Internal Institute of Professional Studies, Devi Ahilya University Indore

August 2025

First Internal Test- Introduction to Data Science

Instructor: Dr. Shaligram Prajapat

Part A: Statistical Characterisation of Feedback Data

1. *Give three more commonly used statistical measures for characterising dispersion in student feedback data (e.g., variation in teaching quality scores, lab facility ratings, or peer collaboration effectiveness). Discuss how these measures can be computed efficiently in large student databases.*
2. *Suppose that the student feedback (rating on a scale of 1–100) on faculty mentoring effectiveness is given in sorted order:* 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.
   (a) *What is the mean rating? What is the median?*
   (b) What is the mode of the data? Comment on its modality (bimodal, trimodal, etc.)
   (c) What is the midrange of the data?
   (d) Estimate the first quartile (Q1) and the third quartile (Q3).
   (e) Give the five-number summary of the ratings.
   (f) Draw a boxplot of the data.
   (g) How is a quantile–quantile plot different from a quantile plot in this context?
3. Suppose student lab–infrastructure satisfaction scores are grouped into intervals with the following distribution:

Score Range	1-5	6-15	16-20	21-50	51-80	81-110
Frequency (students)	200	450	300	1500	700	44

   Compute an approximate median satisfaction score for the students.
4. In the department of Lifelong Learning , A feedback survey collected both student age and their self-assessment of learning productivity (%) for 18 students:

Age	23	23	27	27	39	41	47	49	50	52	54	54	56	57	58	58	60	61
%	9.5	26.5	7.8	17.8	31.4	25.9	27.4	27.2	31.2	34.6	42.5	28.8	33.4	30.2	34.1	32.9	41.2	35.7

   A. Calculate the mean, median, and standard deviation of age and % productivity
   B. Draw the boxplots for both.
   C. Draw a scatter plot and a Q–Q plot to study the relationship.

Part B: Similarity and Dissimilarity in Feedback

5. Briefly outline how to compute dissimilarity between students' feedback profiles described by:
   (a) Nominal attributes (department, elective chosen) (b) Asymmetric binary attributes (attended workshops: yes/no)
   (c) Numeric attributes (e.g., CGPA, attendance %) (d) Term–frequency vectors (keywords in open-text feedback).
6. Suppose two student–faculty interaction profiles are represented as tuples:(22, 1, 42, 10) and (20, 0, 36, 8), where attributes represent [hours spent in mentoring, project reviews attended, assignments submitted, infrastructure usage count].Compute
   (a) Euclidean distance. (b) Manhattan distance
   (c) Minkowski distance with h = 3 (d) Supremum distance
7. The median score is often a key holistic indicator of faculty or infrastructure performance. When dealing with massive student feedback datasets, propose methods for median approximation. Analyse their complexity and suggest heuristics to balance accuracy vs. efficiency.
8. Consider student–faculty similarity analysis: A sample dataset of two attributes (Faculty Responsiveness, Infrastructure Support):

Student	S1	S2	S3	S4	S5
A1	1.5	2	1.6	1.2	1.5
A2	1.7	1.9	1.8	1.5	1.0

Given a new feedback query x = (1.4, 1.6).
(a) Rank existing students' feedback profiles by similarity using Euclidean, Manhattan, Supremum, and Cosine similarity
(b) Normalise each student's vector to unit length and recompute using Euclidean distance

Part C: Data Preprocessing for Feedback Analytics

9. Student–faculty feedback data may suffer from issues like accuracy, completeness, and consistency. Discuss each issue in this context (e.g., missing survey responses, contradictory ratings). Propose two other dimensions of data quality relevant to learning analytics.
10. Handling missing student feedback values: Discuss approaches such as ignoring tuples, filling manually, filling with constants, mean/median/mode imputation, class-specific imputation, or most-probable value methods.
11. Consider the feedback ratings (same as in Q2).
    (a) Smooth the ratings using bin means with bin depth 3. Illustrate steps.
    (b) Suggest ways to detect outliers (e.g., abnormally high or low satisfaction).
    (c) Discuss other methods for data smoothing.
12. In data integration for MCA/MTech analytics, discuss schema integration, handling redundant data (e.g., duplicate entries from multiple surveys), and resolving conflicts in values.
13. Discuss normalisation methods:
    (a) min–max normalisation
    (b) z-score normalization
    (c) z-score normalisation using mean absolute deviation
    (d) decimal scaling
14. Normalise the following faculty workload values (hours per semester): 200, 300, 400, 600, 1000 using (a)–(d).
15. Using the feedback ratings dataset, normalize the value 35 using:
    (a) min–max normalization (0–1 range)
    (b) z-score normalization ($\sigma = 12.94$)
    (c) decimal scaling
    (d) Discuss which method is most appropriate and why.
16. Using age and productivity data (from Q4):
    (a) Normalize both attributes using z-score normalization.
    (b) Compute Pearson's correlation coefficient and covariance. Comment on whether the relationship is positive/negative.
17. Suppose student project–evaluation scores are:
    5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.
    Partition into bins using:
    (a) equal–frequency partitioning (b) equal–width partitioning (c) clustering
18. Use a flowchart to summarize attribute selection techniques for feedback analysis:
    (a) forward selection
    (b) backward elimination
    (c) hybrid method

Part D: Compatibility & Pairing (Collaboration Analytics)

19. A student collaboration dataset contains:

Student	Gender	Skill-1	Skill-2	Skill-3	Skill-4
A	M	N	P	P	N
B	F	N	P	P	N
C	M	P	N	N	P

Here, skills are asymmetric (P = 1, N = 0), gender is symmetric.
(a) Construct contingency matrices for A–B, A–C, B–C. (b) Compute invariant dissimilarity between pairs.
(c) Compute non-invariant dissimilarity. (d) Suggest the best and worst student–collaboration pairs.
(e) Include gender in analysis and re-evaluate compatibility.