

Introduction to machine learning

Maxime Sangnier

December 16, 2020

Contents

Introduction	5
1 Classification	7
1.1 Discriminant analysis	7
1.1.1 The multivariate normal distribution	7
1.1.2 Bayes classifier for multivariate normal distributions	8
1.1.3 Fisher discriminant analysis	11
1.1.4 Kernel Fisher discriminant analysis	15
1.1.5 Multiclass linear discriminant	16
1.2 Logistic regression	17
1.2.1 Model and risk	17
1.2.2 Maximum likelihood estimation	19
1.2.3 Logistic regression versus linear discriminant analysis (LDA)	20
1.3 Boosting	22
1.3.1 Adaboost	22
1.3.2 ERM point of view and remarks	27
1.3.3 Gradient boosting	28
1.4 Support vector machines	32
1.4.1 Large margin classifier	32
1.4.2 RKHS	34
1.4.3 Kernel trick and nonlinear SVM	39
1.4.4 SVM in action	41
1.4.5 Duality in convex optimization	41
1.4.6 Dual problem and support vectors	46
1.4.7 Statistical perspective	49
1.5 A detour to nonparametric regression	50
1.5.1 Least mean squares	50
1.5.2 Least absolute deviations	51
1.5.3 Support vector regression	51
1.6 Other methods	54
1.6.1 k-nearest neighbors	54
1.6.2 Decision trees	54
1.6.3 Bagging	56
1.6.4 Random forests	56

1.7	Exercises	57
1.7.1	Discriminant analysis	57
1.7.2	Boosting	57
1.7.3	SVM	59
1.7.4	Regression	60
2	Clustering	65
2.1	Gaussian mixtures	66
2.1.1	Mixture model	66
2.1.2	A toy example	68
2.1.3	EM for Gaussian mixtures (soft k-means)	70
2.1.4	The general EM algorithm	73
2.1.5	Model selection	77
2.2	Cost minimization methods	78
2.2.1	Center-based objectives	78
2.2.2	k-means algorithm	81
2.2.3	Point-based objectives	86
2.2.4	Similarity graphs	87
2.2.5	Spectral clustering	88
2.2.6	Properties of graph Laplacians	94
2.2.7	Practical details	95
2.3	Hierarchical clustering	96
2.3.1	Agglomerative approaches	96
2.3.2	Connection with minimum spanning trees	97
2.4	Density-based clustering	99
2.5	Clustering evaluation	100
2.5.1	Elbow method	100
2.5.2	Silhouette coefficient	101
2.5.3	Calinski-Harabasz index	102
3	Dimensionality reduction	104
3.1	Linear methods	106
3.1.1	Principal component analysis	106
3.1.2	Link with variance maximization	110
3.1.3	Link with the Gram matrix	111
3.1.4	Link with singular values	112
3.1.5	Random projection	114
3.1.6	Reconstruction of random projections	117
3.2	Nonlinear methods	119
3.2.1	Kernel principal component analysis	119
3.2.2	Classical multidimensional scaling	123
3.2.3	Metric and nonmetric multidimensional scaling	127
3.3	Other methods	130
3.3.1	Spectral embedding	130
3.3.2	Linear discriminant analysis	130

3.4	Exercises	131
3.4.1	Random projection	131
4	Previous exams	132
	Exam 2019	132
	Retake 2019	136
	References	139

List of Algorithms

1	Adaboost.	23
2	Gradient boosting.	30
3	Sequential minimal optimization.	46
4	Sampling of a mixture model.	67
5	EM for Gaussian mixtures (soft k-means).	72
6	EM algorithm.	74
7	EM algorithm (maximization-maximization).	77
8	k-means.	83
9	k-means++.	86
10	Unnormalized spectral clustering.	91
11	Normalized spectral clustering (with L_w).	92
12	Normalized spectral clustering (with L_s).	92
13	DBSCAN.	100
14	Reduced representation by principal component analysis (PCA).	113
15	Classical multidimensional scaling.	127
16	SMACOF.	130

Introduction

This course comes as a complement to Pr Biau's course on statistical learning, and this in two directions:

1. it tackles both supervised (Chapter 1) and unsupervised learning (Chapters 2 and 3);
2. it presents an algorithmic point of view and comes with practical homeworks.

This explains why some major methods, like k-nearest neighbors, decision trees and random forests are only skimmed over.

These lectures notes are organized in three chapters:

Chapter 1: a few classification methods are introduced in details and we bridge quickly the gap between classification and regression:

- ◇ linear and quadratic discriminant analysis (LDA, QDA);
- ◇ Fisher discriminant analysis (FDA);
- ◇ kernel Fisher discriminant analysis (KFDA);
- ◇ multiclass linear discriminant analysis;
- ◇ logistic regression;
- ◇ Adaboost and gradient boosting;
- ◇ support vector machines (SVM) for classification (SVC) and regression (SVR).

Chapter 2: we consider the problem of unobserved labels and present some methods to produce a partition of the input space:

- ◇ expectation-maximization for Gaussian mixtures (soft k-means);
- ◇ k-means algorithm;
- ◇ spectral clustering;
- ◇ hierarchical agglomerative clustering;
- ◇ density-based spatial clustering of applications with noise (DBSCAN).

Chapter 3: the curse of dimensionality is quickly addressed and some dimensionality reduction techniques (linear or not) are presented:

- ◇ principal component analysis (PCA);
- ◇ random projections;
- ◇ kernel principal component analysis (KPCA);
- ◇ multidimensional scaling (MDS).

In all chapters, we start from a generative (or statistical modeling) point of view and gently slide to the discriminative angle, keeping in mind Vapnik's principle: avoiding a more general (and potentially more difficult) task than the one we aim at.

Moreover, many methods are explained with a probabilistic point of view (namely, we consider a random variable X or a pair of random variables (X, Y) , respectively for unsupervised and supervised learning) but in practice, we assume that people are provided with a sample $\{X_1, \dots, X_n\}$ (respectively $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$) and all formulas can be transformed to an empirical twin by considering the empirical distribution $\frac{1}{n} \sum_{i=1}^n \delta_{\{X_i\}}$ (respectively $\frac{1}{n} \sum_{i=1}^n \delta_{\{(X_i, Y_i)\}}$), where δ represents the Dirac measure.


Supervised (X, Y)		Unsupervised X	
Classification Y discrete	Regression Y continuous	Clustering	Reduction
$\text{0-1: } \min_f P(Y \neq f(X))$	<u>Model</u> : $Y = f(X) + \varepsilon$ $\text{L2: } \min_f E[(Y - f(X))^2]$	Minimal distortion (like quantification) 	$\begin{aligned} \phi: \mathbb{R}^d &\rightarrow \mathbb{R}^p && \text{compress} \\ \psi: \mathbb{R}^p &\rightarrow \mathbb{R}^d && \text{reconstruction} \\ \min E[(X - (\psi \circ \phi)(X))^2] \end{aligned}$
Hypothesis on data: LDA, FDA Hypothesis on the Bayes classifier: LR Nonparametric & Ensemble methods: SVM, tree, boosting Geometrical approach: SVM	Non parametric: SVR	Hypothesis on data: ET Non parametric: K-means Graph model: Spectral clustering Geometrical approach: agglomerative clustering, density clustering	Linear ϕ, ψ : PCA Distance saving criterion: → Random approach: random project → Geometrical approach: TRS

Figure 1: The big picture of the course.

Chapter 1

Classification

Classification focuses on a pair of random variables $(X, Y) \in \mathbb{R}^d \times [C]$, where C is a positive integer, and Y is a label characterizing the class of X . The bracket notation is for indexing integers: $[C] = \{1, 2, \dots, C\}$. If there is no ambiguity, with a slight abuse, we may consider that $[2] = \{-1, +1\} = \{\pm 1\}$ (this appears for binary classification). The aim of classification is to predict Y given X with minimal error (*i.e.* finding $f : \mathbb{R}^d \rightarrow [C]$ such that $\mathbb{P}(Y \neq f(X))$ is minimal), based on a sample $\{(X_i, Y_i)\}_{1 \leq i \leq n}$. This is the most pleasant situation of statistical learning since we observe both X and Y , and Y is discrete.

In this chapter, we describe several methods of classification, from a statistical modeling point of view to a discriminative one. We propose to make, at the end of this chapter, a detour to regression. Regression is very similar to classification, but Y is continuous ($Y \in \mathbb{R}$) instead of being discrete. In practice, this boils down to replacing the loss function used in variational formulations used to build estimators.

1.1 Discriminant analysis

1.1.1 The multivariate normal distribution

Definition 1.1.1. Let $\mu \in \mathbb{R}^d$, Σ be a positive definite (PD) matrix. We write $X \sim \mathcal{N}(\mu, \Sigma)$ when the Lebesgue density of X is

$$x \in \mathbb{R}^d \mapsto |2\pi\Sigma|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)} = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)},$$

where $|\Sigma|$ is the determinant of Σ . In addition, we have

$$\mathbb{E} X = \mu, \quad \mathbb{V}(X) = \Sigma,$$

where $\mathbb{V}(X)$ is the covariance matrix of X .

Proposition 1. Let $\mu^* \in \mathbb{R}^d$, Σ^* be a PD matrix and $\{X_1, \dots, X_n\}$ be a sample independent and

identically distributed (iid) according to $\mathcal{N}(\mu^*, \Sigma^*)$.

Then

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

and

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top$$

are maximum likelihood estimators (MLEs) respectively of μ^* and Σ^* .

The proof is a good exercise.

Proposition 2 ([Anderson, 2003, Chapters 3 and 5]). For a positive integer C , let $\{(X_1^j, \dots, X_{n_j}^j)\}_{1 \leq j \leq C}$ be C independent samples such that each sample $(X_1^j, \dots, X_{n_j}^j)$ (for all $j \in [C]$) is iid according to $\mathcal{N}(\mu_j, \Sigma)$, where $\mu_j \in \mathbb{R}^d$ and Σ is a PD matrix of size d .

Then for each $j \in [C]$:

$$\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i^j$$

is an unbiased and normally distributed estimate of μ_j and

$$\hat{\Sigma} = \frac{1}{\sum_{j=1}^C n_j - C} \sum_{j=1}^C \sum_{i=1}^{n_j} (X_i^j - \hat{\mu}_j)(X_i^j - \hat{\mu}_j)^\top$$

is an unbiased estimate of Σ .

1.1.2 Bayes classifier for multivariate normal distributions

Let C be a positive integer and $(X, Y) \in \mathbb{R}^d \times [C]$ be a random pair of variables, where Y is a label characterizing the class of X . We are interested in computing the Bayes classifier when each class $i \in [C]$ is normally distributed: there exists a PD matrix Σ_i and a vector $\mu_i \in \mathbb{R}^d$ such that

$$X|Y = i \sim \mathcal{N}(\mu_i, \Sigma_i).$$

As a reminder, a Bayes classifier for classifying X is defined by:

$$\forall x \in \mathbb{R}^d: \quad g^*(x) \in \arg \max_{i \in [C]} \mathbb{P}(Y = i | X = x).$$

Proposition 3. Let us assume that each class is normally distributed and let $\pi_i = \mathbb{P}(Y = i)$ be class

prior probabilities, for all $i \in [C]$. Then, a Bayes classifier g^* is defined by:

$$\forall x \in \mathbb{R}^d: \quad g^*(x) \in \arg \max_{i \in [C]} \log(\pi_i) - \frac{1}{2} \log |\Sigma_i| - \frac{1}{2} (x - \mu_i)^\top \Sigma_i^{-1} (x - \mu_i).$$

The proof is a good exercise.

Remark 1.1.1. When $\pi_1 = \dots = \pi_C$ and $\Sigma_1 = \dots = \Sigma_C = I_d$, the Bayes classifier g^* boils down to be the minimum distance to center classifier.

Let us now assume that we have only two classes ($C = 2$) and let us analyze a Bayes classifier for multivariate normal distributions. As a reminder, we have

$$g^*: x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } \mathbb{P}(Y = 1|X = x) > \mathbb{P}(Y = -1|X = x) \\ -1 & \text{otherwise.} \end{cases}$$

Proposition 4 (Linear discriminant analysis (LDA)). Let us assume that $C = 2$ and that each class is normally distributed with equal covariance, denoted Σ . Let $\pi_i = \mathbb{P}(Y = i)$ be class prior probabilities, for all $i \in [2]$, and let us denote

$$h: x \in \mathbb{R}^d \mapsto (\mu_1 - \mu_{-1})^\top \Sigma^{-1} x$$

$$b = \frac{1}{2} (\mu_{-1}^\top \Sigma^{-1} \mu_{-1} - \mu_1^\top \Sigma^{-1} \mu_1) + \log \left(\frac{\pi_1}{\pi_{-1}} \right).$$

Then, a Bayes classifier is

$$g^*: x \in \mathbb{R}^d \mapsto \text{sign}(h(x) + b),$$

where

$$\text{sign}: x \in \mathbb{R} \mapsto \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise.} \end{cases}$$

The proof is a good exercise.

Remark 1.1.2. Under the LDA assumptions and when $\pi_1 = \pi_{-1}$, we have:

$$g^*(x) = 1 \iff (x - \mu_1)^\top \Sigma^{-1} (x - \mu_1) < (x - \mu_{-1})^\top \Sigma^{-1} (x - \mu_{-1}),$$

i.e. if and only if x is closer to μ_1 than μ_{-1} with respect to the Mahalanobis distance ruled by Σ . This is similar to whitening the data with $\Sigma^{-\frac{1}{2}}$ and considering the Euclidean distance.

Using such a metric makes sens, as shown on Figure 1.1.

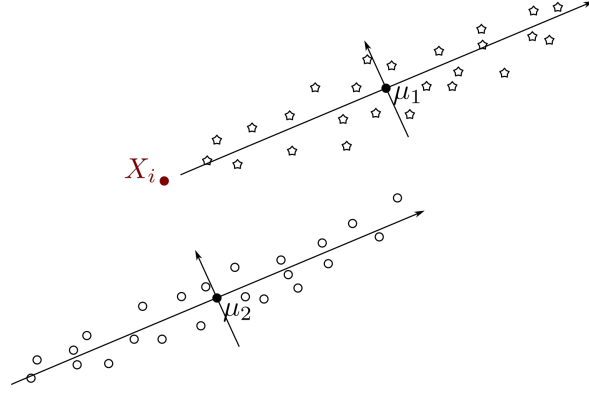


Figure 1.1: Here, the point X_i is closer to μ_2 in the Euclidean distance while it appears naturally that it belongs to the group of data centered in μ_1 . The Mahalanobis distance makes it possible to rectify this misbehavior.

