DATA PREPROCESSING

χ^2 analysis

- $\chi^2 = \sum \frac{(O-E)^2}{E}$ O - Observed Value
  E - Expected Value.

Calculation of E → For ij in contingency table,
compute $\frac{\text{rowsum}(i) \times \text{colsum}(j)}{\text{Total Elements}}$

- It is a test of independence for two attributes (categorical).
- It is a right tailed test
- H_0: The two attributes are independent / not correlated.
- dof: $(r-1)\times(c-1)$ where (r, c) are rows & cols in contingency table
  i.e. They refer to the levels in a category.
- $\chi^2 \leq \chi^2_{critical}$ ⇒ Fail to reject the Null Hypothesis (Independent)
- $\chi^2 > \chi^2_{critical}$ ⇒ Reject the null hypothesis. (Not Independent)

PEARSON'S CORRELATION CO-EFFICIENT

$$r_{xy} = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{n\,\sigma_A\,\sigma_B} = \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}} = \frac{\sum x_i y_i - n\,\bar{x}\,\bar{y}}{n\sqrt{Var(X)\cdot Var(Y)}}$$

$r_{xy} > 0$ ⇒ Positive correl
< 0 ⇒ Negative correl
$r_{xy} \in [-1, 1]$
Sign of r → Direction of relation
Magnitude of r → Strength (linear) of relation)
$r = 0$ ⇒ No linear relation between x & y.

BINNING
- → EQUAL DEPTH: Arrange data in asc. order; det. no. of bins
  Dataset Size / Bin number → Bin Capacity
  Create Bins.
- → EQUAL WIDTH: Arrange data in asc. order; det min & max
  Bin Width = (max - min) / no. of bins
  Formulate bins $(x_{i-1} + w, x_{i-1} + 2w]$
  where x_{i-1} → Prev. bin close & w - bin width.

SMOOTHENING
- By Mean → Replace all elems in a bin by mean of those elems
- By Median → " " median " "
- By Boundaries → Replace elems closer to min of bin by min & closer to max of bin by max.

NORMALIZATION
- Min Max Scaling → $x' = \frac{x - x_{min}}{x_{max} - x_{min}}$
- Decimal Scaling → $x' = \frac{x}{10^j}$ → (j is the order of mag of largest no. in dataset)

DATA EXPLORATION

Attribute types

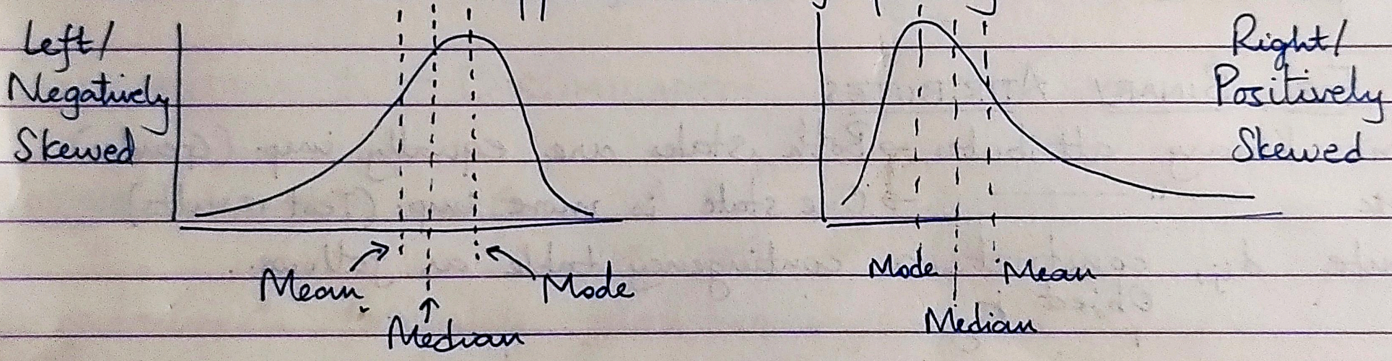
Qualitative → Nominal
- → Binary
- → Ordinal (Order matters)

Quantitative → Interval Scaled (Only δ matters)
- → Ratio Scaled (Ht. & Wt., contains abs. zero)

Discrete Attribute → Has finite or countably infinite set of values.
Continuous Attribute → Has Real numbers as attribute values.

