierarchical Clustering

Linkage Methods

- **Single Linkage** $$ D(A,B) = \min_{i \in A, j \in B} d(i, j) $$
- **Complete Linkage** $$ D(A,B) = \max_{i \in A, j \in B} d(i, j) $$
- **Average Linkage** $$ D(A,B) = \frac{1}{|A||B|} \sum_{i \in A, j \in B} d(i, j) $$

---

6. DBSCAN

Parameters

- `eps` — radius of neighborhood
- `minPts` — density threshold

Type	Description

Type	Description
Core	≥ `minPts` points within `eps`
Border	Reachable but not dense
Noise	Not reachable

---

7. Decision Tree (Entropy & Information Gain)

Entropy

$$ Entropy(S) = -\sum_{i=1}^{k} p_i \log_2(p_i) $$

Information Gain

$$ Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) $$

---

8. Data Preprocessing Techniques

Min–Max Normalization

$$ X' = \frac{X - X_{min}}{X_{max} - X_{min}} $$

Z-Score Normalization

$$ X' = \frac{X - \bar{X}}{\sigma} $$

Binning

- Replace values in a bin with **mean** or **median**

---

9. Association Rule Mining (Apriori)

Support

$$ Support(A \Rightarrow B) = \frac{\text{Transactions containing } (A \cup B)} {\text{Total Transactions}} $$

Confidence

$$ Confidence(A \Rightarrow B) = \frac{\text{Transactions containing } (A \cup B)} {\text{Transactions containing } A} $$

Lift

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

---

10. Summary Table

Algorithm	Type	Works On	Output
Linear Regression	Predictive	Numerical	Continuous Y
Naive Bayes	Predictive	Mixed	Class Label
K-Means	Descriptive	Numerical	Clusters
DBSCAN	Descriptive	Numerical	Clusters + Noise
Decision Tree	Predictive	Mixed	Class Label
Apriori	Descriptive	Categorical	Rules
Correlation	Descriptive	Numerical	Relation Strength