Data Science
Linear Discriminant Analysis

Supervised latent analysis

Stéphane Marchand-Maillet

Department of Computer Science

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Master en Sciences Informatiques - Autumn semester

Table of contents

Motivation

Supervised learning

Linear Discriminant Analysis
- Discriminant direction
- Interclass criterion
- Intraclass criterion
- Generalization
- Discriminant subspaces

Examples

Inference
- Gaussian classes
- Optimally

What is the lecture about?

- ⋆ To introduce a supervised definition of latent factors
- ⋆ To propose a conditional data modeling
- ⋆ To perform classification of unlabeled items

Reading: [1] (chap 4) and [2] (chap 9)

Supervised learning

Given data $\mathcal{X} \subset \Omega \subseteq \mathbb{R}^D$ associated with categories (class labels) $\mathcal{Y} = [\![M]\!] \subset \mathbb{N}$, supervised learning is about finding (learning) the parameters $\boldsymbol{\theta}$ of a learner (function) $\phi_{\boldsymbol{\theta}}$ so as to minimize the loss $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}(\boldsymbol{\theta})$ incurred when predicting class labels using $\phi_{\boldsymbol{\theta}}$

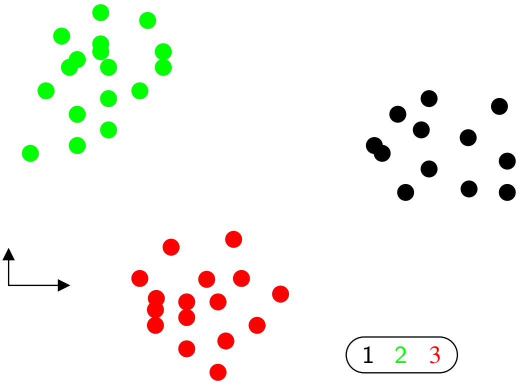
$$\begin{aligned} \phi_{\boldsymbol{\theta}} \quad : \quad & \Omega \to \mathcal{Y} \\ & \boldsymbol{x}_i \mapsto \phi_{\boldsymbol{\theta}}(x_i) = \tilde{y}_i \end{aligned}$$

Examples

- $\phi_{\boldsymbol{\theta}}$ is a logistic regression with parameters $\boldsymbol{\theta}$
- $\phi_{\boldsymbol{\theta}}$ is a neural network with weights $\boldsymbol{\theta}$
- $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}(\boldsymbol{\theta}) = \sum_{\mathcal{X}} \|\phi_{\boldsymbol{\theta}}(x_i) - y_i\|^2_{\text{some}}$
- ...

Linear Discriminant Analysis (LDA)

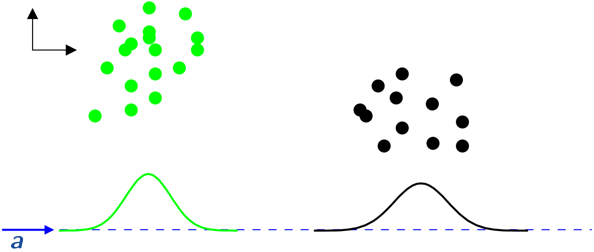
Q: Given multivariate data $\mathcal{X}$, associated with labels from $\mathcal{Y}$, we seek a model for this data



Note: This LDA is not to be mixed with "Latent Dirichlet Allocation" (related to NLP models)

Discriminant direction

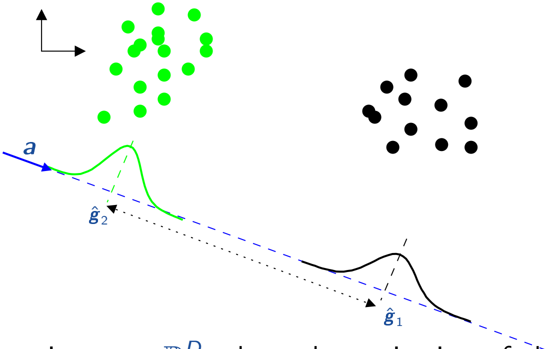
With 2 classes



we seek a direction $\boldsymbol{a} \in \mathbb{R}^D$ where the projection of the data over this direction $\hat{\mathcal{X}} = \mathrm{Proj}_{\boldsymbol{a}}(\mathcal{X})$ shows optimal properties for class discrimination

Discriminant direction

With 2 classes



we seek a direction $\boldsymbol{a} \in \mathbb{R}^D$ where the projection of the data over this direction $\hat{\mathcal{X}} = \text{Proj}_{\boldsymbol{a}}(\mathcal{X})$ shows optimal properties for class discrimination

