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

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 θ of a learner (function) ϕ_θ so as to minimize the loss $\mathcal{L}_{\mathcal{X}\times\mathcal{Y}}(\theta)$ incurred when predicting class labels using ϕ_θ

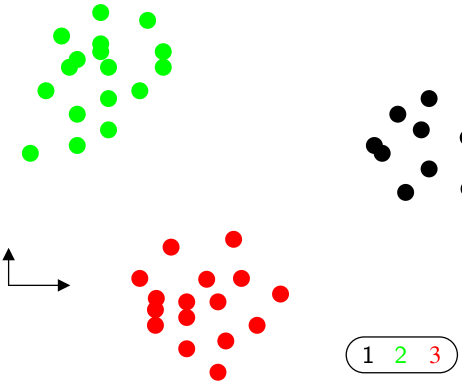
$$\begin{aligned} \phi_\theta \quad : \quad & \Omega \to \mathcal{Y} \\ & x_i \mapsto \phi_\theta(x_i) = \tilde{y}_i \end{aligned}$$

Examples

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

Linear Discriminant Analysis (LDA)

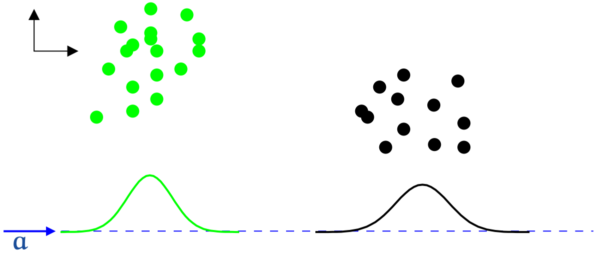
Q: Given multivariate data $\mathcal{X}$, associated with labels from $\mathcal{Y}$, we seek a model for this data



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

Discriminant direction

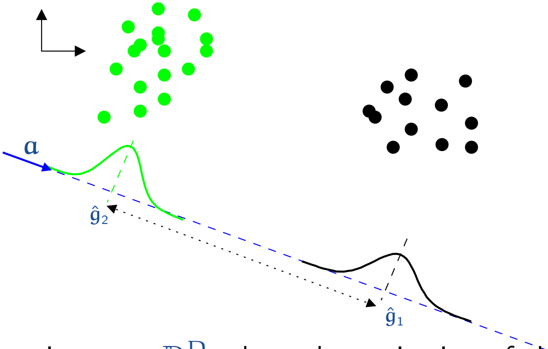
With 2 classes



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

Discriminant direction

With 2 classes



we seek a direction $\mathbf{a} \in \mathbb{R}^D$ where the projection of the data over this direction $\hat{\mathcal{X}} = \mathsf{Proj}_{\mathbf{a}}(\mathcal{X})$ 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 $\mathbf{a}$ where the inter-class discrimination is maximum

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

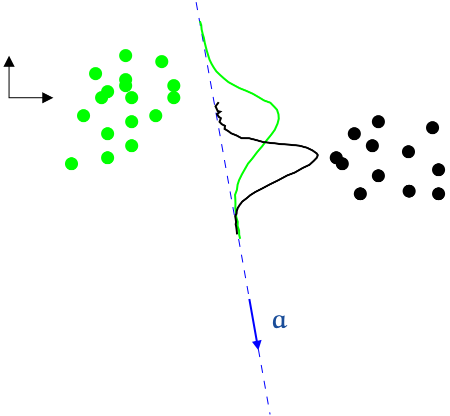
$$\|\hat{\mathbf{g}}_1 - \hat{\mathbf{g}}_2\|^2 = \left\| \frac{\mathbf{a}^T\mathbf{g}_1}{\|\mathbf{a}\|^2}\mathbf{a} - \frac{\mathbf{a}^T\mathbf{g}_2}{\|\mathbf{a}\|^2}\mathbf{a} \right\|^2 = \frac{1}{\|\mathbf{a}\|^2}(\mathbf{a}^T(\mathbf{g}_1 - \mathbf{g}_2))^2$$

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

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

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

$$\underset{\mathbf{a}}{\operatorname{argmax}} \frac{\|\hat{\mathbf{g}}_1 - \hat{\mathbf{g}}_2\|^2}{\hat{\sigma}_1^2 + \hat{\sigma}_2^2}$$

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

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

Generalization: inter-class criteria

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

$$\frac{1}{M}\sum_k (\hat{\mathbf{g}}_k - \hat{\mathbf{g}})^\mathsf{T}(\hat{\mathbf{g}}_k - \hat{\mathbf{g}}) = \frac{1}{\|\mathbf{a}\|^2}\mathbf{a}^\mathsf{T}\mathbf{S}_b\mathbf{a}$$

