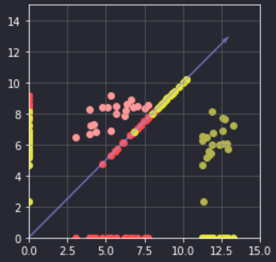
Linear Discriminant Analysis

Rudra Mukhopadhyay

March 11, 2022

Introduction to LDA

- LDA as a *feature extractor*
  - $f_E : \mathbb{R}^n \mapsto \mathbb{R}^m$ via linearly combining the original features
- LDA as a *classification* technique
  - Maximization of some "class discriminatory information"



Introduction to LDA

-Measure of separability

- $\omega_i = \{x_j = (x_{j1}, \ldots, x_{jn}) | j = 1, \ldots, p_i\}$
- Consider a vector w
- $\tilde{\omega}_i = \{y_j := \langle w, x_j \rangle \, | j = 1, \ldots, p_i\}$
- $\mu_i = E[x], \tilde{\mu}_i = E[y]$
  - $\tilde{\mu}_i = E[w^t x] = w^t E[x] = \langle w, \mu_i \rangle$

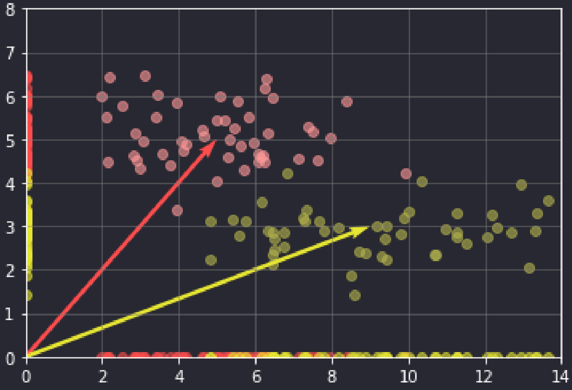
Introduction to LDA

-Measure of separability

- One possible way is to consider the separation between the means.
- $\mathcal{J}(w) = |\tilde{\mu_1} - \tilde{\mu_2}|$, for $i = 1, 2$
- Possibly $\mathcal{J}(w) = \sum_{i,j;i \neq j} |\tilde{\mu_i} - \tilde{\mu_j}|$, for multi-class set-up
- Not a convenient measure, however.
- Well separated means does not necessarily imply well-separated class clusters (fig in the following slide).

Introduction to LDA

-Measure of separability



Drawback of mean-based separation

Fisher's LDA

-Measure of separability: Fisher's LDA

- To account for the *spread* of a class, consider the within-class scatter.
- $\tilde{s}_i^2 := \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$
- Within-class scatter $\tilde{s}_W := \sum_i \tilde{s}_i^2$
- *Fisher* linear discriminant analysis: find w that maximizes
  $$\mathcal{J}(w) := \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_W}$$
- Can be interpreted as maximally separated means with maximally *squeezed* classes.

Fisher's LDA

-Measure of separability: Fisher's LDA

- We try to express $\mathcal{J}$ in terms of w.
- $\tilde{s}_i^2 = \sum_{y\in\omega_i}(y - \tilde{\mu}_i)^2 = \sum_{x\in\omega_i}(w^t x - w^t \mu_i)^2$
  $= \sum_{x\in\omega_i} w^t(x - \mu_i)(w^t(x - \mu_i))^t$
  $= \sum_{x\in\omega_i} w^t(x - \mu_i)(x - \mu_i)^t w = w^t S_i w$
- $\sum_i \tilde{s}_i^2 = \sum_i w^t S_i w = w^t S_W w$
  - $S_W := \sum_i (x - \mu_i)(x - \mu_i)^t$
  - Termed as the within-class scatter matrix
- $(\tilde{\mu}_1 - \tilde{\mu}_2)^2 = (w^t \mu_1 - w^t \mu_2)^2 = w^t(\mu_1 - \mu_2)(\mu_1 - \mu_2)^t w = w^t S_B w$
  - Where S_B is the between-class scatter