Inter-class discrimination criterion

If the class projections are well-separated, then they can be easily discriminated:

$\Rightarrow$ Search for direction $\boldsymbol{a}$ where the inter-class discrimination is maximum

- $\star$ Clearly, $\boldsymbol{a}$ must be colinear to the line $\boldsymbol{g}_1 - \boldsymbol{g}_2$ across the centers of mass of the classes, since

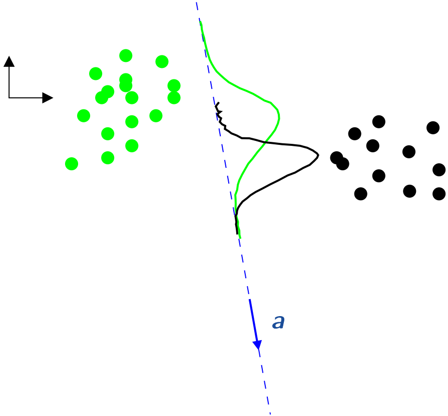
$$(\hat{\boldsymbol{g}}_1 - \hat{\boldsymbol{g}}_2)^2 = \left(\frac{\boldsymbol{a}^\top \boldsymbol{g}_1}{\|\boldsymbol{a}\|} - \frac{\boldsymbol{a}^\top \boldsymbol{g}_2}{\|\boldsymbol{a}\|}\right)^2 = \left(\frac{\boldsymbol{a}^\top}{\|\boldsymbol{a}\|}(\boldsymbol{g}_1 - \boldsymbol{g}_2)\right)^2$$

is maximum when $\boldsymbol{a} = \lambda(\boldsymbol{g}_1 - \boldsymbol{g}_2)$ φ

⚠ The original data ($\boldsymbol{x}_i \in \mathcal{X}$) does not change. It is only the projected data ($\hat{\boldsymbol{x}}_i \in \hat{\mathcal{X}}$ and all subsequent computations) that changes when evolving $\boldsymbol{a}$

Note: φ prove that the center of mass of projected data is the projection of the center of mass of the original data

Intra-class discrimination criterion



We also seek to minimize the variance of the individual projected class

Intra-class discrimination criterion

- We account for the *intra*-class variance of the projected data
- Fisher criterion maximizes

$$\underset{\boldsymbol{a}}{\text{argmax}} \frac{(\hat{\boldsymbol{g}}_1 - \hat{\boldsymbol{g}}_2)^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

where $\hat{\sigma_k}$ is the normalized variance of the projection of class k

$$\hat{\sigma}_k^2 = \frac{1}{N_k} \sum_{y_i = k} (\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{g}}_k)^{\mathsf{T}} (\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{g}}_k)$$

Generalization: inter-class criteria

- Let $\boldsymbol{B} = [\boldsymbol{g}_1 - \boldsymbol{g}, \ldots, \boldsymbol{g}_M - \boldsymbol{g}]$ be the matrix of centered data centers ($\boldsymbol{g} = \frac{1}{N}\sum_N \boldsymbol{x}_i$ and $N = \sum_k N_k$)
- $\boldsymbol{S}_\mathrm{b} = \frac{1}{M}\boldsymbol{B}\boldsymbol{B}^\mathsf{T}$ is the covariance matrix of class centers
- We maximize over $a \in \mathbb{R}^D$

$$\sum_k (\hat{\boldsymbol{g}}_k - \hat{\boldsymbol{g}})^\mathsf{T}(\hat{\boldsymbol{g}}_k - \hat{\boldsymbol{g}}) = \frac{1}{\|\boldsymbol{a}\|^2}\boldsymbol{a}^\mathsf{T}\boldsymbol{S}_\mathrm{b}\boldsymbol{a}$$