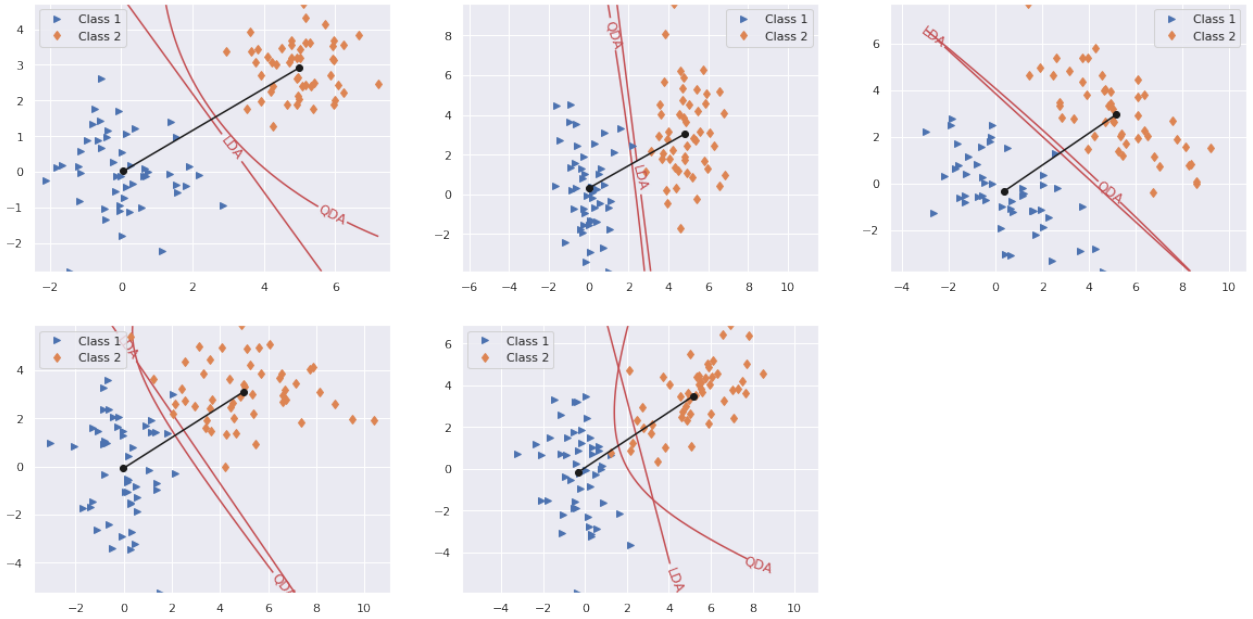


Figure 1.2: Comparison of LDA and QDA on different simulated datasets (Gaussian classes with potentially different covariance matrices).

Proposition 5 (Quadratic discriminant analysis (QDA)). *Let us assume that $C = 2$ and that each class is normally distributed. Let $\pi_i = \mathbb{P}(Y = i)$ be class prior probabilities, for all $i \in [2]$, and let us denote*

$$h: x \in \mathbb{R}^d \mapsto \frac{1}{2}x^\top(\Sigma_{-1}^{-1} - \Sigma_1^{-1})x + (\mu_1^\top \Sigma_{-1}^{-1} - \mu_{-1}^\top \Sigma_{-1}^{-1})x$$

$$b = \frac{1}{2}(\mu_{-1}^\top \Sigma_{-1}^{-1} \mu_{-1} - \mu_1^\top \Sigma_1^{-1} \mu_1) - \frac{1}{2} \log \left(\frac{|\Sigma_1|}{|\Sigma_{-1}|} \right) + \log \left(\frac{\pi_1}{\pi_{-1}} \right).$$

Then, a Bayes classifier is

$$g^*: x \in \mathbb{R}^d \mapsto \begin{cases} 1 & \text{if } h(x) + b > 0 \\ -1 & \text{otherwise.} \end{cases}$$

The proof is a good exercise.

LDA exhibits that for Gaussian data, the optimal classifier is linear. The same kind of result can be obtained for least squares regression, as exemplified by Proposition 6.

Proposition 6 (Linear regression). *Let (X, Y) be a pair of random variables with values in $\mathbb{R}^d \times \mathbb{R}$ such that $\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ m \end{pmatrix}, \begin{pmatrix} \Sigma & \ell \\ \ell^\top & \sigma^2 \end{pmatrix}\right)$, where $\mu \in \mathbb{R}^d$, $m \in \mathbb{R}$, $\Sigma \in \mathbb{R}^{d \times d}$, $\ell \in \mathbb{R}^d$, $\sigma > 0$ such that $\begin{pmatrix} \Sigma & \ell \\ \ell^\top & \sigma^2 \end{pmatrix}$ is PD. Let $w = \Sigma^{-1}\ell$ and $\sigma'^2 = \sigma^2 - \ell^\top \Sigma^{-1}\ell$. Then,*

$$\forall x \in \mathbb{R}^d, \quad [Y|X = x] \sim \mathcal{N}(m + w^\top(x - \mu), \sigma'^2),$$

and in particular, $\mathbb{E}[Y|X = x] = m + w^\top(x - \mu)$.

Proof. Without loss of generality, consider that $\begin{pmatrix} \mu \\ m \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$. Let $\tilde{Y} = w^\top X$ and $A = \begin{pmatrix} I_d & 0 \\ -w^\top & 1 \end{pmatrix}$. Then,

$$\begin{pmatrix} X \\ Y - \tilde{Y} \end{pmatrix} = A \begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N}\left(0, A \begin{pmatrix} \Sigma & \ell \\ \ell^\top & \sigma^2 \end{pmatrix} A^\top\right) \sim \mathcal{N}\left(0, \begin{pmatrix} \Sigma & 0 \\ 0 & \sigma'^2 \end{pmatrix}\right).$$

As a consequence, $\begin{pmatrix} X \\ Y - \tilde{Y} \end{pmatrix}$ is a Gaussian random vector with a block diagonal covariance matrix. As a result, X is independent of $Y - \tilde{Y}$ and for all $x \in \mathbb{R}^d$,

$$\mathbb{E}[Y - \tilde{Y}|X = x] = \mathbb{E}[Y - \tilde{Y}] = 0, \quad \mathbb{V}[Y - \tilde{Y}|X = x] = \mathbb{V}[Y - \tilde{Y}] = \sigma'^2.$$

It follows $\mathbb{E}[Y|X = x] = \mathbb{E}[\tilde{Y}|X = x] = w^\top x$, $\mathbb{V}[Y|X = x] = \mathbb{V}[Y - \tilde{Y}|X = x] = \sigma'^2$ and $[Y|X = x] \sim \mathcal{N}(w^\top x, \sigma'^2)$. \square

1.1.3 Fisher discriminant analysis

Fisher discriminant analysis explores linear regression with weaker assumptions on data than linear discriminant analysis. In practice, it is only assumed that for all $i \in [2]$, $\mathbb{E}[X|Y = i]$ and $\mathbb{V}[X|Y = i]$ exist, meaning that the classes are sufficiently concentrated around their means.

Fisher discriminant analysis aims at finding a direction $w \in \mathbb{R}^d \setminus \{0\}$ such that the projection of X onto that direction maximizes the variance between classes while minimizing the variances within classes (see Figure 1.3).

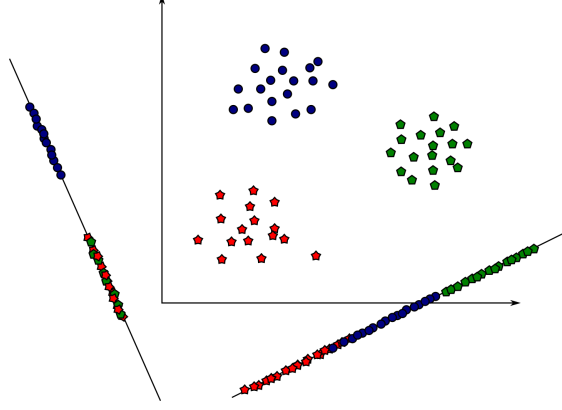


Figure 1.3: Subspace with maximal Rayleigh quotient.

More formally, we are interested in minimizing the Rayleigh quotient:

$$\underset{w \in \mathbb{R}^d \setminus \{0\}}{\text{maximize}} \quad r(w) = \frac{\mathbb{V}(\mathbb{E}(w^\top X|Y))}{\mathbb{E}(\mathbb{V}(w^\top X|Y))}, \quad (\text{P1})$$

where for any random pair of variables $(U, V) \in \mathbb{R}^2$, $\mathbb{V}(U|V) = \mathbb{E}((U - \mathbb{E}(U|V))^2 | V)$. Denoting $\mu = \mathbb{E}X$ and, for each $i \in [C]$, $\mu_i = \mathbb{E}(X|Y = i)$, $\Sigma_i = \mathbb{V}(X|Y = i)$ and $\pi_i = \mathbb{P}(Y = i)$, we remark that:

$$\begin{cases} \mathbb{E}(X|Y) \sim \sum_{i=1}^C \pi_i \delta_{\mu_i} \\ \mathbb{V}(X|Y) \sim \sum_{i=1}^C \pi_i \delta_{\Sigma_i}. \end{cases}$$

Thus, $\forall w \neq 0$:

$$r(w) = \frac{w^\top \left(\sum_{i=1}^C \pi_i (\mu_i - \mu)(\mu_i - \mu)^\top \right) w}{w^\top \left(\sum_{i=1}^C \pi_i \Sigma_i \right) w}.$$

Let us assume that $C = 2$. Then, we have $\mu = \pi_1 \mu_1 + (1 - \pi_1) \mu_{-1}$ and the Rayleigh quotient becomes

$$\begin{aligned} r(w) &= \frac{w^\top \left(\pi_1 (\mu_1 - \mu)(\mu_1 - \mu)^\top + (1 - \pi_1) (\mu_{-1} - \mu)(\mu_{-1} - \mu)^\top \right) w}{w^\top (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w} \\ &= \pi_1 (1 - \pi_1) \frac{(w^\top \mu_1 - w^\top \mu_{-1})^2}{w^\top (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w}. \end{aligned}$$

Proposition 7 (Fisher's linear discriminant). *Let us assume that $C = 2$ and that $\mu_1 \neq \mu_{-1}$. Then*

$$\text{span} \left((\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1})^{-1} (\mu_1 - \mu_{-1}) \right) \setminus \{0\} = \arg \max_{w \in \mathbb{R}^d \setminus \{0\}} r(w).$$

Proof. Let $M = (\mu_1 - \mu_{-1})(\mu_1 - \mu_{-1})^\top$ and $\Sigma = \pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}$. Then $\forall w \in \mathbb{R}^d \setminus \{0\}$, $r(w) = \pi_1 (1 - \pi_1) \frac{w^\top M w}{w^\top \Sigma w}$.

Proof 1

We have, $\forall \lambda \neq 0, r(\lambda w) = r(w)$. As a consequence, we can restrict the search of a maximizer to $\|w\|_{\ell_2} = 1$:

$$\underset{w \in \mathbb{R}^d : \|w\|_{\ell_2} = 1}{\text{maximize}} \quad r(w).$$

For all $w \in \mathbb{R}^d : \|w\|_{\ell_2} = 1$:

1. $r(w) \geq 0$;
2. $r(w) = \pi_1(1 - \pi_1) \frac{w^\top M w}{\|w\|_{\ell_2}^2} \frac{\|w\|_{\ell_2}^2}{w^\top \Sigma w} \leq \pi_1(1 - \pi_1) \frac{\|\mu_1 - \mu_{-1}\|_{\ell_2}^2}{\lambda_{\min}}$, where $\lambda_{\min} > 0$ is the smallest eigenvalue of Σ ;
3. r is continuous and differentiable in w .

Since $\{w \in \mathbb{R}^d : \|w\|_{\ell_2} = 1\}$ is a sphere and r is continuous and bounded on this set, then the restriction of r on this sphere, denoted $r|_{\|w\|_{\ell_2}=1}$, attains its minimum and its maximum. Moreover, they are critical points, that is the gradient of $r|_{\|w\|_{\ell_2}=1}$ is zero for these points. Let P_w be the orthogonal projector on $\text{span}(w)^\perp$. Then, critical points are solution to the equation $P_w \nabla r(w) = 0$. But r is constant on lines coming from 0 so for all $w \in \mathbb{R}^d \setminus \{0\} : \langle \nabla r(w), w \rangle_{\ell_2} = 0$. Consequently, $P_w \nabla r(w) = 0 \iff \nabla r(w) = 0$.

By the chain rule, $\nabla r(w) = 0$ if and only if

$$\begin{aligned} & [w^\top (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w] (\mu_1 - \mu_{-1}) (\mu_1 - \mu_{-1})^\top w \\ & - [w^\top (\mu_1 - \mu_{-1}) (\mu_1 - \mu_{-1})^\top w] (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w = 0, \end{aligned}$$

that is, when reorganizing,

$$[w^\top (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w (\mu_1 - \mu_{-1})^\top w] (\mu_1 - \mu_{-1}) = [(\mu_1 - \mu_{-1})^\top w]^2 (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1}) w,$$

where the terms in brackets in the left and right hand side are scalar. Assuming that w is not orthogonal to $\mu_1 - \mu_{-1}$ (that is the scalar in the right hand side is nonzero), this means that, in any case:

$$w \propto (\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1})^{-1} (\mu_1 - \mu_{-1}).$$

Recapping, an extremum of $r|_{\|w\|_{\ell_2}=1}$ is either orthogonal to $\mu_1 - \mu_{-1}$ or satisfies the previous relation with $\|w\|_{\ell_2} = 1$.

On the one hand, considering w orthogonal to $\mu_1 - \mu_{-1}$ leads to $r(w) = 0$ (which is the minimum of r). On the other hand, all nonzero vectors satisfying the previous relation have the same value of Rayleigh quotient, which is not null. Since $r|_{\|w\|_{\ell_2}=1}$ has a maximum, it comes that vectors proportional to $(\pi_1 \Sigma_1 + (1 - \pi_1) \Sigma_{-1})^{-1} (\mu_1 - \mu_{-1})$ with $\|w\|_{\ell_2} = 1$ are maxima.