Generalization: intra-class criteria

- Let $\mathbf{A}_k = [\mathbf{x}_1 - \mathbf{g}_k, \ldots, \mathbf{x}_{N_k} - \mathbf{g}_k], y_i = k$ be the matrix of centered data
- $\frac{1}{N_k}\mathbf{A}_k\mathbf{A}_k^\mathsf{T}$ is the *intra*-class covariance matrix
- $\mathbf{S}_w = \sum_k \frac{1}{N_k}\mathbf{A}_k\mathbf{A}_k^\mathsf{T}$ is the sum of *intra*-class covariance matrices
- We minimize the overall distances in the projected space

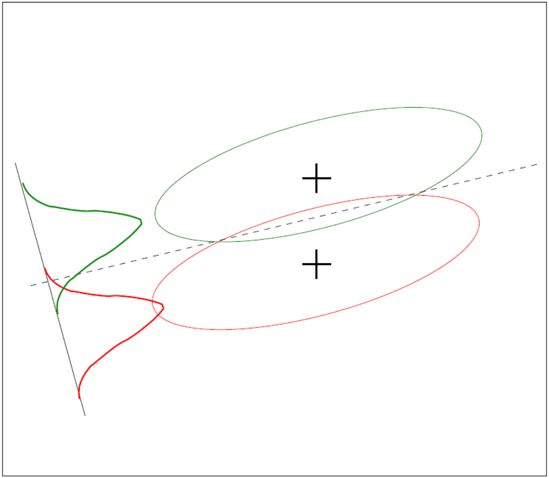
$$\sum_k \frac{1}{N_k} \sum_{y_i = k} (\hat{\mathbf{x}}_i - \hat{\mathbf{g}}_k)^\mathsf{T}(\hat{\mathbf{x}}_i - \hat{\mathbf{g}}_k)$$

$$= \sum_k \frac{1}{N_k} \sum_{y_i = k} \left(\frac{\mathbf{a}^\mathsf{T}(\mathbf{x}_i - \mathbf{g}_k)}{\|\mathbf{a}\|^2}\mathbf{a}\right)^\mathsf{T} \frac{\mathbf{a}^\mathsf{T}(\mathbf{x}_i - \mathbf{g}_k)}{\|\mathbf{a}\|^2}\mathbf{a}$$

$$= \sum_k \frac{1}{N_k}\frac{1}{\mathbf{a}^\mathsf{T}\mathbf{a}}\mathbf{a}^\mathsf{T}\mathbf{A}_k\mathbf{A}_k^\mathsf{T}\mathbf{a} = \frac{1}{\mathbf{a}^\mathsf{T}\mathbf{a}}\mathbf{a}^\mathsf{T}\mathbf{S}_w\mathbf{a}$$

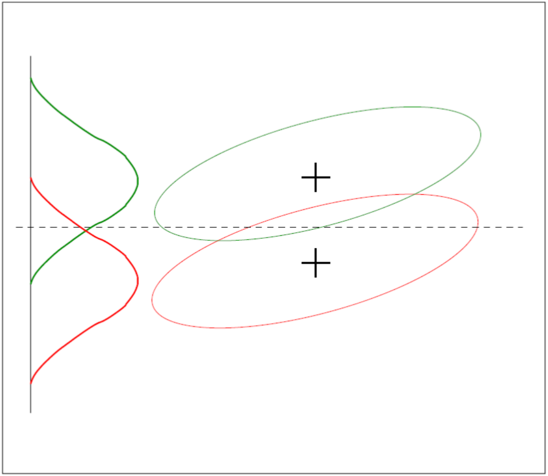
Mixing both criteria

Intra-class:



Mixing both criteria

Inter-class:



Fisher discrimination criteria

⋆ Combining intra- and inter-class criteria

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

which is found if:

φ

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

⇒ $\mathbf{a}$ is solution of the generalized eigen system: $\mathbf{S}_\mathrm{b}\mathbf{a} = \mathcal{L}(\mathbf{a})\mathbf{S}_\mathrm{w}\mathbf{a}$

Hence, $\mathbf{a}$ is the first e.vector of $\mathbf{S}_\mathrm{w}^{-1}\mathbf{S}_\mathrm{b}$ (with e.value $\lambda_1 = \mathcal{L}(\mathbf{a})$) φ

Discriminant subspaces

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

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

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

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

Particular case: $M = 2$

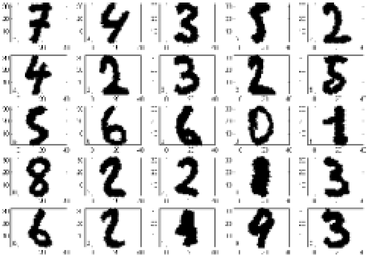
φ $\mathbf{B}\mathbf{B}^\mathsf{T} = (\mathbf{g}_1 - \mathbf{g})(\mathbf{g}_1 - \mathbf{g})^\mathsf{T} + (\mathbf{g}_2 - \mathbf{g})(\mathbf{g}_2 - \mathbf{g})^\mathsf{T} = (\mathbf{g}_1 - \mathbf{g}_2)(\mathbf{g}_1 - \mathbf{g}_2)^\mathsf{T}$