Generalization: intra-class criteria

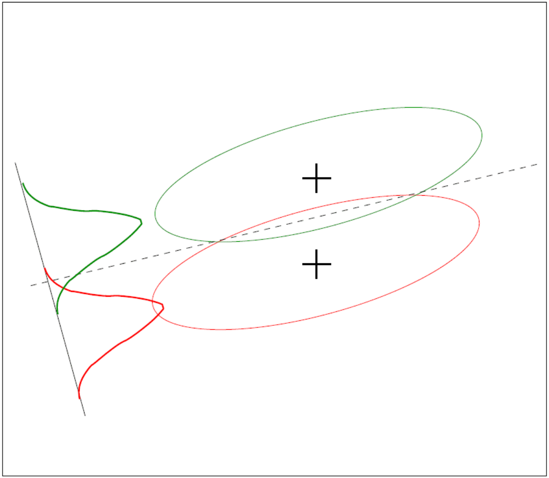
- ⋆ Let $\boldsymbol{A}_k = [\boldsymbol{x}_1 - \boldsymbol{g}_k, \ldots, \boldsymbol{x}_{N_k} - \boldsymbol{g}_k], y_i = k$ be the matrix of centered data
- ⋆ $\frac{1}{N_k}\boldsymbol{A}_k\boldsymbol{A}_k^{\mathsf{T}}$ is the *intra*-class covariance matrix
- ⋆ $\boldsymbol{S}_{\mathrm{w}} = \sum_k \frac{1}{N_k}\boldsymbol{A}_k\boldsymbol{A}_k^{\mathsf{T}}$ is the sum of *intra*-class covariance matrices
- ⋆ We minimize

$$\sum_k \sum_{y_i=k} (\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{g}}_k)^{\mathsf{T}}(\hat{\boldsymbol{x}}_i - \hat{\boldsymbol{g}}_k) = \sum_k \sum_{y_i=k} \frac{(\boldsymbol{a}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{g}_k))^{\mathsf{T}}\boldsymbol{a}^{\mathsf{T}}(\boldsymbol{x}_i - \boldsymbol{g}_k)}{\|\boldsymbol{a}\|^2}$$

$$= \sum_k \frac{1}{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{A}_k\boldsymbol{A}_k^{\mathsf{T}}\boldsymbol{a} = \frac{1}{\boldsymbol{a}^{\mathsf{T}}\boldsymbol{a}}\boldsymbol{a}^{\mathsf{T}}\boldsymbol{S}_{\mathrm{w}}\boldsymbol{a}$$

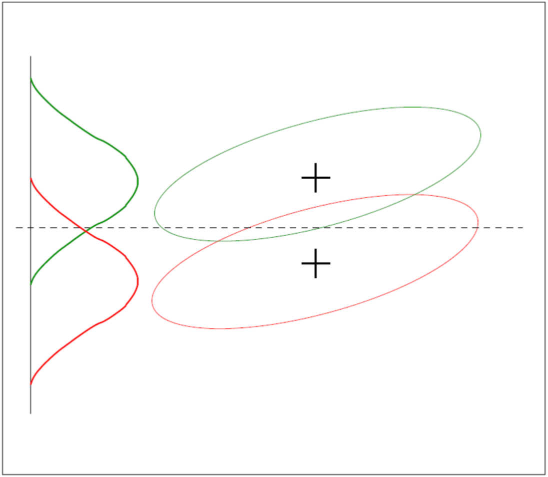
Mixing both criteria

Intra-class:



Mixing both criteria

Inter-class:



Fisher discrimination criteria

⋆ Combining intra- and inter-class criteria

$$\underset{\boldsymbol{a}}{\operatorname{argmax}}\, \mathcal{L}(\boldsymbol{a}) = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, \frac{\boldsymbol{a}^\mathsf{T} \boldsymbol{S}_\mathrm{b} \boldsymbol{a}}{\boldsymbol{a}^\mathsf{T} \boldsymbol{S}_\mathrm{w} \boldsymbol{a}}$$

which is found if:

$$\varphi \qquad \frac{\partial \mathcal{L}(\boldsymbol{a})}{\partial \boldsymbol{a}} = \frac{\boldsymbol{S}_\mathrm{b} \boldsymbol{a}(\boldsymbol{a}^\mathsf{T} \boldsymbol{S}_\mathrm{w} \boldsymbol{a}) - \boldsymbol{S}_\mathrm{w} \boldsymbol{a}(\boldsymbol{a}^\mathsf{T} \boldsymbol{S}_\mathrm{b} \boldsymbol{a})}{(\boldsymbol{a}^\mathsf{T} \boldsymbol{S}_\mathrm{w} \boldsymbol{a})^2} = 0$$

$\Rightarrow$ $\boldsymbol{a}$ is solution of the generalized eigen system: $\boldsymbol{S}_\mathrm{b} \boldsymbol{a} = \mathcal{L}(\boldsymbol{a}) \boldsymbol{S}_\mathrm{w} \boldsymbol{a}$
Hence, $\boldsymbol{a}$ is the first e.vector of $\boldsymbol{S}_\mathrm{w}^{-1} \boldsymbol{S}_\mathrm{b}$ (with e.value $\lambda_1 = \mathcal{L}(\boldsymbol{a})$) φ