Proof 2

Since Σ is non-singular, by the change of variable $u = \Sigma^{\frac{1}{2}} w$, the problem of interest becomes

$$\underset{u \in \mathbb{R}^d : u \neq 0}{\text{maximize}} \quad \frac{u^\top A u}{\|u\|_{\ell_2}^2},$$

where $A = \Sigma^{-\frac{1}{2}} M \Sigma^{-\frac{1}{2}}$. Considering the eigendecomposition $(\lambda_1 \geq \dots \geq \lambda_d)$, (v_1, \dots, v_d) of A and for all $u \in \mathbb{R}^d : u \neq 0$, its decomposition on the orthonormal basis (v_1, \dots, v_d) : $u = \sum_{i=1}^d a_i v_i$, we have

$$\frac{v_1^\top A v_1}{\|v_1\|_{\ell_2}^2} - \frac{u^\top A u}{\|u\|_{\ell_2}^2} = \lambda_1 - \frac{\sum_{i=1}^d a_i^2 \lambda_i}{\sum_{i=1}^d a_i^2} = \frac{\sum_{i=1}^d a_i^2 (\lambda_1 - \lambda_i)}{\sum_{i=1}^d a_i^2} \geq 0.$$

So $\frac{u^\top A u}{\|u\|_{\ell_2}^2}$ is maximized for $\text{span}(v_1) \setminus \{0\}$.

Moreover, with the change of variable $u = \Sigma^{\frac{1}{2}} w$, we have, for any $\lambda \geq 0$, $Au = \lambda u \iff \Sigma^{-1} M w = \lambda w$. In other words, the eigenvectors of $\Sigma^{-1} M$ are (w_1, \dots, w_d) with $w_i = \Sigma^{-\frac{1}{2}} v_i$. Moreover, for $\lambda > 0$

$$\Sigma^{-1} M w = \lambda w \iff \Sigma^{-1} (\mu_1 - \mu_{-1}) (\mu_1 - \mu_{-1})^\top w = \lambda w \iff w = \frac{(\mu_1 - \mu_{-1})^\top w}{\lambda} \Sigma^{-1} (\mu_1 - \mu_{-1}).$$

As a consequence and since $\lambda_1 > 0$, r is maximized by

$$\text{span}(w_1) \setminus \{0\} = \text{span}(\Sigma^{-1} (\mu_1 - \mu_{-1})) \setminus \{0\}.$$

□

Remark 1.1.3. When covariance matrices are equal, Fisher's discriminant direction is the same as that of LDA.

Projection of X on the direction w is given by:

$$h(X) = w^\top X.$$

In order to classify, we may apply different rules like assigning to the class of the nearest center or thresholding based on an intercept. Such an intercept b can be defined by:

$$b \in \arg \min_{a \in \mathbb{R}} \mathbb{P}(Y \neq g_a(X)),$$

where

$$g_a: x \in \mathbb{R}^d \mapsto \text{sign}(h(x) + a).$$

Let us remark that, in its empirical version (that is replacing expected values by their means computed with the sample $\{(X_i, Y_i)\}_{1 \leq i \leq n}$), an intercept can be defined by

$$b \in \arg \min_{a \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \neq g_a(X_i)},$$

where $a \in \mathbb{R} \mapsto \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \neq g_a(X_i)}$ is a piecewise constant function, for which the steps are at $\{-h(X_1), \dots, -h(X_n)\}$. This means that only n values have to be evaluated to determine an empirical threshold b .

1.1.4 Kernel Fisher discriminant analysis

Let $\{X_i\}_{1 \leq i \leq n} \subset \mathbb{R}^d$ be *iid* copies of X and $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ a kernel (see Section 1.4.2) with feature map $\phi: \mathbb{R}^d \rightarrow \mathcal{G}$, where \mathcal{G} is an appropriate Hilbert space (of dimension D , potentially infinite). As a reminder, we have $\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d: k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$.

We aim at applying the kernel trick (see Section 1.4.3) to Fisher's approach. For this purpose, let us consider the problem of Fisher's linear discriminant analysis for the random pair $(\phi(X), Y)$. Denoting, for each $i \in [C]$, $\pi_i = \mathbb{P}(Y = i)$, $\mu_i^\phi = \mathbb{E}(\phi(X)|Y = i)$, $\Sigma_i^\phi = \mathbb{V}(\phi(X)|Y = i)$, we have, for $w \in \mathcal{G}$,

$$r(w) = \pi_1(1 - \pi_1) \frac{\langle w, \mu_1^\phi - \mu_{-1}^\phi \rangle_{\mathcal{G}}^2}{\langle w, (\pi_1 \Sigma_1^\phi + (1 - \pi_1) \Sigma_{-1}^\phi) w \rangle_{\mathcal{G}}}.$$

Since \mathcal{G} may be infinite-dimensional, the Rayleigh quotient cannot be maximized numerically. However, in its empirical version, it involves the estimators $\hat{\mu}_1^\phi, \hat{\mu}_{-1}^\phi \in \text{span}(\{\phi(X_1), \dots, \phi(X_n)\})$. Thus, $\forall w \in \text{span}(\{\phi(X_1), \dots, \phi(X_n)\})^\perp$, $r(w) = 0$ and we can restrict the maximization of r to w in $\text{span}(\{\phi(X_1), \dots, \phi(X_n)\})$. In other words, we look for solutions w such that there exists $\alpha \in \mathbb{R}^n$ with $w = \sum_{i=1}^n \alpha_i \phi(X_i)$. Then, we get

$$r(w) = \pi_1(1 - \pi_1) \frac{(\alpha^\top (\nu_1 - \nu_{-1}))^2}{\alpha^\top (\pi_1 \Psi_1 + (1 - \pi_1) \Psi_{-1}) \alpha},$$

where for each $i \in \{1, 2\}$,

$$\nu_i = \left(\langle \mu_i^\phi, \phi(X_1) \rangle_{\mathcal{G}}, \dots, \langle \mu_i^\phi, \phi(X_n) \rangle_{\mathcal{G}} \right) \in \mathbb{R}^n$$

and

$$\Psi_i = \left(\langle \phi(X_l), \Sigma_i^\phi \phi(X_j) \rangle_{\mathcal{G}} \right)_{1 \leq l, j \leq n} \in \mathbb{R}^{n \times n}.$$

Let $\mathcal{I}_i = \{l \in [n] : Y_l = i\}$. Replacing μ_i^ϕ and Σ_i^ϕ by their estimates $\hat{\mu}_i^\phi = \frac{1}{|\mathcal{I}_i|} \sum_{\ell \in \mathcal{I}_i} \phi(X_\ell)$ and $\hat{\Sigma}_i^\phi$ gives a practical method for nonlinear discriminant analysis: on the first hand, for each $i \in \{1, 2\}$,

$$\hat{\nu}_i = \left(\frac{1}{|\mathcal{I}_i|} \sum_{l \in \mathcal{I}_i} k(X_l, X_1), \dots, \frac{1}{|\mathcal{I}_i|} \sum_{l \in \mathcal{I}_i} k(X_l, X_n) \right).$$

On the other hand, let \mathbf{X} be the sample matrix in the feature space \mathcal{G} :

$$\mathbf{X} = [\phi(X_1) | \dots | \phi(X_n)]^\top \in \mathbb{R}^{n \times D}.$$

Then, the matrix of centered data is

$$\mathbf{Z} = \mathbf{X} - \left[\frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) | \dots | \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right]^\top = \mathbf{X} - \mathbf{1} \left(\frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right)^\top = (I_n - M) \mathbf{X} = H_n \mathbf{X},$$

where I_n is the identity matrix of size n , $M = \mathbf{1}\mathbf{1}^\top/n \in \mathbb{R}^{n \times n}$, $\mathbf{1}$ is the all-ones vector of adequate size and $H_n = I_n - M$.

Let, for all $i \in \{1, 2\}$, \mathbf{X}_i be the submatrix of \mathbf{X} containing only the rows indexed by \mathcal{I}_i , and

$$\mathbf{Z}_i = H_{|\mathcal{I}_i|} \mathbf{X}_i,$$

the matrix of the centered data from class i . Then,

$$\hat{\Sigma}_i^\phi = \mathbf{Z}_i^\top \mathbf{Z}_i = \mathbf{X}_i^\top H_{|\mathcal{I}_i|}^2 \mathbf{X}_i = \mathbf{X}_i^\top H_{|\mathcal{I}_i|} \mathbf{X}_i.$$

Moreover, we easily see that $\Psi_i = \mathbf{X} \hat{\Sigma}_i^\phi \mathbf{X}^\top$, which leads to

$$\hat{\Psi}_i = \mathbf{X} \hat{\Sigma}_i^\phi \mathbf{X}^\top = \mathbf{X} \mathbf{X}_i^\top H_{|\mathcal{I}_i|} \mathbf{X}_i \mathbf{X}^\top = K_i H_{|\mathcal{I}_i|} K_i^\top,$$

where $K_i = \mathbf{X} \mathbf{X}_i^\top \in \mathbb{R}^{n \times |\mathcal{I}_i|}$ is also defined by $K_i = (k(X_j, X_l))_{\substack{1 \leq j \leq n \\ 1 \leq l \leq |\mathcal{I}_i|}}$.

Similarly to previously, solutions are given by

$$\alpha \propto (\pi_1 \hat{\Psi}_1 + (1 - \pi_1) \hat{\Psi}_{-1})^{-1} (\hat{\mathbf{v}}_1 - \hat{\mathbf{v}}_{-1}).$$

In addition, knowing the direction $w = \sum_{i=1}^n \alpha_i \phi(X_i)$, projection of X can be computed by:

$$h(X) = \langle w, \phi(X) \rangle_{\mathcal{G}} = \sum_{i=1}^n \alpha_i k(X, X_i).$$

1.1.5 Multiclass linear discriminant

Let us denote $\mu = \mathbb{E} X$, $\mu_i = \mathbb{E}(X|Y = i)$ and $\Sigma_i = \mathbb{V}(X|Y = i)$ for each $i \in [C]$, and $\Sigma = \sum_{i=1}^C \pi \Sigma_i$. Then, the Rayleigh quotient reads:

$$r(w) = \frac{w^\top M w}{w^\top \Sigma w},$$

where $M = \sum_{i=1}^C \pi_i (\mu_i - \mu)(\mu_i - \mu)^\top$ is *a priori* a rank- $(C - 1)$ matrix of size $d \times d$.

Similarly to previously, it appears that if w maximizes r , then w is an eigenvector of $\Sigma^{-1} M$ and then the Rayleigh quotient equals the corresponding eigenvalue. Since Σ^{-1} is a rank- d matrix, if M is a rank- $(C - 1)$ matrix, then $\Sigma^{-1} M$ is a rank- $(C - 1)$ matrix. Thus, the $(C - 1)$ leading eigenvectors of $\Sigma^{-1} M$, denoted (w_1, \dots, w_{C-1}) (with non-increasing eigenvalues), concentrate the variability between features.

At this step, if for $C = 2$, it is sufficient to find an intercept to separate the data, it is more complicated for multiclass problems. The idea is thus to apply a simple classifier in the feature space described by eigenvectors (w_1, \dots, w_{C-1}) : let $P \in \mathbb{R}^{(C-1) \times d}$ be the row matrix of normalized eigenvectors $w_i / \|w_i\|_{\ell_2}$. One can choose the classifier given by:

$$g(X) \in \arg \min_{i \in [C]} \|PX - P\mu_i\|_{\ell_2}.$$

1.2 Logistic regression

1.2.1 Model and risk

Similarly to LDA, let us consider normally distributed classes with equal variances:

$$\forall i \in [C]: \quad X|Y = i \sim \mathcal{N}(\mu_i, \Sigma).$$

Then, for each class $i \in [C]$, the log posterior ratio is given by:

$$\forall x \in \mathbb{R}^d: \quad \log \left(\frac{\mathbb{P}(Y = i|X = x)}{\mathbb{P}(Y = C|X = x)} \right) = w_i^\top x + b_i,$$

where

$$\begin{aligned} w_i &= \Sigma^{-1}(\mu_i - \mu_C) \\ b_i &= \log \left(\frac{\pi_i}{\pi_C} \right) - \frac{1}{2} \log \left(\frac{|\Sigma_i|}{|\Sigma_C|} \right) + \frac{1}{2} \mu_C^\top \Sigma^{-1} \mu_C - \frac{1}{2} \mu_i^\top \Sigma^{-1} \mu_i. \end{aligned}$$

This linear form of the log ratio (also called log-odds or logit transformations) results from Gaussian assumption but motivates, in a more general framework, to model the log ratio as a linear function of x . Thus, without any other assumption, logistic regression assumes that, for each class $i \in [C - 1]$, there exists $(b_i^*, w_i^*) \in \mathbb{R} \times \mathbb{R}^d$ such that:

$$\forall x \in \mathbb{R}^d: \quad \log \left(\frac{\mathbb{P}(Y = i|X = x)}{\mathbb{P}(Y = C|X = x)} \right) = (w_i^*)^\top x + b_i^*.$$

In particular, for $C = 2$, it is assumed that there exists $(b^*, w^*) \in \mathbb{R} \times \mathbb{R}^d$ such that:

$$\forall x \in \mathbb{R}^d: \quad \log \left(\frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = -1|X = x)} \right) = (w^*)^\top x + b^*.$$

Remark 1.2.1 (Hypothesis on the Bayes classifier). *The point of view adopted here is that, contrarily to LDA, logistic regression does not directly make an assumption on the data distribution, but on the Bayes classifier: for two classes ($C = 2$), the Bayes classifier is assumed to be linear (i.e. the decision frontier is a hyperplane). By simplicity, the common decision function $f : x \in \mathbb{R}^d \mapsto \log \left(\frac{\mathbb{P}(Y=1|X=x)}{\mathbb{P}(Y=-1|X=x)} \right)$ is assumed to be affine. The forthcoming derivation exhibits that this results in making an assumption on the distribution of $Y|X = x$.*

Example 1.2.1. *A motivating example is the case where $X \in \mathbb{R}^2$ and $X|Y$ has density $x \in \mathbb{R}^2 \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_1 - \mu_Y)^2}{2}} \mathbb{1}_{[0,1]}(x_2)$, i.e. the first coordinate of $X|Y$ is Gaussian and the second is independent from the first and uniform on $[0, 1]$. In this case, it is easy to see that the decision function f is linear but $X|Y$ is definitely not Gaussian.*

Now, optimal parameters w^*, b^* have to be estimated. For this purpose, we resort to empirical risk minimization based on the following result.

Theorem 8. *Let us consider that $C = 2$ and that the logit-transformation is affine with parameters (b^*, w^*) . Let $f^*: x \in \mathbb{R}^d \mapsto (w^*)^\top x + b^*$.*

Assuming that $X \in L^1$, then f^ is a minimizer of the risk functional $f \mapsto \mathbb{E}[\log(1 + \exp(-Yf(X)))]$ over all affine functions and*

$$g^*: x \in \mathbb{R}^d \mapsto \text{sign}(f(x))$$

is a Bayes classifier.