Fisher's LDA

-Solving for Fisher's linear discriminant

- $\mathcal{J}(w) = \dfrac{w^t S_B w}{w^t S_W w}$
- In order to maximize $\mathcal{J}$ wrt w, we equate $\dfrac{\partial \mathcal{J}}{\partial w}$ to 0.
- $\dfrac{\partial \mathcal{J}}{\partial w} = 0$

$$\implies \frac{\partial (w^t S_B w)}{\partial w} \frac{(w^t S_W w)^2}{(w^t S_W w)} = (w^t S_B w) \frac{\partial (w^t S_W w)}{\partial w}$$

$$\implies (2 S_B w) \cdot (w^t S_W w) = (w^t S_B w) \cdot (2 S_W w)$$

$$\implies S_B w = \frac{w^t S_B w}{w^t S_W w} S_W w$$

$$\implies S_B w = \mathcal{J}(w) S_W w$$

Fisher's LDA

-Solving for Fisher's linear discriminant

S_B is singular

- $S_B = uu^t$, where $u = (\mu_1 - \mu_2)$
- Say, $u = (a_1, \ldots, a_n)^t$
- $r_2[uu^t] = a_2(a_1, \ldots, a_n) = \frac{a_2}{a_1} a_1(a_1, \ldots, a_n) = \frac{a_2}{a_1} r_1[uu^t]$
- Thus $\text{rank}(S_B) \leq 1$ and $\text{rank}(S_B) = 1$, when $\mu_1 \neq \mu_2$
- $\det S_B = 0 \implies S_B^{-1}$ is undefined

Fisher's LDA

-Solving for Fisher's linear discriminant

- Continuing with $S_B w = \mathcal{J} S_W w$...
- $S_W^{-1} S_B w = \mathcal{J}(w) w$
  - S_B is not invertible
  - $\mathcal{J}(w)$ is a scalar
  - S_W has to be nonsingular
- Note that $\text{rank}(S_W^{-1} S_B) \leq \min\left[\text{rank}(S_W^{-1}), \text{rank}(S_B)\right]$
- Also, $\#$ distinct nonzero eigenvalues $\leq$ rank

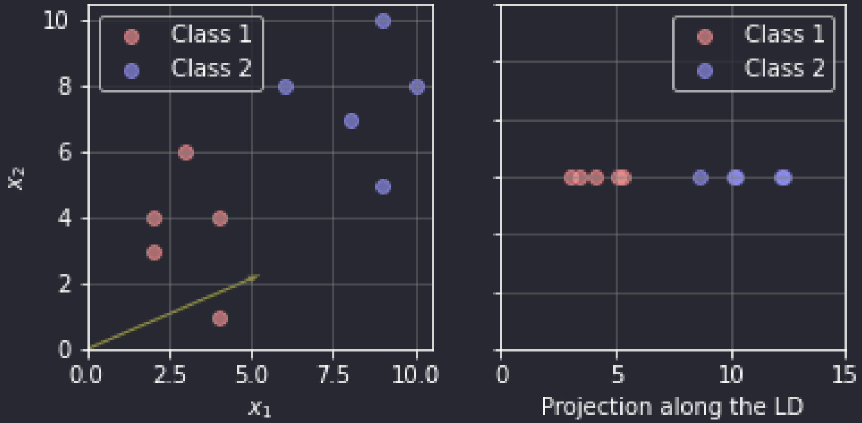
Fisher's LDA

-Solving for Fisher's linear discriminant

- w^* be the solution
- For a given distribution $(\mu_1 - \mu_2)^t w^* = \lambda$ (say)
- $S_W^{-1} S_B w^* = S_W^{-1}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^t w^* = \mathcal{J}(w^*) w^*$
  $\implies \frac{\lambda}{\mathcal{J}(w)} S_W^{-1}(\mu_1 - \mu_2) = w^*$
- As only the direction is of concern, we find-
  $w^* = S_W^{-1}(\mu_1 - \mu_2)$

Fisher's LDA

-A 1D example



Digression: Bayes Discriminant

-Gaussian with homogeneity in Σ

From Bayesian classification, we invoke the following-

Discriminant function: $g_i = -\dfrac{1}{2}(x-\mu_i)^t\Sigma^{-1}(x-\mu) + \ln P(\omega_i)$

Class boundary $(\mathcal{B}_{ij}) : v^t(x - x_0) = 0, v = \Sigma^{-1}(\mu_i - \mu_j)$

Note the solution to Fisher's LDA: $w^* = S_W^{-1}(\mu_1 - \mu_2)$

$S_W = S_1 + S_2 = \sum_{x\in\omega_i}(x-\mu_{\omega_i})(x-\mu_{\omega_i})^t = (n_1+n_2)\Sigma = N\Sigma$ (say)

$\therefore v = \left(\dfrac{1}{N}S_W\right)^{-1}(\mu_1-\mu_2) = NS_W^{-1}(\mu_1-\mu_2)$

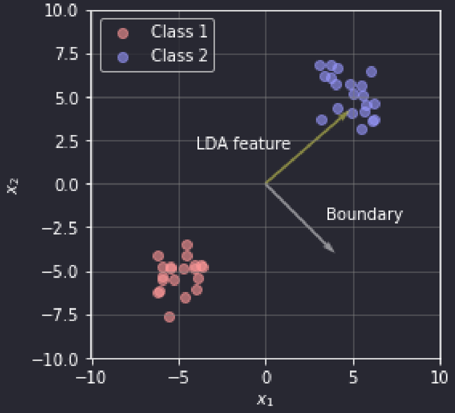
The Bayesian boundary can be re-written as-

