# Linear Discriminant Analysis - LDA

ANNKAH, Abigail
aannkah@aimsammi.org

Quantum Leap Africa
African Institute for Mathematical Sciences, Rwanda

December 24, 2018

AIMS | African Institute for Mathematical Sciences RWANDA

# Dimensionality Reduction

1. **Objective**:
   To reduce the number of random variables under consideration by removing redundant and dependent features of the data.
2. Generally has two approaches:
   - **Unsupervised**
     Reduce data dimensions without considering class labels (PCA).
     Useful for data visualization and noise removal
   - **Supervised**
     Class labels are considered in feature reduction (LDA).
     Useful in Biometrics and Bioinformatics
3. Essentially, we want to select the best features that represent the data and project our data unto this new feature subspace.

# Discriminant Analysis

1. The analysis of categorical dependent variables of continuous independent variables.

2. It defines a **Discriminant Function**
   - This function is a linear combination of the independent variables that discriminate between classes of the dependent variable in a perfect manner.
   - The number of categories of the dependent variable influences the description of the Discriminant function.

# Discriminant Analysis

## Assumptions

1. Sample size
2. Samples from Gaussian Distribution
3. Homogeneity of variables
4. No multicollinearity

# LDA

1. **Objective:**
   - Perform dimensionality reduction of a feature space
   - Preserve class separability information as much as possible.
2. LDA defines a linear combination of features used as a projection that maximizes separation between 2 or more classes.
3. What about PCA?

# LDA vs PCA

## PCA

1. Recall for PCA that:
    - Ignoring class labels of the data, PCA projects the feature space unto a lower dimension.
    - This is achieved by finding the axes or directions of the feature space that maximize variance of the data set.
    - These axes are the **Principal Components** of the feature space.

# LDA vs PCA
## LDA

1. Now, LDA:
   - Finds a projection for the feature subspace by taking into consideration the class labels
   - This is achieved by finding the axes or directions of the feature space that maximize class separability information.
   - Then we define a feature subspace on these axes.

# PCA vs LDA

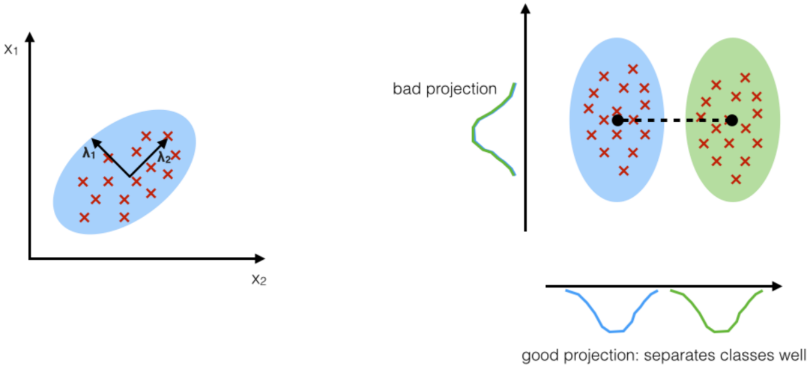


Figure: PCA maximimizing variance on the left and LDA maximizing class separability on the right

# Back to LDA

## LDA Technique

1. In order to find a good feature subspace, we need a measure of the separation between the features in the subspace.
   - Between class separability $S_B$
     This is the distance between the means of the different class and the total mean.
   - Within class separability $S_W$
     The distance between the means and samples of each class
2. These two measures are called **Scatter Matrices** of the feature space.

We define the between class variance $S_B$ and within class variance $S_W$ as

$$S_B = \sum_{i=1}^{K} n_i(\mu_i - \mu_T)(\mu_i - \mu_T)^T$$

$$S_W = \sum_{j=1}^{K} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)(x_{ij} - \mu_j)^T$$

# LDA

## Fisher's Criterion

1. Fisher proposes a solution: to maximize a function that represents the difference between the means, normalized by a measure of the within-class variability.
2. Given a class size of 2, with mean and covariances of each class as $\mu_1, S_1^2$ and $\mu_2, S_2^2$, Fisher's criterion is defined as

# LDA

## Fisher's Criterion

$$\mathcal{J}(\mathbf{w}) = \frac{\mid \mu_1 - \mu_2 \mid^2}{S_1^2 + S_2^2} = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$$

Our solution is in finding the optimum value of $\mathcal{J}$ with respect to $\mathbf{w}$ by solving

$$argmax_{\mathbf{w}} \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \iff \mathbf{w} = S_W^{-1} S_B$$

This is the transformation vector (matrix for $K \geq 3$).

# LDA

## Fisher's Criterion

1. The matrix $\mathbf{W}$ captures the goal of LDA: to minimize within class variance and increase between class variance.
2. The eigen values of $\mathbf{W}$ is used then to determine how well particular features differentiates a class.
3. By choosing the eigen vectors associated with largest K eigen values we select a good feature subspace for dimensionality reduction.

# LDA in 5 steps

1. Given a dataset of N samples $[x_i]_{i=1}^{N}$ each of which is of row of length D, that is $\mathbf{X} = [x_1, x_2, ..., x_N]^T$ of classes $K \geq 2$, we perform LDA as follows:
   - Compute the d-dimensional mean vectors for the K classes from the dataset.
   - Compute the scatter matrices
   - Compute the eigen vectors and corresponding eigen values of the ratio of the scatter matrices.
   - Sort the eigen values in decreasing order and select eigen vectors associated with the largest M to form a $d \times m$ matrix $\mathbf{W}$.
   - Transform the samples onto the new subspace using the equation $\mathbf{Y} = \mathbf{X} \times \mathbf{W}$

# Limitations of LDA

1. Parametric method assumes unimodal Gaussian distribution. If the distribution is significantly non-Gaussian, LDA may not preserve the structure need for classification.
2. Fails if the discriminatory information is not in the mean but is in the variance.
3. LDA can produce at most K-1 feature subspace for K features.

# Summary

1. LDA is used for dimensionality reduction as a preprocessing step for classification.
2. In LDA, we want to find a projection of the data unto a feature subspace where a class separation is maximized.
3. LDA uses the Fisher's criterion of maximizing between class variance and minimizing within class variance.
4. Both PCA and LDA project the data unto a lower dimensional feature space. Whiles PCA finds a projection that maximises the variance in the data set, LDA finds projection that maximize class separability information.
5. LDA is applied in areas of Agriculture, Biomedical Studies, Biometrics
6. LDA fails if classes are non-separable linearly, classification test require more features and small sample sizes are used.

# Further Reading


A. Tharwat, T. Gaber, A. Ibrahim, and A.E. Hassanien. *Linear Discriminant Analysis: A detailed tutorial*. Ai Communications (2017)

S. Balakrishnama, A. Ganapathiraju Hassanien.*Linear Discriminant Analysis: A brief tutorial*. Institute for Signal and Information Processing (2017)

S. Raschka *Linear Discriminant Analysis- Bit by Bit* (2014)