Proof. First, let us remark that g^* , as previously defined, is a Bayes classifier, since $f^*(x) > 0$ if and only if $\mathbb{P}(Y = +1|X = x) > \mathbb{P}(Y = -1|X = x)$.

Second, by summation to 1, we have:

$$\forall (x, y) \in \mathbb{R}^d \times \{\pm 1\} : \quad \mathbb{P}(Y = y|X = x) = \frac{1}{1 + \exp(-y((w^*)^\top x + b^*))}.$$

Third, without loss of generality, we consider that b vanishes and we denote

$$\psi: (x, y, w) \mapsto \log(1 + \exp(-yw^\top x)).$$

Then, $\forall w \in \mathbb{R}^d$:

- ◇ $0 \leq \mathbb{E}[\psi(X, Y, w)] \leq \mathbb{E}[| -Yw^\top X |] + 1 = \mathbb{E}[|w^\top X|] + 1 < +\infty$ since $X \in L^1$;
- ◇ ψ is differentiable in w ;
- ◇ $\nabla_w \psi(X, Y, w) = -\frac{Y \exp(-Yw^\top X)}{1 + \exp(-Yw^\top X)} X$ and $\|\nabla_w \psi(X, Y, w)\|_{\ell_2} \leq \left\| \frac{\exp(-Yw^\top X)}{1 + \exp(-Yw^\top X)} X \right\|_{\ell_2} \leq \|X\|_{\ell_2} \in L^1$.

Thus, by Leibniz integral rule, $\nabla_w \mathbb{E}[\psi(X, Y, \cdot)](w) = \mathbb{E}[\nabla_w \psi(X, Y, w)] = \mathbb{E}[\mathbb{E}[\nabla_w \psi(X, Y, w)|X]]$.

Moreover,

$$\begin{aligned} & \mathbb{E}[\nabla_w \psi(X, Y, w)|X] \\ &= \mathbb{P}(Y = +1|X) \nabla_w \psi(X, 1, w) + \mathbb{P}(Y = -1|X) \nabla_w \psi(X, -1, w) \\ &= \frac{1}{1 + \exp(-(w^*)^\top X)} \left(-\frac{\exp(-w^\top X)}{1 + \exp(-w^\top X)} X \right) + \frac{1}{1 + \exp((w^*)^\top X)} \left(\frac{\exp(w^\top X)}{1 + \exp(w^\top X)} X \right) \\ &= \frac{1}{1 + \exp(-(w^*)^\top X)} \left(-\frac{\exp(-w^\top X)}{1 + \exp(-w^\top X)} X \right) + \frac{\exp(-(w^*)^\top X)}{1 + \exp(-(w^*)^\top X)} \left(\frac{1}{1 + \exp(-w^\top X)} X \right). \end{aligned}$$

Thus, $\mathbb{E}[\nabla_w \psi(X, Y, w^*)|X] = 0$ and $\nabla_w \mathbb{E}[\psi(X, Y, \cdot)](w^*) = 0$. Since ψ is convex in w , this proves that w^* is a minimizer of $\mathbb{E}(\psi(X, Y, \cdot))$. \square

Let $\{(X_i, Y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \{\pm 1\}$ be an *iid* sample distributed as (X, Y) . Theorem 8 illustrates that parameters of logistic regression can be estimated by minimizing an empirical risk defined by the logistic loss:

$$R_n(w, b) = \frac{1}{n} \sum_{i=1}^n \log \left(1 + e^{-Y_i(w^\top X_i + b)} \right)$$

with respect to the hyperplan parameters $(w, b) \in \mathbb{R}^d \times \mathbb{R}$. This is a new exemple of loss function for classification, to be compared to the exponential and the hinge functions (see Figure 1.4).

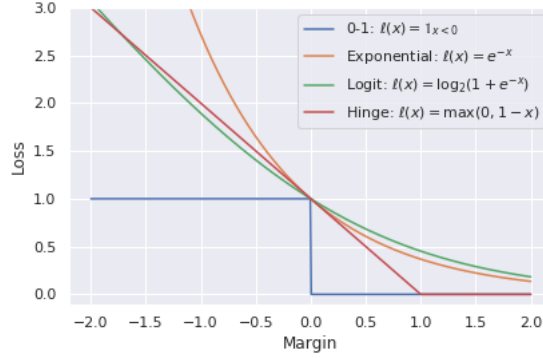


Figure 1.4: Example of convex losses.

Remark 1.2.2. *Logistic regression can also be seen as a latent-variable model: consider the latent variable $Z = f^*(X) + \epsilon$, where $\epsilon \sim \mathcal{L}(0, 1)$ is a random variable independent from X , and set $Y = \text{sign}(Z)$. Then, for any $x \in \mathbb{R}^d$, we have:*

$$\begin{aligned} p &= \mathbb{P}(Y = 1 | X = x) \\ &= \mathbb{P}(f^*(X) + \epsilon > 0 | X = x) \\ &= \mathbb{P}(-\epsilon < f^*(x)) && \text{(by independence)} \\ &= \mathbb{P}(\epsilon < f^*(x)) && \text{(by symmetry of } \mathcal{L}(0, 1)) \\ &= \frac{1}{1 + e^{-f^*(x)}} && \text{(by definition).} \end{aligned}$$

As a result, $\log \left(\frac{p}{1-p} \right) = f^*(x)$, which is the assumption of logistic regression.

1.2.2 Maximum likelihood estimation

The use of the logistic loss in the logistic risk $R(w, b) = \mathbb{E} \left[\log \left(1 + e^{-Y(w^\top X + b)} \right) \right]$ is consistent when thinking to maximum likelihood estimation of (w^*, b^*) . Let $f_{(X, Y)}$ et f_X be respectively the joint density of (X, Y) and the marginal density of X . Since, with a slight abuse of notation, $\mathbb{P}(Y|X) = \frac{1}{1 + \exp(-Y(b + w^\top X))}$, the full log-likelihood of any $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ is:

$$\log(f_{(X, Y)}(X, Y)) = \log(\mathbb{P}(Y|X)) + \log(f_X(X)) = -\log \left(1 + e^{-Y(w^\top X + b)} \right) + \log(f_X(X)),$$

and the conditional log-likelihood (i.e. in the statistical model associated to $Y|X$) is:

$$\log(\mathbb{P}(Y|X)) = -\log \left(1 + e^{-Y(w^\top X + b)} \right).$$

Given our assumption, we do not know f_X , which motivates considering the conditional log-likelihood instead of the full log-likelihood in order to estimate w^* , b^* . Up to the sign, the conditional log-likelihood is exactly the term under the expectation in the logistic risk $R(w, b)$. Going to estimation, it becomes clear that the empirical conditional log-likelihood of any $(w, b) \in \mathbb{R}^d \times \mathbb{R}$ is linked to the empirical logistic risk:

$$\log \left(\prod_{i=1}^n \mathbb{P}(Y_i|X_i) \right) = -\sum_{i=1}^n \log \left(1 + e^{-Y_i(w^\top X_i + b)} \right) = -nR_n(w, b).$$

This point is a big difference between LDA and logistic regression: LDA fits the parameters by maximizing the full log-likelihood

$$\log(f_{(X,Y)}(X, Y)) = \log(\mathbb{P}(Y|X)) + \log(f_X(X)),$$

while logistic regression leaves the marginal density of X aside and maximizes the conditional log-likelihood. In some sense, the marginal likelihood can be thought of as a regularizer.

Remark 1.2.3. *If the dataset in a two-class logistic regression model is linearly separable, the maximum likelihood estimates of the parameters are undefined (i.e., infinite): let $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ be an iid sample according to (X, Y) such that:*

$$\exists (w_0, b_0) \in \mathbb{R}^d \times \mathbb{R} : \forall i \in [n], Y_i(w_0^\top X_i + b_0) > 0.$$

Then, $\varphi: \lambda \in \mathbb{R} \mapsto F(\lambda w_0, \lambda b_0) = \frac{1}{n} \sum_{i=1}^n \log \left(1 + e^{-\lambda[Y_i(w_0^\top X_i + b_0)]} \right)$ is decreasing and converges to 0. In addition, for any $(w, b) \in \mathbb{R}^d \times \mathbb{R}$, $F(w, b) > \frac{F(w, b)}{2} > 0$. Since φ is decreasing and converges to 0, we can find $\bar{\lambda} \in \mathbb{R}$ such that $\frac{F(w, b)}{2} \geq \varphi(\bar{\lambda}) = F(\bar{\lambda} w_0, \bar{\lambda} b_0)$, so $F(w, b) > F(\bar{\lambda} w_0, \bar{\lambda} b_0)$. We conclude that there is no solution.

However, the LDA coefficients for the same data will be well defined.

As a result of this remark and in order to enhance the generalization properties of logistic regression, it is common to estimate (w^*, b^*) by minimization of a regularized empirical risk (or negative log-likelihood):

$$\underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{n} \sum_{i=1}^n \log \left(1 + e^{-Y_i(w^\top X_i + b)} \right) + \frac{\lambda}{2} \|w\|_{\ell_2}^2,$$

where $\lambda > 0$ be a regularization parameter. It is easy to see that the ℓ_2 regularization on w helps solving both caveats of vanilla logistic regression.

1.2.3 Logistic regression versus LDA

Besides the maximum likelihood difference enlightened in the previous section, we give here some empirical conclusions borrowed from [Hastie et al. \[2013\]](#).

Power of logistic regression

If in fact the classes are Gaussian, then in the worst case ignoring this marginal part of the likelihood constitutes a loss of efficiency of about 30% asymptotically in the error rate. Paraphrasing: with 30% more data, the conditional likelihood will do as well.

Outliers

Observations far from the decision boundary are down-weighted by logistic regression while they play a role in estimating the common covariance matrix. It means that LDA is not robust to gross outliers (see Figure 1.6).

In practice

In practice the normal assumption is never correct, and often some covariates are qualitative. It is generally felt that logistic regression is a safer, more robust bet than the LDA model, relying on fewer assumptions (see Figure 1.5).

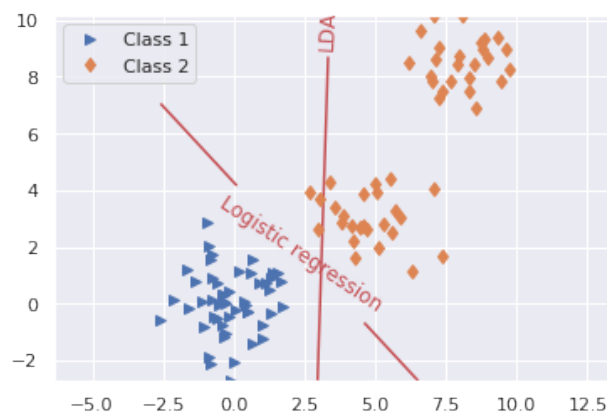


Figure 1.5: Comparison of logistic regression and LDA with non-Gaussian classes.



Figure 1.6: Comparison of logistic regression and LDA with a single outlier.

1.3 Boosting

1.3.1 Adaboost

Adaboost (and boosting in a more general way) was designed to expand the expressiveness of linear predictors by composing them on top of other functions. More formally, it came up to answer a novel theoretical question, that of designing a strong learning algorithm using a weak learning one.

The boosting approach addresses two major issues raised against linear predictors:

1. the bias-complexity tradeoff: the error of an ERM learner can be decomposed into an approximation error and an estimation error (see Figure 1.7). The more expressive the hypothesis class the learner is searching over, the smaller the approximation error is, but the larger the estimation error becomes. In the boosting paradigm, the learning starts with a basic class (that might have a large approximation error), and as it progresses, the class that the predictor may belong to grows richer. This procedure allows to have a smooth control of the tradeoff between approximation and estimation errors.
2. computational complexity of learning: boosting is very cheap, particularly with decision stumps.

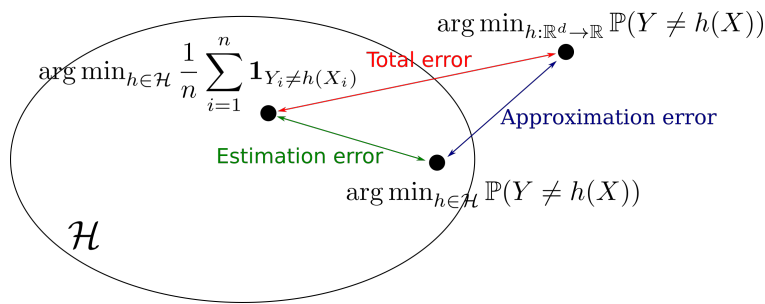


Figure 1.7: Types of errors in the empirical risk minimization process.

Let \mathcal{C} be a class of $\{\pm 1\}$ -classifiers. In the forthcoming paragraphs, we describe Adaboost (Algorithm 1) and its generalization ability. As an appetizer, the iteration of Adaboost can be wrapped up in the following three steps (detailed in Algorithm 1):

- ◇ find a classifier $g_t \in \mathcal{C}$ with small weighted error;
- ◇ weight g_t with w_t such that $f_t = f_{t-1} + w_t g_t$ has a small empirical risk;
- ◇ update the point weights according to how they are recognized by f_t .

Property 9. In Algorithm 1, we have for each iteration $t \in [T]$:

- ◇ for all $i \in [n]$, $D_{t+1}(i) = \frac{\exp(-Y_i f_t(X_i))}{n \prod_{j=1}^t Z_j}$;
- ◇ $w_t = \frac{1}{2} \log \left(\frac{1-\epsilon_t}{\epsilon_t} \right)$;
- ◇ $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$.

Proof. First, let us express the adaptive empirical distribution. By induction, we have for all $i \in [n]$ and

Algorithm 1 Adaboost.

Input: $T \in \mathbb{N}$ (number of iterations), $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ (training sample).

for $i = 1$ **to** n **do**

$$D_1(i) \leftarrow \frac{1}{n}$$

end for

$f_0 = 0$ (null function)

for $t = 1$ **to** T **do**

$g_t \leftarrow$ base $\{\pm 1\}$ -classifier from \mathcal{C} with small error $\epsilon_t = \sum_{i=1}^n D_t(i) \mathbf{1}_{Y_i \neq g_t(X_i)}$

$$w_t \leftarrow \arg \min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \exp(-Y_i(f_{t-1}(X_i) + w g_t(X_i))) = \frac{1}{2} \log \left(\frac{1-\epsilon_t}{\epsilon_t} \right) \text{ (ERM)}$$

$$Z_t \leftarrow \sum_{i=1}^n D_t(i) \exp(-w_t Y_i g_t(X_i)) = 2\sqrt{\epsilon_t(1-\epsilon_t)} \text{ (normalization)}$$

for $i = 1$ **to** n **do**

$$D_{t+1}(i) \leftarrow D_t(i) \exp(-w_t Y_i g_t(X_i)) / Z_t$$

end for

$$f_t = \sum_{j=1}^t w_j g_j$$

end for

Output: $g_n^T = \text{sign}(f_T)$.

