

---

# Machine Learning (Problem set 5)

**Ali Alipour**, University of Tehran

---

**T**his report discusses methods for selecting optimal features and performing dimensionality reduction using Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). The report also includes model performance evaluation through cross-validation with different numbers of components.

## 1 Feature Selection and Dimensionality Reduction

### 1.1 Principal Component Analysis (PCA)

PCA helps identify the optimal number of components by balancing the retention of meaningful information and reducing dimensionality. The method finds an "elbow point" where the eigenvalues decrease significantly, indicating that most of the variance is captured by the components before this point.

### 1.2 Cumulative Variance

The cumulative variance plot helps visualize how much total variance is explained as more components are added. This ensures that we keep enough components to capture the majority of the data's variance.

### 1.3 Model Evaluation Using Cross-Validation

Cross-validation is used to evaluate the performance of the model with different numbers of components. It helps determine the number of components that yield the best accuracy.

## 2 Linear Discriminant Analysis (LDA)

LDA focuses on maximizing class separability by applying a linear transformation. It selects the components that contribute the most to between-class variance while minimizing within-class variance.

### 2.1 Feature Selection in LDA

LDA automatically selects the most important features that maximize class separability. This selection process enhances classification accuracy and reduces the time required for convergence.

## 3 Evaluation of Gaussian Mixture Models (GMM)

Gaussian Mixture Models (GMM) are used to classify the dataset. Cross-validation helps determine the optimal number of components for the GMM, ensuring that the model minimizes classification errors.

## 4 Question 1

The evaluation of different techniques for dimensionality reduction is crucial in ensuring that meaningful information is preserved while minimizing unnecessary data. Methods like PCA and LDA offer distinct advantages depending on the dataset and the goals of the analysis.

## 5 Question 2

Cross-validation serves as a powerful tool for evaluating model performance across different configurations, including the number of components or features selected. By comparing results obtained from various settings, we can choose the configuration that offers the best trade-off between complexity and accuracy.

## References

[Alipour Fraydani, 2024] Alipour Fraydani, A. (2024). Homework on Machine Learning problem set 5, University of Tehran. *Unpublished Manuscript*, Department of Electrical Engineering, University of Tehran.