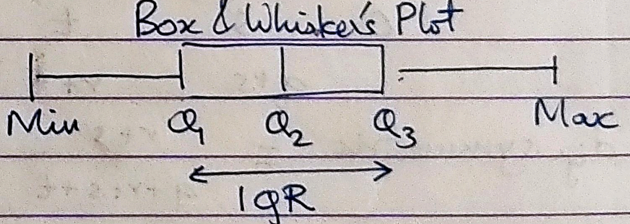
Mean → $\frac{\sum x_i}{N}$; Trimmed Mean $\frac{\sum x_i}{N}$ $\forall x_i \notin \{\text{outliers set}\}$

Median → Arrange data in ascending order; Middle element → Median.
Mode → Item which appears most frequently in the dataset.



FIVE NUMBER SUMMARY

Min Q_1 Q_2 (Median) Q_3 Max



Q_1 → Element @ $\frac{(n+1)}{4}$ position

Q_2 → Element @ $\frac{n+1}{2}^{th}$ position $\quad Q_3$ → Element @ $\frac{3(n+1)}{4}^{th}$ position.

VARIANCE

$$\text{Variance } (\sigma^2) = \frac{1}{n}\sum_{1}^{n}(x-\bar{x})^2$$

$$\text{Variance } (s^2) = \frac{1}{n-1}\sum_{1}^{n}(x-\bar{x})^2$$

} Population v/s Sample

Population stdev $\sigma = \sqrt{\sigma^2}$

Sample stdev $S = \sqrt{s^2}$ → Sqrt. of Sample Variance

Similarity / Dissimilarity (Proximity Measures)

Data matrix - Input data (with all attributes but the target)

Dissimilarity Matrix → A Δ^r matrix with d_{ij} = dissim. between record i & record j

→ Technically it's a symmetric matrix with $d_{ij} = 0 \ \forall \ i = j$

Proximity For Nominal Attributes

$$d_{ij} = \frac{p - m}{p}$$

p - No. of nominal attributes

m - No. of matches

$d_{ij} \in [0, 1]$

★ If you're asked to weigh all attributes equally in a mixed attribute dataset, don't use this. Only compare match/no match.

Proximity For Binary Attributes

Symmetric binary attributes → Both states are equally imp. (Gender)

Asymmetric —"— → One state is more imp. (Test results)

To compute d_{ij}, construct a contingency table as follows.

Object i \ Object j	1	0	
1	q	r	$q+r$
0	s	t	$s+t$
	$q+s$	$r+t$	

$$d_{ij\ \text{symmetric}} = \frac{r+s}{q+r+s+t} = \frac{\text{Items which don't match}}{\text{Total Items.}}$$

$$d_{ij\ \text{asymm.}} = \frac{r+s}{q+r+s}$$

↳ Don't care about (0,0) match for similarity.

Proximity For Numeric Data

$$d_{ij} = \sqrt[n]{\sum (x_i - x_j)^n}$$

If $n = 2 \Rightarrow$ Euclidean

$n = \mathbb{R} - \{0, 1, 2\} \Rightarrow$ Minkowski

$n = \infty \Rightarrow$ Supremum → Max row sum

$n = 1 \Rightarrow$ L1-norm distance (Manhattan)

p^1	0	2
p^2	2	0
p^3	3	1
p^4	5	1

$(p^1, p^2)_1 = |2| + |2| = 2$

$(p^1, p^3)_2 = \sqrt{3^2 + 1^2} = \sqrt{10}$

$(p^1, p^4)_\infty = \max(5, 1) = 5$

PROXIMITY MEASURE OF ORDINAL DATA

→ Replace each ordinal by rank

→ Normalize the ranking $r = \frac{r_{if} - 1}{M_f - 1}$

r_{if} → Rating of record i for ordinal feature f

M_f → Number of levels in feature f.