$t \in [T]$:

$$\begin{aligned} D_{t+1}(i) &= \frac{1}{n} \prod_{j=1}^t \exp(-w_j Y_i g_j(X_i)) / Z_j \\ &= \frac{1}{n} \frac{\exp\left(-Y_i \sum_{j=1}^t w_j g_j(X_i)\right)}{\prod_{j=1}^t Z_j} \\ &= \frac{\exp(-Y_i f_t(X_i))}{n \prod_{j=1}^t Z_j}. \end{aligned}$$

Then, w_t is the minimizer of a strictly convex function. It is achieved when the derivative vanishes, that is:

$$\begin{aligned} 0 &= \sum_{i=1}^n Y_i g_t(X_i) e^{-Y_i(f_{t-1}(X_i) + w_t g_t(X_i))} \\ &= \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = 1}} e^{-Y_i f_{t-1}(X_i)} e^{-w_t} - \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = -1}} e^{-Y_i f_{t-1}(X_i)} e^{w_t} \\ &= \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = 1}} e^{-Y_i f_{t-1}(X_i)} - e^{2w_t} \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = -1}} e^{-Y_i f_{t-1}(X_i)}. \end{aligned}$$

Rearranging, we have:

$$\begin{aligned}
w_t &= \frac{1}{2} \log \left(\frac{\sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = 1}} e^{-Y_i f_{t-1}(X_i)}}{\sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = -1}} e^{-Y_i f_{t-1}(X_i)}} \right) \\
&= \frac{1}{2} \log \left(\frac{\sum_{i=1}^n e^{-Y_i f_{t-1}(X_i)} \mathbf{1}_{Y_i = g_t(X_i)}}{\sum_{i=1}^n e^{-Y_i f_{t-1}(X_i)} \mathbf{1}_{Y_i \neq g_t(X_i)}} \right) \\
&= \frac{1}{2} \log \left(\frac{\sum_{i=1}^n n \prod_{j=1}^{t-1} Z_j D_t(i) \mathbf{1}_{Y_i = g_t(X_i)}}{\sum_{i=1}^n n \prod_{j=1}^{t-1} Z_j D_t(i) \mathbf{1}_{Y_i \neq g_t(X_i)}} \right) \\
&= \frac{1}{2} \log \left(\frac{\sum_{i=1}^n D_t(i) \mathbf{1}_{Y_i = g_t(X_i)}}{\sum_{i=1}^n D_t(i) \mathbf{1}_{Y_i \neq g_t(X_i)}} \right) \\
&= .
\end{aligned}$$

In addition, for all $t \in [T]$, the normalization factor is:

$$\begin{aligned}
Z_t &= \sum_{i=1}^n D_t(i) \exp(-w_t Y_i g_t(X_i)) \\
&= \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = 1}} D_t(i) \exp(-w_t) + \sum_{\substack{1 \leq i \leq n \\ Y_i g_t(X_i) = -1}} D_t(i) \exp(w_t) \\
&= (1 - \epsilon_t) \exp(-w_t) + \epsilon_t \exp(w_t) \\
&= (1 - \epsilon_t) \sqrt{\frac{\epsilon_t}{1 - \epsilon_t}} + \epsilon_t \sqrt{\frac{1 - \epsilon_t}{\epsilon_t}} \\
&= 2\sqrt{\epsilon_t(1 - \epsilon_t)}.
\end{aligned}$$

□

Theorem 10. Assume that there exists $\gamma > 0$ such that $\forall t \in [T]$, $\epsilon_t \leq \frac{1}{2} - \gamma$ almost surely and let $g_n^T: \mathcal{X} \rightarrow \{\pm 1\}$ be the classifier returned by Adaboost. Then,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \neq g_n^T(X_i)} \leq \exp(-2\gamma^2 T).$$

Proof. Since we have for all $x \in \mathbb{R}$ $\mathbf{1}_{x \leq 0} \leq \exp(-x)$, we can write

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \neq \text{sign } f_T(X_i)} &\leq \frac{1}{n} \sum_{i=1}^n \exp(-Y_i f_T(X_i)) \\
&= \frac{1}{n} \sum_{i=1}^n \left(\prod_{t=1}^T Z_t \right) D_{T+1}(i) \\
&= \prod_{t=1}^T Z_t \quad \left(\sum_{i=1}^n D_{T+1}(i) = 1 \right) \\
&= \prod_{t=1}^T 2\sqrt{\epsilon_t(1-\epsilon_t)} \\
&\leq \prod_{t=1}^T 2\sqrt{\left(\frac{1}{2} - \gamma\right) \left(1 - \left(\frac{1}{2} - \gamma\right)\right)} \quad \left(\text{we want to sharpen the bound } \epsilon_t(1-\epsilon_t) \leq \frac{1}{4} \right) \\
&= \prod_{t=1}^T \sqrt{1 - 4\gamma^2} \\
&\leq \prod_{t=1}^T \exp(-4\gamma^2)^{\frac{1}{2}} \quad \left(\text{since } 1 - x \leq \exp(-x), \forall x \in \mathbb{R} \right) \\
&\leq \exp(-2\gamma^2 T).
\end{aligned}$$

□

Definition 1.3.1. Let \mathcal{F} be a class of function from \mathbb{R}^d to \mathbb{R} and (Z_1, \dots, Z_n) be an iid sample of random vectors from \mathbb{R}^d . Let also $(\sigma_1, \dots, \sigma_n)$ be iid Rademacher random variables, independent from (Z_1, \dots, Z_n) .

The Rademacher complexity of \mathcal{F} is

$$R_n(\mathcal{F}(Z_1^n)) = \mathbb{E} \left(\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \sigma_i f(Z_i) \right| \mid Z_1, \dots, Z_n \right).$$

Let us consider $\mathcal{F} = \{f = \sum_{j=1}^T w_j g_j : T \in \mathbb{N}, (g_1, \dots, g_T) \in \mathcal{C}^T, \|w\|_{\ell_1} = 1\}$ be the class of hypotheses.

Theorem 11. Let $\gamma > 0$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$,

$$\forall f \in \mathcal{F}: \quad \mathbb{P}(Y \neq \text{sign}(f(X))) \leq \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i f(X_i) < \gamma} + \frac{4}{\gamma} R_n(\mathcal{C}(X_1^n)) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Refer to Gérard Biau's class (Slides 45 to 47) or [Mohri et al., 2012, Theorem 4.4]. \square

Lemma 12. Assume that there exists $\gamma \in (0, 1/2)$ such that $\forall t \in [T]$, $\epsilon_t \leq \frac{1}{2} - \gamma$ almost surely and let $f_T: \mathcal{X} \rightarrow \mathbb{R}$ be the Adaboost classifier at the last iteration. Then,

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{\gamma_i \frac{f_T(X_i)}{\|w\|_{\ell_1}} < \gamma} \leq \left((1 - 4\gamma^2) \left(\frac{1 + 2\gamma}{1 - 2\gamma} \right)^\gamma \right)^{T/2}.$$

In addition, we have $(1 - 4\gamma^2) \left(\frac{1 + 2\gamma}{1 - 2\gamma} \right)^\gamma < 1$.

The proof is a good exercise.

Theorem 13. Assume that there exists $\gamma \in (0, 1/2)$ such that $\forall t \in [T]$, $\epsilon_t \leq \frac{1}{2} - \gamma$ almost surely and let $g_n^T: \mathcal{X} \rightarrow \{\pm 1\}$ be the classifier returned by Adaboost. Let also $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$,

$$\mathbb{P}(Y \neq g_n^T(X) | X_1, \dots, X_n) \leq \left((1 - 4\gamma^2) \left(\frac{1 + 2\gamma}{1 - 2\gamma} \right)^\gamma \right)^{T/2} + \frac{4}{\gamma} R_n(\mathcal{C}(X_1^n)) + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. It comes from applying Theorem 11 with $\mathcal{F} = \{f = \sum_{j=1}^T \frac{w_j}{\|w\|_{\ell_1}} g_j : (g_1, \dots, g_T) \in \mathcal{C}^T, w \in \mathbb{R}^d\}$ \square

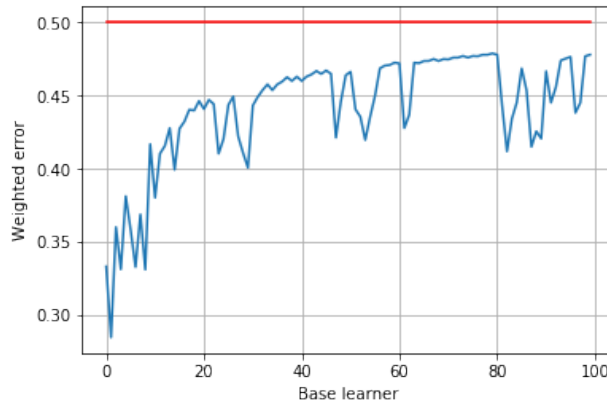


Figure 1.8: Weighted error of each weak learner. As expected, it tends to 0.5 since Adaboost focuses on hard examples.

1.3.2 ERM point of view and remarks

ERM

We have seen in Algorithm 1 that for all $t \in [T]$, each weight w_t is obtained by the rule

$$w_t \leftarrow \arg \min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \exp \left(-Y_i \left(\sum_{j=1}^{t-1} w_j g_j(X_i) + w g_t(X_i) \right) \right).$$

Defining $\phi: x \in \mathbb{R}^d \mapsto (g_1(x), \dots, g_T(x)) \in \{\pm 1\}^T$, this update rule can be seen as a coordinate descent for the empirical risk

$$w \in \mathbb{R}^T \mapsto \frac{1}{n} \sum_{i=1}^n \exp \left(-Y_i w^\top \phi(X_i) \right).$$

In that sens, Adaboost performs a coordinate descent on the convex risk previously defined, and learns a linear model penalized by the exponential loss. Using the logistic loss $x \in \mathbb{R} \mapsto \log(1 + e^{-x})$ instead leads to something very close to logistic regression.

Weak learners

In practice, it is common to use *stumps* as base classifiers, that is, decision trees of depth one. Thus, at each iteration, decision stumps quantize the most discriminative coordinate in ± 1 . The coordinate is then weighted by minimization of the exponential empirical risk.

Noise

It has been shown empirically that noise severely damages the performance of Adaboost. That is the most serious disadvantage of boosting.

In practice, we observe that the examples that are harder to classify end up dominating the selection of the base classifiers, which play a detrimental role in the definition of the final classifier.

Multiclass classification

Even though Adaboost with trees has been awarded with the “best off-the-shelf classifier in the world” title for binary classification problems, its natural extension to multiclass problems ($C > 2$) turns out to be very poor.

For this reason, an efficient extension of Adaboost has been proposed for multiclass problems, called Stagewise Additive Modeling using a Multiclass Exponential loss (SAMME). In accordance with its name, SAMME uses a multi-class exponential loss to compute the weights w_t :

$$f \mapsto \frac{1}{n} \sum_{i=1}^n \exp \left(-\frac{\tilde{Y}_i^\top f(X_i)}{C} \right),$$

where \tilde{Y}_i has 1 in its Y_i^{th} component and $-\frac{1}{C-1}$ otherwise, and $f: \mathcal{X} \rightarrow \mathbb{R}^C$ is a vector-valued decision function, each component of which corresponding to a class. At iteration t , a weak classifier $g_t: \mathcal{X} \rightarrow$

\mathbb{R}^C is learned, such that $g_t(x)$ does have the form of the coding vectors \tilde{Y}_i (that is, components are either 1 or $-\frac{1}{C-1}$) and $f_t = \sum_{j=1}^t w_j g_j$ is a function from \mathcal{X} to \mathbb{R}^C . The new updates are $w_t \leftarrow \frac{(C-1)^2}{C} \left(\log \left(\frac{1-\epsilon_t}{\epsilon_t} \right) + \log(C-1) \right)$ and $D_{t+1}(i) = D_t(i) \exp \left(-\frac{w_t \tilde{Y}_i^\top g_t(X_i)}{C} \right)$. Let us remark that this update is consistent with the situation in which $C = 2$. Besides this difference, weak classifiers are required to have an error better than random guessing, that is $\epsilon_t < \frac{C-1}{C}$. Moreover, the final decision rule becomes $g_n^T(x) = \arg \max_{1 \leq j \leq C} (f_T(x))_j = \arg \max_{1 \leq j \leq C} \sum_{t=1}^T w_t \mathbf{1}_{(g_t(x))_j=1}$. This last inequality can be seen by remarking that for any $1 \leq j \neq k \leq C$:

$$(f_T(x))_k - (f_T(x))_j = \sum_{t=1}^T w_t \left[(g_t(x))_k - (g_t(x))_j \right],$$

where

$$(g_t(x))_k - (g_t(x))_j = \begin{cases} 1 + \frac{1}{C-1} & \text{if } (g_t(x))_k = 1 \\ -\frac{1}{C-1} + \frac{1}{C-1} & \text{if } (g_t(x))_k \neq 1 \text{ and } (g_t(x))_j \neq 1 \\ -\frac{1}{C-1} - 1 & \text{if } (g_t(x))_j = 1. \end{cases}$$

Thus

$$(f_T(x))_k - (f_T(x))_j = \left(1 + \frac{1}{C-1} \right) \sum_{t=1}^T w_t \mathbf{1}_{(g_t(x))_k=1} - \left(1 + \frac{1}{C-1} \right) \sum_{t=1}^T w_t \mathbf{1}_{(g_t(x))_j=1},$$

and

$$(f_T(x))_k - (f_T(x))_j > 0 \iff \sum_{t=1}^T w_t \mathbf{1}_{(g_t(x))_k=1} > \sum_{t=1}^T w_t \mathbf{1}_{(g_t(x))_j=1}.$$

In addition, SAMME comes with a variant of it, called SAMME.R (R for real), which makes use of class probability estimates instead of classifiers g_t . SAMME.R is generally even more efficient than SAMME with respect to the classification accuracy.