$\mathcal{B}_{ij} : [(NS_W^{-1})(\mu_1-\mu_2)]^t(x-x_0) = 0$

$\implies Nw^{*t}(x-x_0) = 0$

Digression: Bayes Discriminant

-Gaussian with homogeneity in Σ



Fisher's LDA as a classifier

Multi-class LDA: Generalization

-For c-class set-up

- As before, $\omega_i = \{x_j = (x_{j1}, \dots, x_{jn}) | j = 1, \dots, p_i\}, i = 1, \dots, c$
- Consider $y_j = (y_{j1}, \dots, y_{j(c-1)})$ where $y_{ji} = \langle w_i, x_j \rangle$
  - Thus, $y = W^t x$, $W = [w_1 | w_2 | \dots | w_{c-1}]$
- $S_W = \sum_{i=1}^{c} S_i = \sum_{i=1}^{c} \sum_{x \in \omega_i} (x - \mu_i)(x - \mu_i)^t$
  - $\mu_i = \frac{1}{p_i} \sum_{k=1}^{p_i} x_k$
- $S_B := \sum_{i=1}^{c} p_i (\mu_i - \mu)(\mu_i - \mu)^t$
  - $\mu = \frac{1}{c} \sum_i \mu_i, i = 1, \dots, c$
- We proceed to find $\tilde{S_W}, \tilde{S_B}$

Multi-class LDA: Generalization

-For c-class set-up

- $y = W^t x$
- $\tilde{S_W} = \sum_{i=1}^{c} (y - \tilde{\mu}_i)(y - \tilde{\mu}_i)^t = \sum_{i=1}^{c} W^t S_i W = W^t S_W W$
  - $\tilde{\mu}_i = \bar{y}_{\omega_i} = W^t \bar{x}_{\omega_i} = W^t \mu_i$
- $\tilde{S_B} = \sum_{i=1}^{c} p_i (\tilde{\mu}_i - \tilde{\mu})(\tilde{\mu}_i - \tilde{\mu})^t$
  $= \sum_{i=1}^{c} p_i W^t (\mu_i - \mu)(\mu_i - \mu)^t W = W^t S_B W$
- $\mathcal{J}(W) = \frac{|\tilde{S_B}|}{|\tilde{S_W}|} = \frac{|W^t S_B W|}{|W^t S_W W|}$
- $W^* = \mathrm{argmax}_W \mathcal{J}(W)$

Multi-class LDA: Solution

-For c-class set-up

- The optimal solution for $W^* = [w_1^* | \dots | w_{c-1}^*]$ is given by- $(S_B - \lambda_i S_W) w_i^* = 0$
  - How so?
- If S_W is non-singular, W^* would contain columns that are eigenvectors of $S_W^{-1} S_B$

Multi-class LDA: Solution

-For c-class set-up

- As discussed earlier, $\text{rank}(uu^t) \leq 1$ where u is any arbitrary vector
- $\text{rank}(S_B) = \text{rank}(\sum_{i=1}^{c} p_i(\mu_i - \mu)(\mu_i - \mu)^t) \leq c - 1$
  - $a_i := (\mu_i - \mu)$
  - $\sum_{i=1}^{c} p_i a_i a_i^t = \sum_{i=1}^{c-1} p_i a_i a_i^t + p_c(\mu_c - \mu)(\mu_c - \mu)^t$
  - $\mu_c - \mu = (c - 1)\mu - \sum_{i=1}^{c-1} \mu_i$
  - $\text{rank}(S_B) \leq c - 1$
- $\text{rank}(S_W) = \text{rank}(\sum_{i=1}^{c} \sum_{x \in \omega_i}(x - \mu_i)(x - \mu_i)^t) \leq N - c$
  - $N = \sum_i p_i$

Multi-class LDA: Solution and limitaions

-For c-class set-up

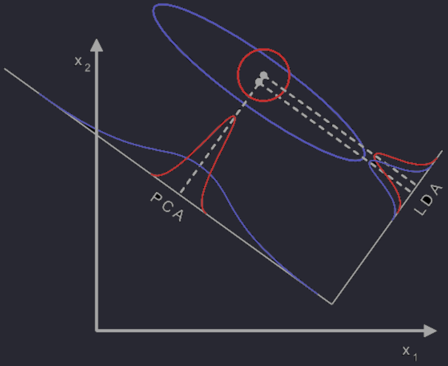
- # distinct nonzero eigenvalues $\leq$ rank $\leq c - 1$
- LDA, thus, can project upon maximum $c - 1$ features, considering $c - 1$ eigenvectors, corresponding to the largest eigenvalues as w_i's

