Lecture : 07

Measure of proximity and similarity:

proximity refers to how close one item is to another
similarity describes how alike two items are
These words are use interchangeably--

➡ Importance of Similarity in Data science
- Feature Relationship
- Improved Data understanding

→ Measure of similarity relevant to numeric attributes
→ For similarity calculation, convert attributes into vectors

proximity / similarity Measures:

① **Dot product**: (how align and how strong)
- It multiplies corresponding entries and sum the results
- The output is single scaler value
- If the dot product is zero, vectors are orthogonal meaning no similarity
- A higher dot product implies greater similarity

② Pros:
Easy to calculate

cons:
changing the unit of measurement alter the dot product value

② **cosine similarity**: (only direction)
- Focuses solely on the direction of vectors
- Ressolves the magnitude issue of dot product
- Always in the range of -1 to +1
- +1 implies perfect allignment
- -1 implies opposition

$$sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

③ Covalience

- It measures how two variables vary together.
- It consider the magnitude by subtracting the mean from each observation
- A positive covariance means both variables tends to move together
- A negetive covariance means that they move in opposite directions.

$$cov(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

→ $-\infty$ to $+\infty$

Cons:

It doesnot provide strength of vector

④ Correlation

- Quantifies both strength & direction
- ± 1 indicates perfect relationship
- values bounded b/w -1 and $+1$
- Derived by dividing covariance by standard deviation

$$corr(X, Y) = \frac{cov(X, Y)}{\sigma X \sigma y}$$

Lecture: 08

Data Preprocessing

Importance of Data Quality

- Good models
- Good data → Good models
- Direct impact on model performance

What is data Quality?

→ Accuracy of data
→ Consistency in attributes
→ No unnecessary duplicates
→ Minimal missing values
→ Timliness (up to date info for current needs)

Real world Data challenges

→ Missing values
→ Duplicate entries
→ Inconsistent formatting
→ Irrelevant features

GARBAGE IN, GARBAGE OUT

Data Preprocessing

1) clean (remove errors and incnsistencies)
↓
2) Integrate (combine multiple sources)
↓
3) Reduce
↓
4) Transform

→ Data cleaning:

Remove duplicates, Handling Missing values, Fixing incorrect entries, Removing Irrelevant features et

→ Data Integration

Merging data from different origins into a single dataset

→ Data Reduction

Feature Aggregation, Feature selection, Principal component Analysis, strategic sampling etc

→ Data Transformation

Min Max scaling (0-1), zscore normalization (mean=0, std=

Logarithmeic scaling etc