1.3.3 Gradient boosting

Let $F: \mathbb{R}^n \rightarrow \mathbb{R}$ be a real-valued function, that is bounded from below. It is known that under some assumptions (convexity, differentiability of F and Lipschitz continuity of ∇F), the sequence defined by any $x_0 \in \mathbb{R}^n$ and for all positive integer t by:

$$x_t = x_{t-1} - w_t \nabla F(x_{t-1}),$$

where

$$w_t \in \arg \min_{w \in \mathbb{R}} F(x_{t-1} - w \nabla F(x_{t-1})),$$

converges to a minimizer of F . This is called *gradient descent with exact line search*. This procedure can be wrapped-up in three steps:

1. finding a direction of descent (here, $\nabla F(x_{t-1})$);
2. computing a step of descent w_t (minimizing $F(x_{t-1} - w \nabla F(x_{t-1}))$ with respect to $w \in \mathbb{R}$);

3. updating the optimization variable $x_t = x_{t-1} - w_t \nabla F(x_{t-1})$.

Gradient boosting occurred from the similarity between Adaboost and gradient descent. To observe it, let us remark that at each step $t > 0$ of Adaboost,

$$f_t = \sum_{j=1}^t w_j g_j = f_{t-1} + w_t g_t,$$

where

$$w_t \in \arg \min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \exp(-Y_i(f_{t-1}(X_i) + w g_t(X_i))).$$

This line search is point-wise in the sense that it only depends on the evaluations of f_{t-1} and g_t at $\{X_1, \dots, X_n\}$. Thus, let

$$F : x \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^n e^{-Y_i x_i}$$

be the empirical risk minimized in Algorithm 1 and consider the notation

$$x_t = (f_t(X_1), \dots, f_t(X_n)) \in \mathbb{R}^n \quad \text{and} \quad d_t = (-g_t(X_1), \dots, -g_t(X_n)) \in \mathbb{R}^n.$$

Then, the line search reads:

$$w_t \in \arg \min_{w \in \mathbb{R}} F(x_{t-1} - w d_t),$$

and at each iteration t , g_t (or equivalently d_t) is learned so as to minimize the weighted error

$$\begin{aligned} \sum_{i=1}^n D_t(i) \mathbf{1}_{Y_i \neq g_t(X_i)} &\propto \frac{2}{n} \sum_{i=1}^n e^{-Y_i f_{t-1}(X_i)} \mathbf{1}_{Y_i \neq g_t(X_i)} \\ &= \frac{2}{n} \sum_{i=1}^n e^{-Y_i f_{t-1}(X_i)} \frac{1 - Y_i g_t(X_i)}{2} \\ &= \frac{1}{n} \sum_{i=1}^n e^{-Y_i f_{t-1}(X_i)} - \sum_{i=1}^n \left(\frac{-Y_i e^{-Y_i f_{t-1}(X_i)}}{n} \right) (-g_t(X_i)) \\ &= F(x_{t-1}) - \langle \nabla F(x_{t-1}), d_t \rangle_{\ell_2}, \end{aligned}$$

that is so as to maximize $\langle \nabla F(x_{t-1}), d_t \rangle_{\ell_2}$. In other words, Adaboost learns at each iteration a base classifier, that is close to the gradient of F (in the correlation sense).¹

Consequently, Adaboost

1. finds a direction of descent $(-g_t)$, which is a function;
2. computes a step of descent w_t according to a point-wise rule (minimizing $F(x_{t-1} - w d_t)$ with respect to $w \in \mathbb{R}$, where $d_t = (-g_t(X_1), \dots, -g_t(X_n))$);
3. updates the optimization variable $f_t = f_{t-1} - w_t(-g_t)$, which is a function.

¹This way to find a direction of descent is related to the Frank–Wolfe algorithm, also known as the conditional gradient method.

It becomes clear that Adaboost is very similar to a gradient descent with exact line search except that:

- ◇ the direction is not the gradient of F at x_{t-1} but $d_t = (-g_t(X_1), \dots, -g_t(X_n))$;
- ◇ gradient descent updates a vector x_t while Adaboost maintains a functional variable f_t .

In a more general setting, we can consider

$$F: x \in \mathbb{R}^n \mapsto \frac{1}{n} \ell_i(x_i),$$

where $\ell_i: \mathbb{R} \rightarrow \mathbb{R}$ is a loss function, which may be, for example:

- ◇ the exponential loss (classification): $\ell_i(x) = \exp(-Y_i x)$ (similar to Adaboost);
- ◇ the logistic loss (classification): $\ell_i(x) = \log(1 + e^{-Y_i x})$ (similar to logistic regression);
- ◇ the squared loss (regression): $\ell_i(x) = \frac{1}{2}(Y_i - x)^2$;
- ◇ the absolute loss (regression): $\ell_i(x) = |Y_i - x|$ (not differentiable in $x = Y_i$).

This is the first improvement of gradient boosting (described in Algorithm 2) over Adaboost. As a second difference, gradient boosting does not build a weak learner g_t highly correlated with $-\nabla F(x_{t-1})$ but such that for all $i \in [n]$,

$$g_t(X_i) \approx -\frac{1}{n} \ell'_i(f_{t-1}(X_i)).$$

Thus, g_t is a base regressor picked in a given class \mathcal{R} . In practice, we get rid of the constant term $\frac{1}{n}$ (this is redundant with the line search), so that $g_t(X_i) \approx -\ell'_i(f_{t-1}(X_i))$.

Algorithm 2 Gradient boosting.

Input: $T \in \mathbb{N}$ (number of iterations), $\nu \in (0, 1]$ (shrinkage coefficient), $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ (training sample).

```

 $f_0 \in \arg \min_{\gamma \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n L(Y_i, \gamma)$  (constant function)
for  $t = 1$  to  $T$  do
  for  $i = 1$  to  $n$  do
     $r_{i,t} \leftarrow -\ell'_i(f_{t-1}(X_i))$  (pseudo-residuals)
  end for
   $g_t \leftarrow$  base regressor from  $\mathcal{R}$  for the training set  $\{(X_i, r_{i,t})\}_{1 \leq i \leq n}$ 
   $w_t \leftarrow \arg \min_{w \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n \ell_i(f_{t-1}(X_i) + w g_t(X_i))$  (line search)
   $f_t = f_{t-1} + \nu w_t g_t$ 
end for

```

Output: $\text{sign}(f_T)$ for classification, f_T for regression.

Example 1.3.1. Let us consider the case where $\ell_i(x) = \frac{1}{2}(Y_i - x)^2$. Then

$$\ell'_i(x) = x - Y_i,$$

and $g_t(X_i) \approx Y_i - f_{t-1}(X_i)$. It appears that $g_t(X_i)$ approximates the quantity (the residue) that is

missing to $f_{t-1}(X_i)$ in order to reach Y_i . With $w_t \approx 1$, the update rule becomes

$$f_t(X_i) \approx f_{t-1}(X_i) + w_t(Y_i - f_{t-1}(X_i)) = (1 - w_t)f_{t-1}(X_i) + w_t Y_i \approx Y_i.$$

To sum up, let for $i \in [n]$, $\ell_i : \mathbb{R} \rightarrow \mathbb{R}$ be a convex and differentiable loss function (adapted to classification or regression), $\mathcal{R} \subset \mathbb{R}^{\mathbb{R}^d}$ be a class of real-valued functions and $F : x \in \mathbb{R}^n \mapsto \frac{1}{n} \sum_{i=1}^n \ell_i(x_i)$.

Then, gradient boosting (Algorithm 2) is an algorithm similar to gradient descent aimed at minimizing the empirical risk:

$$A: f \mapsto F \left(\begin{pmatrix} f(X_1) \\ \vdots \\ f(X_n) \end{pmatrix} \right) = \frac{1}{n} \sum_{i=1}^n \ell_i(f(X_i)),$$

over the linear combinations of functions in \mathcal{R} , denoted $\text{span}(\mathcal{R})$. The iteration of this algorithm reads, for any $t > 0$:

$$f_t = f_{t-1} + w_t g_t,$$

where $g_t \in \mathcal{R}$ is a weak learner such that for all $i \in [n]$, $g_t(X_i) \approx -\frac{1}{n} \ell'_i(f_{t-1}(X_i))$.

Theorem 14 (Linear convergence). *Assume that:*

1. F is differentiable with L -Lipschitz continuous gradient ∇F ($L > 0$);
2. F is μ -strongly convex (with $\mu > 0$);
3. there exists $\gamma \in [0, 1]$ such that at each iteration $t > 0$,

$$\sum_{i=1}^n \left(g_t(X_i) + \frac{1}{n} \ell'_i(f_{t-1})(X_i) \right)^2 \leq (1 - \gamma) \sum_{i=1}^n \left(\frac{1}{n} \ell'_i(f_{t-1})(X_i) \right)^2;$$

4. A has a minimizer in $\text{span}(\mathcal{R})$, denoted f^* .

Let us denote f_T the output of Algorithm 2, then:

$$A(f_T) - A(f^*) \leq \left(1 - \frac{\gamma\mu}{2L} \right)^T (A(f_0) - A(f^*)).$$

The proof is a good exercise.

The importance of the shrinkage coefficient

Given a class of regressors \mathcal{R} , the general problem of gradient boosting is to minimize the empirical risk $A: f \mapsto \frac{1}{n} \sum_{i=1}^n \ell_i(f(X_i))$ over the linear combinations of functions in \mathcal{R} , denoted $\text{span}(\mathcal{R})$. Gradient boosting is a greedy procedure that performs T iterations and outputs a final estimator f_T , where T controls:

- ◇ the number of gradient steps performed to minimize the risk A ;
- ◇ the size of the subspace of $\text{span}(\mathcal{R})$ in which lies f_T , since f_T is a linear combination of at most $T + 1$ functions in \mathcal{R} .

Let us remark that, nothing guarantees that f_T is a minimizer of A over linear combinations of at most $T + 1$ functions in \mathcal{R} . We can even be pretty sure of the converse.

That being said, it is now clear that T controls at the same time the convergence of the optimization algorithm and the complexity of the final estimator. In this sense, T acts as:

- ◇ an *iterative regularizer* (controlling the number of iterations for minimizing the empirical risk A);
- ◇ a *statistical regularizer* (controlling the complexity of the hypothesis space).

However, it is obvious that these two complex regularization mechanisms cannot be monitored by a single parameter. That is why the shrinkage coefficient $\nu \in (0, 1]$ comes into play: by rescaling the contribution of each gradient step, it impacts the convergence of the optimization algorithm while leaving the size of the subspace of $\text{span}(\mathcal{R})$ in which lies f_T unchanged.

To wrap off, gradient boosting benefits from:

- ◇ an iterative regularization, controlled by the pair (ν, T) ;
- ◇ a statistical regularization, controlled by T .

1.4 Support vector machines

1.4.1 Large margin classifier

In its empirical version, logistic regression is computed by minimizing a regularized empirical risk defined by the logistic loss. Such a loss may be replaced by any convex surrogate of the $0 - 1$ loss (Figure 1.4) and in particular by the hinge loss: $(x, x') \in \mathbb{R}^2 \mapsto \max(0, 1 - xx')$. Let $\lambda > 0$ be a regularization parameter. This gives rise to a novel classifier $g_n = \text{sign}(\langle w^*, \cdot \rangle_{\ell_2} + b^*)$, where the decision function parameters (w^*, b^*) are solutions to:

$$\underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} \quad \frac{\lambda}{2} \|w\|_{\ell_2}^2 + \frac{1}{n} \sum_{i=1}^n \max(0, 1 - Y_i(w^\top X_i + b)), \quad (\text{P2})$$

where $\{(X_i, Y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \{\pm 1\}$ is an *iid* sample distributed as (X, Y) . Such a classifier is called a linear support vector machine (SVM) or soft SVM.

The main interest of trading the logistic loss for the hinge loss is to provide a geometrical interpretation. To explain it, let us rewrite the previous optimization problem by replacing the hinge loss by a linear constraint.

Lemma 15. *One has*

$$\forall x \in \mathbb{R}: \quad \max(0, 1 - x) = \inf_{\xi \in \mathbb{R}_+: x \geq 1 - \xi} \xi.$$

Proof. Let $x \in \mathbb{R} : x \geq 1$. Then $1 - x \leq 0$, so $\max(0, 1 - x) = 0$. Besides,

$$\inf_{\xi \in \mathbb{R}_+: x \geq 1 - \xi} \xi = \inf_{\xi \in \mathbb{R}_+: \xi \geq 1 - x} \xi = \inf_{\xi \in \mathbb{R}_+} \xi = 0 = \max(0, 1 - x).$$

Conversely, let $x \in \mathbb{R} : x \leq 1$. Then $1 - x \geq 0$, so $\max(0, 1 - x) = 1 - x$. Moreover,

$$\inf_{\xi \in \mathbb{R}_+ : \xi \geq 1-x} \xi = \inf_{\xi \geq 1-x} \xi = 1 - x = \max(0, 1 - x).$$

□

Let $C = 1/(\lambda n)$ ($C > 0$). Then, (P2) can be rewritten equivalently with slack variables (rescaling the objective function):

$$\begin{aligned} & \underset{\substack{w \in \mathbb{R}^d, b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}}{\text{minimize}} && \frac{1}{2} \|w\|_{\ell_2}^2 + C \sum_{i=1}^n \xi_i \\ & \text{s. t.} && \begin{cases} \forall i \in [n], Y_i(w^\top X_i + b) \geq 1 - \xi_i \\ \forall i \in [n], \xi_i \geq 0. \end{cases} \end{aligned} \quad (\text{P3})$$

In Problem (P3), each slack variable ξ_i represents the uncertainty ($0 < \xi_i \leq 1$) or the error ($\xi_i > 1$) of the decision ($w^\top X_i + b$) given the true label Y_i .

Now, let us assume that the training dataset is linearly separable:

$$\exists (w, b) \in \mathbb{R}^d \times \mathbb{R} : \quad \forall i \in [n], Y_i(w^\top X_i + b) > 0.$$

By rescaling w and b by $\min_{1 \leq i \leq n} Y_i(w^\top X_i + b)$, the previous assumption is equivalent to:

$$\exists (w, b) \in \mathbb{R}^d \times \mathbb{R} : \quad \forall i \in [n], Y_i(w^\top X_i + b) \geq 1.$$

Thus, it is quite natural and legitimate to focus only on classifiers able to classify correctly and with high confidence the training sample (that is with $Y_i(w^\top X_i + b) \geq 1$ for all $i \in [n]$, or equivalently null slack variables: $\forall i \in [n], \xi_i = 0$). The new optimization problem of interest is obtained by increasing C to infinity in (P3) and by remarking that for all $\xi \in \mathbb{R}_+^n$, $\lim_{C \rightarrow \infty} C \sum_{i=1}^n \xi_i = \chi_{\xi=0}$ (with convention $0 \times \infty = 0$):

$$\begin{aligned} & \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\text{minimize}} && \frac{1}{2} \|w\|_{\ell_2}^2 \\ & \text{s. t.} && \forall i \in [n], Y_i(w^\top X_i + b) \geq 1. \end{aligned} \quad (\text{P4})$$

The classifier defined by solving (P4) is called hard margin linear SVM or a large margin classifier because the direction of the decision function achieves the highest *margin*. The legitimacy of Problem (P4) comes from the existence of a solution when the training dataset is linearly separable (which is assumed for now).