Limitations

- The earlier discussion brings us to the first limitation: *feature-space with restricted dimensions*
- LDA might not be very effective in *separating complex distributions*
- It is bound to fail when the discriminatory information is *not explained by the means*
- S_W will be non-invertible if not $N >> c$

Multi-class LDA: Limitations

-For c-class set-up

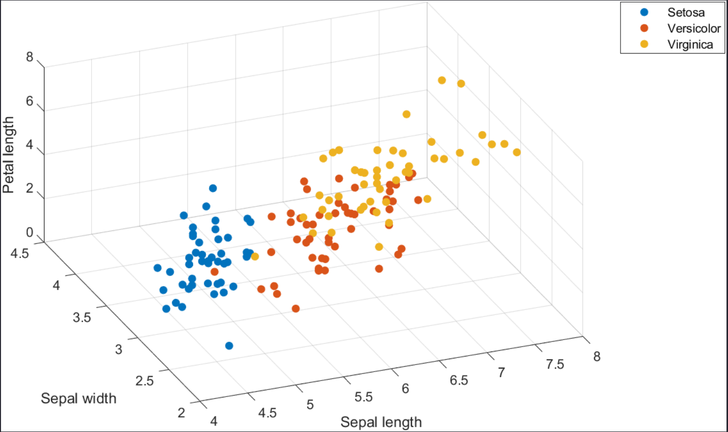


Discriminatory information in variance[1]

[1] Credit: LDA. CSCE 666 Pattern Analysis. Ricardo Gutierrez-Osuna

Multi-class LDA:

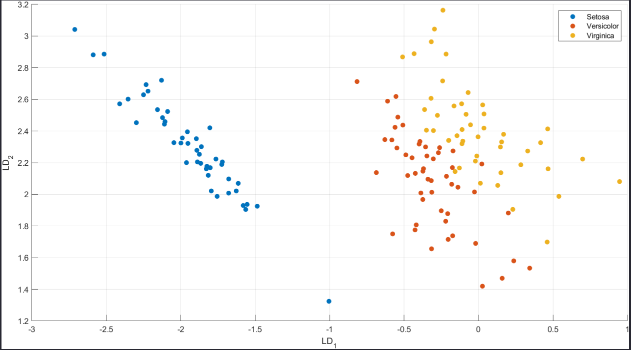
-A 3-class example on the Iris data-set



Param: sepal length, width, petal length

Multi-class LDA:

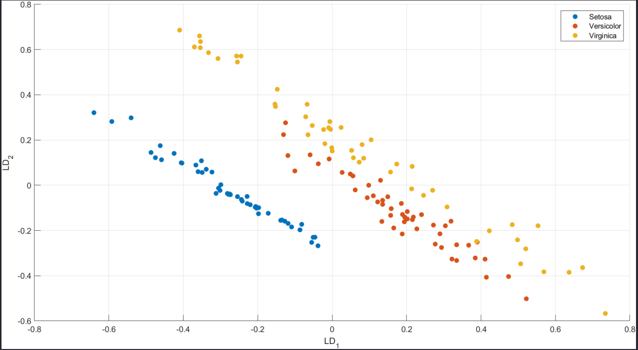
-A 3-class example on the Iris data-set



LDA considering three parameters mentioned earlier

Multi-class LDA:

-A 3-class example on the Iris data-set



Considering sepal length, width; petal length, width

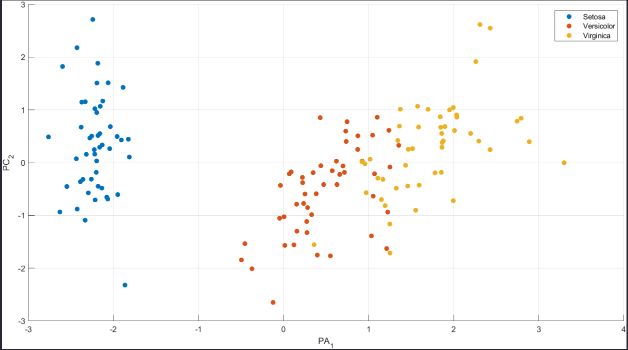
PCA against LDA:

- PCA is unsupervised; LDA needs class-labels
- PCA represents the *maximum variance*;
  LDA is about *maximizing the discriminatory information*
- *"PCA does more of feature classification and LDA does data classification"* [2]

[2] Comparision of PCA and LDA for Face Recognition. Begum, Sajjan. 2013.

PCA against LDA

-On the Iris data-set



Considering sepal length, width; petal length, width