Discriminant subspaces

- ⋆ eigenvectors corresponding to the largest eigenvalues λ_i are the most discriminative dimensions

$$\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots \boldsymbol{a}_p \quad \text{with} \quad \lambda_1 > \lambda_2 > \ldots \lambda_p$$

- ⋆ M classes may be discriminated in a (at most) $(M-1)$-dimensional subspaces (iterative projections)

$\Rightarrow$ only $M-1$ non-zero eigenvalues

Particular case: $M = 2$

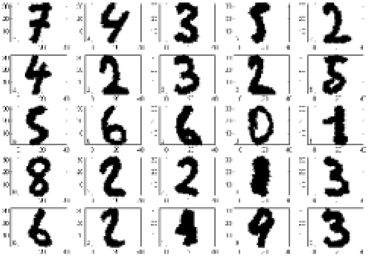
φ $\boldsymbol{B}\boldsymbol{B}^{\mathsf{T}} = (\boldsymbol{g}_1 - \boldsymbol{g})(\boldsymbol{g}_1 - \boldsymbol{g})^{\mathsf{T}} + (\boldsymbol{g}_2 - \boldsymbol{g})(\boldsymbol{g}_2 - \boldsymbol{g})^{\mathsf{T}} = (\boldsymbol{g}_1 - \boldsymbol{g}_2)(\boldsymbol{g}_1 - \boldsymbol{g}_2)^{\mathsf{T}}$

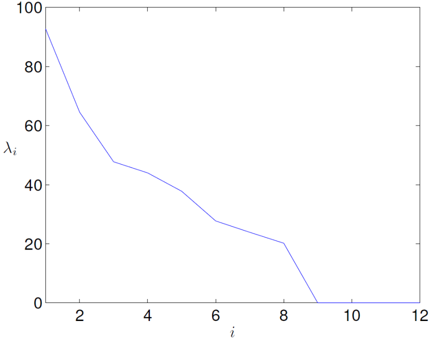
- ⋆ hence $\boldsymbol{B}\boldsymbol{B}^{\mathsf{T}}\boldsymbol{a}$ is a vector along direction $(\boldsymbol{g}_1 - \boldsymbol{g}_2)$
- ⋆ hence $\boldsymbol{a} = \lambda \boldsymbol{S}_{\mathrm{w}}^{-1}(\boldsymbol{g}_1 - \boldsymbol{g}_2)$

Character recognition

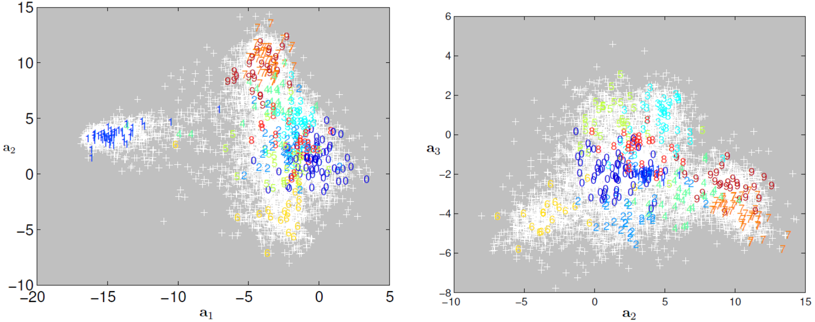
7291 images 16×16 (8 bits) numbers from 0 to 9

$\Rightarrow \{\boldsymbol{x}_i, y_i\}$ with $\boldsymbol{x}_i \in \mathbb{R}^{256}$ and $y_i \in [\![10]\!]$, $i \in [\![7291]\!]$





Projection



⇒ LDA finds the optimal subspace to (linearly) separate data along labels y_i.