Proposition 16. Let $(w, b) \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$ and $x \in \mathbb{R}^d$. Then $\frac{|w^\top x + b|}{\|w\|_{\ell_2}}$ is the distance between the hyperplane $\{z \in \mathbb{R}^d : w^\top z + b = 0\}$ and the point x .

The proof is a good exercise.

Let $(w, b) \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$. The margin of the hyperplane $\{z \in \mathbb{R}^d : w^\top z + b = 0\}$ is defined by:

$$\mu(w, b) = \min_{1 \leq i \leq n} \frac{|w^\top X_i + b|}{\|w\|_{\ell_2}}.$$

Proposition 17. Let us assume that $\exists i \neq j \in [n] : Y_i \neq Y_j$ and let (w_n^*, b_n^*) be a solution to (P4). Then, $\mu(w_n^*, b_n^*) = \frac{1}{\|w_n^*\|_{\ell_2}}$ and (w_n^*, b_n^*) is solution to

$$\begin{aligned} & \underset{w \in \mathbb{R}^d \setminus \{0\}, b \in \mathbb{R}}{\text{maximize}} && \mu(w, b) \\ & \text{s. t.} && \forall i \in [n], Y_i(w^\top X_i + b) \geq 0. \end{aligned} \quad (\text{P5})$$

Proof. First, we have

$$\forall i \in [n], Y_i(\langle w_n^*, X_i \rangle_{\ell_2} + b_n^*) \geq 1 > 0.$$

Besides, since $\exists i \neq j \in [n] : Y_i \neq Y_j$, so $w_n^* \neq 0$. Therefore (w_n^*, b_n^*) is admissible for (P5).

Now, by rescaling, it is easy to see that $\min_{1 \leq i \leq n} Y_i(\langle w_n^*, X_i \rangle_{\ell_2} + b_n^*) = 1$, so $\mu(w_n^*, b_n^*) = \frac{1}{\|w_n^*\|_{\ell_2}}$.

Then, for all $(v, a) \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$ such that $\forall i \in [n], Y_i(v^\top X_i + a) \geq 0$, let $\gamma = \min_{1 \leq i \leq n} Y_i(v^\top X_i + a) = \min_{1 \leq i \leq n} |v^\top X_i + a|$. Either $\gamma = 0$ and $\mu(v, a) = \frac{\gamma}{\|v\|_{\ell_2}} = 0 \leq \mu(w_n^*, b_n^*)$, or $\gamma \neq 0$ and denoting $(w, b) = (v/\gamma, a/\gamma)$, we have $\forall i \in [n], Y_i(w^\top X_i + b) \geq 1$, that is (w, b) is admissible for (P4). Moreover,

$$\mu(v, a) = \frac{\gamma}{\|v\|_{\ell_2}} = \frac{1}{\|w\|_{\ell_2}} \leq \frac{1}{\|w_n^*\|_{\ell_2}} = \mu(w_n^*, b_n^*).$$

This proves that (w_n^*, b_n^*) is solution to (P5). \square

Thus, SVM is said to maximize the margin, that is the distance between the separating hyperplane and the nearest training points. The equivalence holds when the dataset is linearly separable. When this is not true, SVM still maximizes the margin but accepts classification errors embodied by non-zero slack variables ξ_i .

1.4.2 RKHS

The aim of this section is to introduce a class of (potentially) nonlinear functions, that may be used as decision functions in order to build nonlinear classifiers and regressors. The underlying decision functions will have the form of a kernel estimator $\sum_{i=1}^{\infty} \alpha_i k(\cdot, x_i)$ (where $(\alpha_i)_i, (x_i)_i$ and k will be clarified latter), well-known in the statistics community.

In the whole section, we consider a non-empty input set \mathcal{X} .

Definition 1.4.1 (Kernel). A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is called a kernel if there exists a Hilbert space $(\mathcal{G}, \langle \cdot, \cdot \rangle_{\mathcal{G}})$ and a map $\phi : \mathcal{X} \rightarrow \mathcal{G}$ such that

$$\forall (x, x') \in \mathcal{X}^2 : \quad k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}.$$

We call ϕ a feature map and \mathcal{G} a feature space.

Example 1.4.1. Let $\phi_1 : x \in \mathbb{R}^d \mapsto x$ and $\phi_2 : x \in \mathbb{R}^2 \mapsto (x_1^2, x_2^2, \sqrt{2}x_1x_2)$. Then $k_1(x, x') = \phi_1(x)^\top \phi_1(x') = x^\top x'$ and $k_2(x, x') = \phi_2(x)^\top \phi_2(x') = (x^\top x')^2$ are two kernels.

Remark 1.4.1 (Feature map and feature space are not unique). Let us consider $\mathcal{X} = \mathbb{R}^d$ and the kernel $k(x, x') = x^\top x'$. $\phi_1 : x \in \mathbb{R}^d \mapsto x \in \mathbb{R}^d$ and $\phi_2 : x \mapsto \left(\frac{x}{\sqrt{2}}, \frac{x}{\sqrt{2}}\right) \in \mathbb{R}^{2d}$ are two feature maps, respectively with feature spaces $\mathcal{G}_1 = \mathbb{R}^d$ and $\mathcal{G}_2 = \mathbb{R}^{2d}$.

Property 18. *Restriction of kernels* Let k be a kernel on \mathcal{X} , $\tilde{\mathcal{X}}$ be a set and $A : \tilde{\mathcal{X}} \rightarrow \mathcal{X}$. Then $(x, x') \in \tilde{\mathcal{X}}^2 \mapsto k(A(x), A(x'))$ is a kernel on $\tilde{\mathcal{X}}$.

The proof is a good exercise.

Property 19. *Sum of kernels* Let k_1 and k_2 be two kernels on \mathcal{X} and $\alpha \geq 0$. Then αk_1 and $k_1 + k_2$ are kernels.

The proof is a good exercise.

Property 20. *Sum of kernels* Let k_1 and k_2 be two kernels on \mathcal{X} . Then $k_1 k_2$ is a kernel.

The proof is a good exercise.

Property 21 (Polynomial kernels). Assume that $\mathcal{X} \subset \mathbb{R}^d$ and let $p : \mathbb{R} \rightarrow \mathbb{R}$ be a polynomial function with non-negative coefficients. Then $(x, x') \in \mathcal{X}^2 \mapsto p(x^\top x')$ is a kernel.

The proof is a good exercise.

Computing the feature map ϕ is merely needed to define and to evaluate a kernel. We now present a characterization of kernels based on inequalities.

Definition 1.4.2 (Positive definite function). A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is said positive definite if for all $n \in \mathbb{N}$, $\alpha \in \mathbb{R}^n$ and $\{x_1, \dots, x_n\} \subset \mathcal{X}$, we have:

$$\sum_{1 \leq i, j \leq n} \alpha_i \alpha_j k(x_i, x_j) \geq 0.$$

Furthermore, k is said strictly positive definite if for all $n \in \mathbb{N}$, $\alpha \in \mathbb{R}^n \setminus \{0\}$ and $\{x_1, \dots, x_n\} \subset \mathcal{X}$

such that $x_i = x_j \implies i = j$, we have:

$$\sum_{1 \leq i, j \leq n} \alpha_i \alpha_j k(x_i, x_j) > 0.$$

Finally, k is said symmetric if for all $(x, x') \in \mathcal{X}^2$, $k(x, x') = k(x', x)$.

Remark 1.4.2. The definition of a positive definite function k can be trivially linked to positive semi-definiteness of the kernel matrix $K = (k(x_i, x_j))_{1 \leq i, j \leq n}$. In addition, K is often called Gram (or Gramian) matrix.

Let us remark that we obviously have that kernels are symmetric positive definite functions. The following theorem states that symmetric positive definite functions are all kernels.

Theorem 22 (Symmetric positive definite functions are kernels). A function $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof. See [Steinwart and Christmann, 2008, Page 118]. □

Corollary 23 (Limits of kernels). Let $(k_n)_{n \geq 0}$ be a sequence of kernels that converges pointwise to $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$, i.e. for all $(x, x') \in \mathcal{X}^2$, $\lim_{n \rightarrow \infty} k_n(x, x') = k(x, x')$. Then k is a kernel.

Example 1.4.2 (Example of kernels). Let us consider $\mathcal{X} \subset \mathbb{R}^d$. The following functions are common kernels (defined for all $(x, x') \in \mathcal{X}^2$):

linear : $k(x, x') = x^\top x'$;

polynomial : $k(x, x') = (1 + cx^\top x')^d$, $c > 0$, $d \in \mathbb{N}$;

exponential : $k(x, x') = e^{\gamma x^\top x'}$, $\gamma > 0$;

Laplacian : $k(x, x') = e^{-\gamma \|x - x'\|_{\ell_2}}$, $\gamma > 0$;

Gaussian : $k(x, x') = e^{-\gamma \|x - x'\|_{\ell_2}^2}$, $\gamma > 0$.

Kernels are mainly interesting because they define a function space, called a reproducing kernel Hilbert space (RKHS).

Definition 1.4.3 (RKHS). Let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be a Hilbert space and k a kernel. \mathcal{H} is an RKHS with kernel k (or k is a reproducing kernel of \mathcal{H}) if for all $x \in \mathcal{X}$:

- ◇ $k(\cdot, x) \in \mathcal{H}$;
- ◇ $\forall f \in \mathcal{H} : \langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x)$ (reproducing property).

Remark 1.4.3. Let $f \in \mathcal{H}$. Intuitively, f can be described by the infinite dimensional vector of its evaluations $(f(x))_{x \in \mathcal{X}}$, the coordinates of which are $f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$. Even though $\{k(\cdot, x), x \in \mathcal{X}\}$ may not be an orthonormal basis of \mathcal{H} , the reproducing property suggests that we aim at describing f by its expansion on $\{k(\cdot, x), x \in \mathcal{X}\}$: $f = \sum_{x \in \mathcal{X}} \alpha_x k(\cdot, x)$ for some $(\alpha_x)_{x \in \mathcal{X}}$ to be defined.

Proposition 24. Let \mathcal{H} be an RKHS with kernel k . Then, for all $x \in \mathcal{X}$, the evaluation function $E_x : f \in \mathcal{H} \mapsto f(x)$ is continuous.

Proof. For all $x \in \mathcal{X}$, $f, g \in \mathcal{H}$:

$$|E_x(f) - E_x(g)| = |\langle f - g, k(\cdot, x) \rangle_{\mathcal{H}}| \leq \|f - g\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}}.$$

□

In particular, this proposition leads to a remarkable property of RKHSs: norm convergence implies pointwise convergence. Formally, let \mathcal{H} be an RKHS, $f \in \mathcal{H}$ and $(f_n)_n \subset \mathcal{H}$ such that $\|f - f_n\|_{\mathcal{H}} \rightarrow 0$ for $n \rightarrow \infty$. Then, for all $x \in \mathcal{X}$, by continuity of E_x , $f_n(x) = E_x(f_n) \rightarrow E_x(f) = f(x)$ for $n \rightarrow \infty$.

We have just seen that we can build an RKHS with a kernel. We now answer the two questions: given a kernel, is the associated RKHS unique? Given an RKHS, is the associated kernel unique?

Theorem 25 (Uniqueness of the reproducing kernel). *An RKHS \mathcal{H} has a unique reproducing kernel.*

Proof. Let k and k' be two reproducing kernels of \mathcal{H} . For all $x \in \mathcal{X}$, for all $f \in \mathcal{H}$,

$$0 = f(x) - f(x) = \langle f, k(\cdot, x) - k'(\cdot, x) \rangle_{\mathcal{H}}.$$

Since this is true for all $f \in \mathcal{H}$, we have $k(\cdot, x) - k'(\cdot, x) = 0$. This last statement is also true for all $x \in \mathcal{X}$. □

Theorem 26 (Uniqueness of the RKHS, or Moore-Aronszajn theorem). *Let k be a kernel. Then, there exists a unique RKHS \mathcal{H} associated to k .*

Furthermore, let $\mathcal{H}_0 = \text{span} \{k(\cdot, x), x \in \mathcal{X}\}$ associated to the inner product

$$\left\langle \sum_{i=1}^n \alpha_i k(\cdot, x_i), \sum_{j=1}^m \beta_j k(\cdot, x'_j) \right\rangle_{\mathcal{H}_0} = \sum_{\substack{1 \leq i \leq n \\ 1 \leq j \leq m}} \alpha_i \beta_j k(x_i, x'_j),$$

for any integers n and m , any points $x_1, \dots, x_n, x'_1, \dots, x'_m \in \mathcal{X}$ and any vectors $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^m$.

Then, \mathcal{H} is the closure of \mathcal{H}_0 for $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, i.e. \mathcal{H} is the set of functions that are pointwise limits of

Cauchy sequences from \mathcal{H}_0 :

$$\mathcal{H} = \left\{ x \in \mathcal{X} \mapsto \lim_{n \rightarrow \infty} f_n(x) : (f_n)_n \subset \mathcal{H}_0 \text{ Cauchy sequence} \right\},$$

with inner product $\langle \lim_{n \rightarrow \infty} f_n, \lim_{n \rightarrow \infty} g_n \rangle_{\mathcal{H}} = \lim_{n \rightarrow \infty} \langle f_n, g_n \rangle_{\mathcal{H}_0}$.

As a reminder, $(f_n)_n \subset \mathcal{H}_0$ is a Cauchy sequence if $\forall \epsilon > 0, \exists N \in \mathbb{N} : m, n \geq N \implies \|f_m - f_n\|_{\mathcal{H}_0} \leq \epsilon$.

Example 1.4.3. Assume that $k = \langle \cdot, \cdot \rangle_{\ell_2}$. Then $\mathcal{H}_0 \subseteq \{ \langle \cdot, w \rangle_{\ell_2}, w \in \mathbb{R}^d \} \subseteq \mathcal{H}_0$, so $\mathcal{H}_0 = \{ \langle \cdot, w \rangle_{\ell_2}, w \in \mathbb{R}^d \}$ and since \mathcal{H}_0 is already complete, then $\mathcal{H} = \{ \langle \cdot, w \rangle_{\ell_2}, w \in \mathbb{R}^d \}$, i.e. \mathcal{H} is the set of linear functions from \mathbb{R}^d to \mathbb{R} .

