Week 2 Summary: Correlation, Distance Measures, and Data Quality

Z3003 Data Mining: Scaling to Large Data · 23–25 September 2026 · Dr. Innocent Nyalala

Lecture 4: Covariance, Correlation, and Graphical Displays

Covariance $\mathrm{Cov}(x,y) = \frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{n}\sum x_i y_i - \bar{x}\bar{y}$ depends on unit scale, whereas Pearson's correlation $r = \mathrm{Cov}(x,y)/(\sigma_x\sigma_y) \in [-1,1]$ is unit-free and computable in one distributed pass from six running sums $(n, \sum x, \sum y, \sum xy, \sum x^2, \sum y^2)$. Because r detects only linear trends and can reflect confounding (crop type drives both price and yield to $r = -0.42$ on the 500 farms), plot scatter, Q-Q, and boxplots before modeling.

Lecture 5: Similarity and Distance Measures

The Minkowski distance family $d_h(x,y) = \left(\sum_f |x_f - y_f|^h\right)^{1/h}$ specializes to Manhattan ($h=1$), Euclidean ($h=2$), and Chebyshev ($h\to\infty$, $\max_f |x_f - y_f|$), always ordering Manhattan $\geq$ Euclidean $\geq$ Chebyshev.

- **Binary:** Use Simple Matching Coefficient $(q+t)/(q+r+s+t)$ for symmetric attributes and Jaccard $q/(q+r+s)$ for asymmetric baskets or symptoms.
- **Text & Ratings:** Cosine similarity $(x\cdot y)/(\|x\|\|y\|)$ compares direction rather than document length.
- **Mixed Types:** Gower's distance averages per-attribute normalized dissimilarities in $[0,1]$.

Lecture 6: Data Quality, Cleaning, and Redundancy Testing

Evaluate accuracy, completeness, consistency, timeliness, believability, and interpretability. Impute missing numeric values using the within-group median (such as median yield per crop) or deterministic domain equations (revenue = yield × 1000 × price), never blind zeros. Apply Tukey IQR fences inside each crop group rather than globally.

Worked Check: Pearson's r, Distances, and χ^2 Redundancy

Points differing by $(3,4)$ give Euclidean 5, Manhattan 7, Chebyshev 4. On 200 farms cross-tabulating fertilizer and irrigation with observed counts $(60,40,40,60)$ against expected 50 per cell, $\chi^2 = 4\times(10^2/50) = 8.0 > 3.84$ ($\mathrm{df}=1, \alpha=0.05$), rejecting independence.