CSC 4792 Week 3 Summary & Core Concepts: Tools, Technologies & Dimensionality Reduction

---

1. Fundamental Tools & Environments - **Jupyter Notebook:** Interactive environment combining Markdown, executable code, equations, and visualizations. - **Google Colaboratory (Colab):** Cloud-based Jupyter environment with free GPU/TPU access, pre-installed libraries, and easy sharing.

2. Datasets Used - Predicting Student Learning Outcomes dataset - UNZA Institutional Repository Research Output dataset

3. Core Python Libraries - **Pandas:** Data manipulation with DataFrames and Series. - **Matplotlib:** Visualizations (bar, line, histogram, scatter). - **Scikit-learn:** Machine learning and preprocessing tools, including PCA and LDA.

4. Data Exploration & Visualization - Pandas for loading, inspecting (`.head()`, `.info()`, `.describe()`), filtering, and transforming data. - Matplotlib for visual summaries of datasets.

Dimensionality Reduction Fundamentals

What is Dimensionality Reduction? The process of reducing the number of features in a dataset while retaining important information. It helps to: - Solve the *curse of dimensionality* - Improve model efficiency and generalization - Reduce storage needs - Enable easier visualization

Feature Selection: - Selects a subset of original features without changing them. - **Techniques:** - *Filter Methods:* Statistical tests (correlation, Chi-square) - *Wrapper Methods:* Model-based evaluation (Recursive Feature Elimination) - *Embedded Methods:* Selection during training (Lasso regression)

Feature Extraction: - Transforms original features into new features (components). - **PCA:** Unsupervised method finding components that explain most variance. - **LDA:** Supervised method maximizing class separability.

PCA Process: 1. Standardize data 2. Compute covariance matrix 3. Calculate eigenvectors & eigenvalues 4. Select top-k components

LDA Difference: - Considers class labels when finding new axes, unlike PCA.

---

5. Key Code Concepts - Pandas for filtering and transformations. - Matplotlib for visualizations. - Scikit-learn for implementing PCA and LDA.

Takeaway: Week 3 strengthens both practical tool usage (Jupyter, Colab, Pandas, Matplotlib) and theoretical understanding of dimensionality reduction, equipping you to process high-dimensional data efficiently and effectively.