LDA as a support for decision making

Conditional modeling

- ⋆ New data $j \rightarrow \boldsymbol{x}_j$ known, y_j unknown
- ⋆ To which class k point j belongs? (classification)
- ⋆ Declare class (categorical) random variable C

⇒ Predict $\mathbb{P}(C = k|X = \boldsymbol{x}_j)$ (Bayes rule):

$$\mathbb{P}(C = k|\boldsymbol{x}_j) = \frac{\mathbb{P}(\boldsymbol{x}_j|C = k)\,\mathbb{P}(C = k)}{\mathbb{P}(\boldsymbol{x}_j)}$$

Gaussian approximation

Conditional modeling

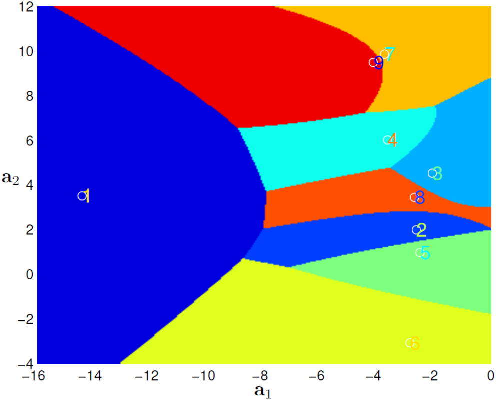
- ⋆ Each class is modeled by $\mathbb{P}(\boldsymbol{x}|C=k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \Sigma_k)$
- ⋆ Prior: $\mathbb{P}(C=k) = 1/M$
- ⋆ Evidence $\mathbb{P}(\boldsymbol{x}_j)$ is ignored

$\Rightarrow$ Maximum likelihood

$$\mathbb{P}(\boldsymbol{x}|C=k) \simeq \exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_k)^{\mathsf{T}}\Sigma_k^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right)$$

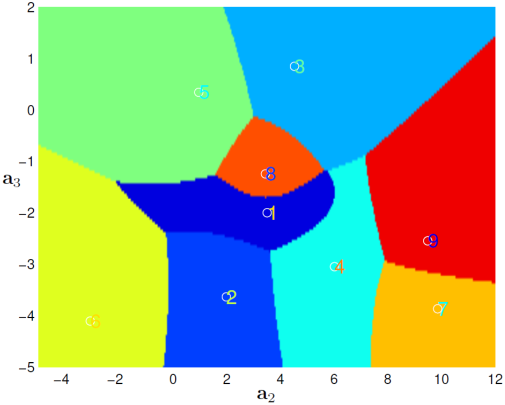
Decision (classification)

$$\delta(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, \mathbb{P}(\boldsymbol{x}|C = k)$$



Decision (classification)

$$\delta(\boldsymbol{x}) = \underset{k}{\operatorname{argmax}}\, \mathbb{P}(\boldsymbol{x}|C = k)$$

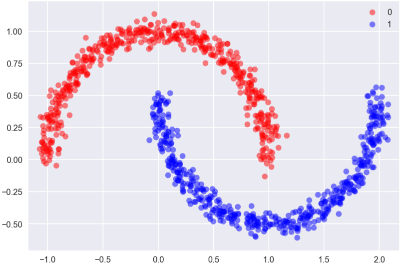


Optimality

- ⋆ LDA is optimal when the M classes are each Gaussian distributed

⇒ because of the discrimination criteria based on covariance matrices S_{w} and S_{b}

- ⋆ Linear Discriminant Analysis → does not account for non-linear relationships between variables (linear discrimination)



⇒ Non-linear classification

Summary

- ⋆ LDA is a supervised technique
- ⋆ LDA finds the optimal linear latent factors explaining (linear) discrimination
- ⋆ Fisher linear discrimination criterion combines within- and between-scatter
- ⋆ LDA is resolved by finding the eigenvalues of the covariance matrix ratio ($\boldsymbol{S}_{\mathrm{w}}^{-1}\boldsymbol{S}_{\mathrm{b}}$)
- ⋆ These latent factors may be used for visualization (accounting for labels)
- ⋆ LDA (as a classification) is an example of conditional modeling: every class has a Normal distribution

Example questions [mostly require formal – mathematical – answers]

- ⋆ Explain the setup of supervised learning
- ⋆ Why is LDA a linear method?
- ⋆ What characterizes a discriminant direction? What does $\boldsymbol{a}$ represent?
- ⋆ What is the Fisher criteria?
- ⋆ What is the inter-class (between) criteria? Explain its principle
- ⋆ How to compute the inter-class covariance matrix?
- ⋆ What is the intra-class (within) criteria? Explain its principle
- ⋆ How to compute the intra-class covariance matrix?
- ⋆ How are they both combined?
- ⋆ Can you justify and explain the derivation of the maximum for $\boldsymbol{a}$?
- ⋆ Can you describe the multidimensional situation ($D > 2$)?
- ⋆ How to do inference using LDA?
- ⋆ Provide an example of conditional data model using LDA

References I

[1] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. (available online).

[2] Kevin P. Murphy. *Probabilistic Machine Learning: an Introduction*. MIT Press, 2022. (available online).