→ Compute the distance using Euclidean dist. formula

MIXED DATATYPE

Find dissimilarity based on individual attribute type as we discussed so far & then weight them

$$\delta_{ij} = \frac{\sum w_f \cdot \delta_{ijf}}{\sum w_f}$$

w_f → Feature weight

δ_{ij} → Distance values.

SIMILARITY = 1 − DISSIMILARITY

COSINE SIMILARITY

$$\rightarrow \cos\theta = \frac{A \cdot B}{||A||\,||B||} = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2 \cdot \sum b_i^2}}$$

Given documents formulate term frequency matrix

	d_1	d_2	d_3	d_4	...
t_1					
t_2					
t_3					
t_4					
⋮					

Each column is now a vector for document

Each row is now a vector for terms

CLASSIFICATION

DECISION TREES

Algorithms: ID3, CART, C4.5 etc.

Principle → Split records based on an attribute that optimizes a criterion

Splits –
- → Binary Split → Straightforward for binary/categorical
  → 2 bins for continuous
- → Multi-way Split → As many splits as levels in categorical
  → Discretize bins for continuous attributes

Measures of Node Impurity

Gini Index → $GI = 1 - \sum_{1}^{r} p_j^2$ where $p_j = \frac{n_j}{n_{dataset\ (node)}}$ & r is number of target classes.

$$= 1 - \sum_{j=1}^{r} \left(\frac{j}{t}\right)^2$$

$$\text{Gini Index}_{(\text{Split})}^{(\text{node})} = \sum w_i \cdot \text{Gini}_i$$

Here w_i → Proportion of i^{th} level in the node split
Gini_i → Corresponding Gini Index of that partition.

$$\Delta GI = \text{Gini (Node)} - \text{Gini}_A\text{(Node Split)}$$

For every attribute A, compute ΔGI and select the one with highest value of $\Delta GI \Rightarrow$ Lowest value of Gini_A(Node Split)

Entropy (ϵ) $\epsilon = -\sum p_j \log p_j$ (Everything else is same as above)

IG i.e. Information Gain = Entropy (Node) – Entropy (Node Split)
Select attribute with high IG for splitting at any given pt.

Decision boundaries are always // to the axes
If Split involves multiple attributes then boundaries can become oblique
⇒ Oblique Decision Trees.

$$\text{Gain Ratio}_A = \frac{\text{Information Gain (A)}}{\text{Entropy (Node Split A)}}$$

→ Attribute with maximum gain ratio is selected as splitting attribute.

Rule-Based Classifier

Rule (Condition) → y
Antecedent Consequent

A rule r covers an instance x if attributes of x satisfy cond. in r

Coverage → Fraction of records which satisfy the rule antecedent
Accuracy → " both antecedent & consequent

Mutually Exclusive Rules

→ If rules are independent of each other
→ Every record is covered by at most 1 rule.

Exhaustive Rules

→ If it accounts for all possible combinations of attribute values.
→ Each record is covered by at least 1 Rule.
→ May lead to clash of ≥ 1 Rule
  → Ordered Set
  → Majority Voting

Rule Quality measures

FOIL Gain → $Gain(R_0, R_1) = t\left\{\log \frac{p_1}{p_1+n_1} - \log \frac{p_0}{p_0+n_0}\right\}$

REGRESSION

$Y = a + bX$

- Y: Dep (Predicted)
- a: Y-intercept
- b: Slope
- X: Indep (Predictor)

LS method (Least Squares Sum)

$$b = \frac{\Sigma xy - n\bar{x}\bar{y}}{\Sigma x^2 - n \cdot \bar{x}^2}$$

$$a = \bar{y} - b\bar{x}$$

$$\text{Std Error } S_{xy} = \sqrt{\frac{\Sigma (y_i - y')^2}{n-2}}$$

EVALUATION METRICS

Actual ↓ / Predicted →	C1	C0
C1	TP	FP
C0	FN	TN

$$\text{Recall} = TP/(TP+FN)$$

$$\text{Sensitivity} = \text{Recall}$$

$$\text{Precision} = TP/(TP+FP)$$

$$\text{Specificity} = TN/(TN+FP)$$

$$\text{F-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{recall}}{\beta^2 \times (\text{Precision}) + \text{recall}}$$