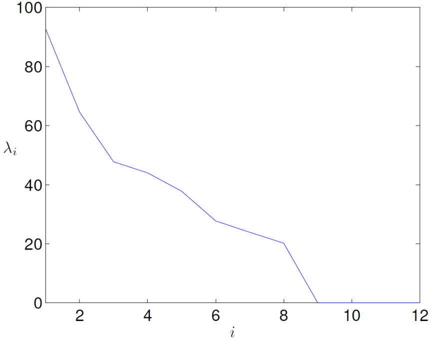
- ⋆ hence $\mathbf{B}\mathbf{B}^\mathsf{T}\boldsymbol{\alpha}$ is a vector along direction $(\mathbf{g}_1 - \mathbf{g}_2)$
- ⋆ hence $\boldsymbol{\alpha} = \lambda S_w^{-1}(\mathbf{g}_1 - \mathbf{g}_2)$

Character recognition

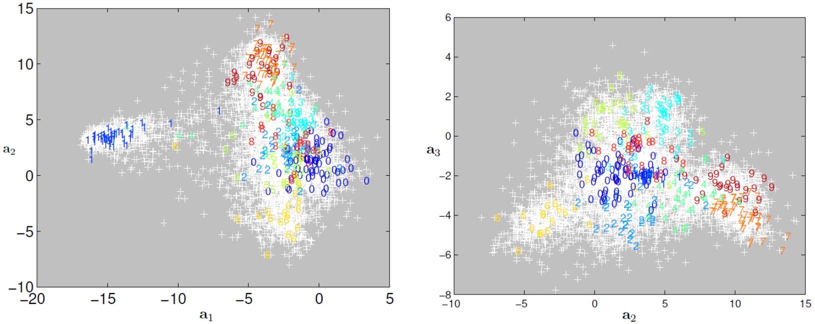
7291 images 16×16 (8 bits) numbers from 0 to 9

$\Rightarrow \{x_i, y_i\}$ with $x_i \in \mathbb{R}^{256}$ 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 $j \rightarrow x_j$ known, y_j unknown
- ⋆ To which class k point j belongs? (classification)
- ⋆ Declare class (categorical) random variable C

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

$$\mathbb{P}(C = k|x_j) = \frac{\mathbb{P}(x_j|C = k)\mathbb{P}(C = k)}{\mathbb{P}(x_j)}$$

Gaussian approximation

Conditional modeling

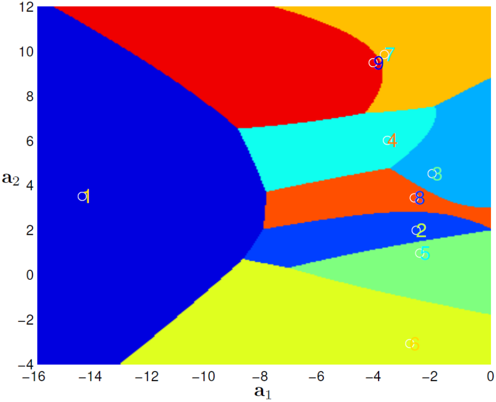
- ⋆ Each class is modeled by $\mathbb{P}(x|C=k) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
- ⋆ Prior: $\mathbb{P}(C=k) = 1/M$
- ⋆ Evidence $\mathbb{P}(x_j)$ is ignored

$\Rightarrow$ Maximum likelihood

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

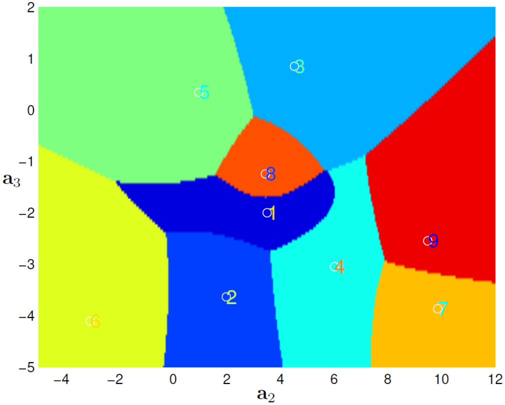
Decision (classification)

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



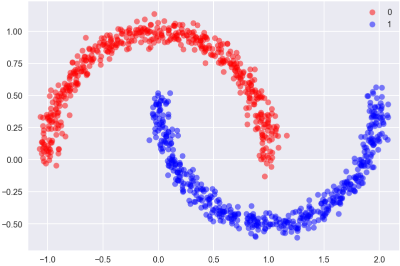
Decision (classification)

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



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 ($\mathbf{S}_{\mathrm{w}}^{-1}\mathbf{S}_{\mathrm{b}}$)
- ⋆ 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 α 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 α?
- ⋆ 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).

License