Corollary 27. Let \mathcal{H} be an RKHS with kernel k . Then

$$\mathcal{H} = \left\{ \sum_{i=1}^{\infty} \alpha_i k(\cdot, x_i) : \sum_{i=1}^{\infty} \alpha_i^2 k(x_i, x_i) < \infty, (x_i)_i \subset \mathcal{X}, (\alpha_i)_i \subset \mathbb{R} \right\}.$$

Theorem 28. Let \mathcal{G} be a Hilbert space, $\phi : \mathcal{X} \rightarrow \mathcal{G}$ and k the kernel associated to the feature map ϕ : for all $(x, x') \in \mathcal{X}^2$, $k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$.

Then, the RKHS associated to k is:

$$\mathcal{H} = \{ \langle w, \phi(\cdot) \rangle_{\mathcal{G}} : w \in \mathcal{G} \},$$

equipped with the norm:

$$\|f\|_{\mathcal{H}} = \inf \{ \|w\|_{\mathcal{G}} : w \in \mathcal{G}, f = \langle w, \phi(\cdot) \rangle_{\mathcal{G}} \}.$$

In particular, both previous definitions are independent of the feature map ϕ .

Remark 1.4.4 (Canonical feature map). Let k be a kernel and \mathcal{H} its associated RKHS. The canonical feature map of k is defined as

$$\phi : x \in \mathcal{X} \mapsto k(\cdot, x),$$

with the canonical feature space $\mathcal{G} = \mathcal{H}$. Obviously, $\langle \phi(x), \phi(x') \rangle_{\mathcal{G}} = \langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}} = k(x, x')$ for all $(x, x') \in \mathcal{X}^2$.

Definition 1.4.4 (Universal kernel). Assume that \mathcal{X} is compact and let us denote $C(\mathcal{X})$ the set of all continuous and bounded functions on \mathcal{X} . A continuous kernel k on \mathcal{X} is said universal if its RKHS \mathcal{H} is dense in $C(\mathcal{X})$.

Example 1.4.4 (Examples of universal kernels). *The exponential and the Gaussian kernels are universal.*

1.4.3 Kernel trick and nonlinear SVM

Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a kernel and $\{(X_i, Y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \{\pm 1\}$ be an *iid* sample. Let us denote \mathcal{G} and $\phi : \mathbb{R}^d \rightarrow \mathcal{G}$ respectively the feature space and the feature map associated to k .

We would like to find the linear SVM (with tradeoff parameter $C > 0$) for the dataset $\{(\phi(X_i), Y_i)\}_{1 \leq i \leq n}$. Accordingly to Problem (P3), the parameters of such an SVM are solution to the optimization problem:

$$\begin{aligned} & \underset{\substack{w \in \mathcal{G}, b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}}{\text{minimize}} && \frac{1}{2} \|w\|_{\mathcal{G}}^2 + C \sum_{i=1}^n \xi_i \\ & \text{s. t.} && \begin{cases} \forall i \in [n], Y_i(\langle w, \phi(X_i) \rangle_{\mathcal{G}} + b) \geq 1 - \xi_i \\ \forall i \in [n], \xi_i \geq 0. \end{cases} \end{aligned} \quad (\text{P6})$$

Let now \mathcal{H} be the RKHS associated to k . Thanks to Theorem 28, we know that for all $w \in \mathcal{G}$, $h = \langle w, \phi(\cdot) \rangle_{\mathcal{G}} \in \mathcal{H}$ and $\|h\|_{\mathcal{H}} = \inf \left\{ \|w'\|_{\mathcal{G}} : w' \in \mathcal{G}, f = \langle w', \phi(\cdot) \rangle_{\mathcal{G}} \right\}$. Therefore, by a change of variable, (P6) can be written:

$$\begin{aligned} & \underset{\substack{h \in \mathcal{H}, b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}}{\text{minimize}} && \frac{1}{2} \|h\|_{\mathcal{H}}^2 + C \sum_{i=1}^n \xi_i \\ & \text{s. t.} && \begin{cases} \forall i \in [n], Y_i(h(X_i) + b) \geq 1 - \xi_i \\ \forall i \in [n], \xi_i \geq 0. \end{cases} \end{aligned} \quad (\text{P7})$$

(P7) reveals that by transforming the data with the feature map ϕ , a linear SVM can be used to estimate a decision function $f = h + b$, $h \in \mathcal{H}$, $b \in \mathbb{R}$, which is nonlinear as soon as the kernel k is not the linear kernel. On the other hand, when k is the linear kernel, ϕ boils to be the identity and \mathcal{H} the set of linear functions (i.e. (P3) and (P7) are strictly the same).

A central question with nonlinear SVM is to compute it in practice. This is not trivial since, on the one hand, solving (P6) involves computing the feature map ϕ (which is unknown for certain kernels, even infinite dimensional for some kernels such as the Gaussian kernel). On the other hand, (P7) involves a nonparametric optimization variable $h \in \mathcal{H}$.

The next theorem states that solutions to (P7) are supported by the data. It is quite reassuring since it provides a way to solve (P7): restricting h to be of the form $\sum_{i=1}^n \alpha_i k(\cdot, X_i)$, for $\alpha \in \mathbb{R}^n$.

Theorem 29 (Representer theorem). *Let $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ be a kernel and \mathcal{H} the associated RKHS. Let also $\psi : \mathbb{R}_+ \rightarrow \mathbb{R}$ be a non-decreasing function and $\ell : \mathbb{R}^{2n} \rightarrow \mathbb{R}$ be any loss function. Given a training sample $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ from $(\mathcal{X} \times \mathbb{R})^n$, if the optimization problem*

$$\underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \quad \psi(\|h\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b) \quad (\text{P8})$$

has a solution, then there exists a solution (h^*, b^*) such that h^* has the form

$$h^* = \sum_{i=1}^n \alpha_i k(\cdot, X_i),$$

where $\alpha \in \mathbb{R}^n$.

In addition, if ψ is an increasing function, then all solutions of (P8) can be written in the form described above.

Proof. Let $(h, b) \in \mathcal{H} \times \mathbb{R}$ be a solution to (P8). Let us consider $V = \text{span} \{k(\cdot, X_i), i \in [n]\} \neq \emptyset$ as well as its orthogonal complement V_\perp . Let $(h_\parallel, h_\perp) \in V \times V_\perp$ such that $h = h_\parallel + h_\perp$ and let us remark that

- ◇ by the Pythagorean theorem, $\|h\|_{\mathcal{H}} = \sqrt{\|h_\parallel\|_{\mathcal{H}}^2 + \|h_\perp\|_{\mathcal{H}}^2} \geq \|h_\parallel\|_{\mathcal{H}}$;
- ◇ for all $i \in [n]$, $h_\perp(X_i) = \langle h_\perp, k(\cdot, X_i) \rangle_{\mathcal{H}} = 0$.

Consequently,

$$\begin{aligned} & \psi(\|h\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b) \\ & \geq \psi(\|h_\parallel\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b) \quad (\psi \text{ non-decreasing and } \|h\|_{\mathcal{H}} \geq \|h_\parallel\|_{\mathcal{H}}) \\ & = \psi(\|h_\parallel\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h_\parallel(X_1) + b, \dots, h_\parallel(X_n) + b) \quad (h(X_i) = h_\parallel(X_i) + h_\perp(X_i) = h_\parallel(X_i)), \end{aligned}$$

which ensures that (h_\parallel, b) is solution to (P8), with $h_\parallel \in V$, that is $h_\parallel = \sum_{i=1}^n \alpha_i k(\cdot, X_i)$ for some $\alpha \in \mathbb{R}^n$.

In addition, since (h, b) and (h_\parallel, b) are solutions to (P8), we also have

$$\begin{aligned} & \psi(\|h\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b) \\ & = \psi(\|h_\parallel\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h_\parallel(X_1) + b, \dots, h_\parallel(X_n) + b) \\ & = \psi(\|h_\parallel\|_{\mathcal{H}}) + \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b), \end{aligned}$$

which leads to $\psi(\|h_\parallel\|_{\mathcal{H}}) = \psi(\|h\|_{\mathcal{H}})$. Therefore, as soon as ψ is increasing, one has $\|h_\parallel\|_{\mathcal{H}}^2 = \|h\|_{\mathcal{H}}^2 = \|h_\parallel\|_{\mathcal{H}}^2 + \|h_\perp\|_{\mathcal{H}}^2$. So $\|h_\perp\|_{\mathcal{H}} = 0$, i.e. $h_\perp = 0$ and $h = h_\parallel$. \square

Remark 1.4.5. Even if ψ is an increasing function, (P8) may not have a solution. For instance, let $R : (h, b) \in \mathcal{H} \times \mathbb{R} \mapsto \|h\|_{\mathcal{H}}^2 + e^{-(h(X_1)+b)}$. Then, for any pair (h, b) , $R(h, b) > 0$ and $R(0, b) \xrightarrow{b \rightarrow +\infty} 0$. So (P8) has no minimizer.

Another example is $R : (h, b) \in \mathcal{H} \times \mathbb{R} \mapsto \|h\|_{\mathcal{H}}^2 - h(X_1)^4$. Let $h_\lambda = \lambda k(\cdot, X_1)$. Then $R(h_\lambda, 0) = \lambda^2 k(X_1, X_1) - \lambda^4 k(X_1, X_1) \xrightarrow{\lambda \rightarrow \infty} -\infty$. So (P8) has no minimizer.

Remark 1.4.6. If ψ and $(h, b) \mapsto \ell(Y_1, \dots, Y_n, h(X_1) + b, \dots, h(X_n) + b)$ are strictly convex, then the pair (h^*, b^*) is unique but the expansion of h^* may not be (it is the case if the kernel matrix $(k(X_i, X_j))_{1 \leq i, j \leq n}$ is rank deficient).

In practice, the duality theory of convex optimization is preferred to the representer theorem in order to solve (P7). It simultaneously exhibit the same result as the representer theorem and a novel optimization problem to determine the optimal weights α . The next sections are devoted to deriving a dual optimization problem to (P7).

Let us remark that the forthcoming derivation could also be executed based on (P6), i.e. using only the feature space notation ($w \in \mathcal{G}$ and $\phi(X_i) \in \mathcal{G}$). The final result would be exactly the same: in order to compute the optimal decision function, we only need to evaluate the kernel k but never to compute the feature map ϕ . This is known as the kernel trick.

1.4.4 SVM in action

- ◇ A fancy demo of polynomial kernel.
- ◇ An applet for playing with SVM.

1.4.5 Duality in convex optimization

Definition 1.4.5. A function $f: \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \forall t \in (0, 1): \quad f(tx + (1 - t)y) \leq tf(x) + (1 - t)f(y).$$

Lemma 30. Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex et differentiable function. Then

$$\forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, : \quad f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle_{\ell_2}.$$

Proof. We have

$$\begin{aligned} \forall t \in (0, 1): f(x + t(y - x)) - f(x) &= f((1 - t)x + ty) - f(x) \\ &\leq (1 - t)f(x) + tf(y) - f(x) && \text{by convexity} \\ &= t(f(y) - f(x)). \end{aligned}$$

Thus,

$$\begin{aligned} \langle \nabla f(x), y - x \rangle_{\ell_2} &= \lim_{t \rightarrow 0} \frac{f(x + t(y - x)) - f(x)}{t} \\ &= \lim_{t \rightarrow 0^+, t < 1} \frac{f(x + t(y - x)) - f(x)}{t} \\ &\leq \lim_{t \rightarrow 0^+, t < 1} \frac{t(f(y) - f(x))}{t} \\ &= f(y) - f(x). \end{aligned}$$

□

Theorem 31 (Fermat's rule). *Let $f: \mathbb{R}^d \rightarrow \mathbb{R}$ be a convex et differentiable function. Then*

$$x^* \in \arg \min_{x \in \mathbb{R}^d} f(x) \iff \nabla f(x^*) = 0.$$

Proof. By convexity,

$$\forall x \in \mathbb{R}^d: f(x) - f(x^*) \geq \langle \nabla f(x^*), x - x^* \rangle_{\ell_2}.$$

Thus, $\nabla f(x^*) = 0 \implies x^* \in \arg \min_{x \in \mathbb{R}^d} f(x)$.

On the other hand, if $\nabla f(x^*) \neq 0$, then $\|\nabla f(x^*)\|_{\ell_2}^2 = \langle \nabla f(x^*), \nabla f(x^*) \rangle_{\ell_2} > 0$, that is

$$\lim_{t \rightarrow 0^-} \frac{f(x^* + t \nabla f(x^*)) - f(x^*)}{t} > 0.$$

So, $\exists t < 0$ such that $\frac{f(x^* + t \nabla f(x^*)) - f(x^*)}{t} > 0$. This implies $f(x^* + t \nabla f(x^*)) - f(x^*) < 0$. In other words, x^* is not a minimizer of f . □

From now on, we consider the optimization problem

$$\begin{aligned} & \underset{x \in \mathbb{R}^d}{\text{minimize}} && f(x) \\ & \text{s. t.} && \begin{cases} \forall i \in [n]: g_i(x) \leq 0 \\ \forall i \in [m]: h_i(x) = 0, \end{cases} \end{aligned} \tag{P9}$$

where $f: \mathbb{R}^d \rightarrow \mathbb{R}$, $(g_i: \mathbb{R}^d \rightarrow \mathbb{R})_{1 \leq i \leq n}$ and $(h_i: \mathbb{R}^d \rightarrow \mathbb{R})_{1 \leq i \leq m}$ are $n + m + 1$ convex and differentiable functions. let $\mathcal{C} = \{x \in \mathbb{R}^d : \forall i \in [n], g_i(x) \leq 0, \forall i \in [m], h_i(x) = 0\}$ be the set of constraints.

Definition 1.4.6 (Lagrangian function). *The Lagrangian function of (P9) is:*

$$L: (x, \lambda, \nu) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m \mapsto f(x) + \sum_{i=1}^n \lambda_i g_i(x) + \sum_{i=1}^m \nu_i h_i(x).$$

Property 32. *Let us consider (P9) along with its Lagrangian L . Then*

$$\forall x \in \mathbb{R}^d: \sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(x, \lambda, \nu) = \begin{cases} f(x) & \text{if } x \in \mathcal{C} \\ \infty & \text{otherwise.} \end{cases}$$

In addition, if $\mathcal{C} \neq \emptyset$, then

$$\inf_{x \in \mathcal{C}} f(x) = \inf_{x \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(x, \lambda, \nu).$$

The proof is a good exercise.

Remark 1.4.7 (Variational formulation of the characteristic function). *Considering $f = 0$ leads to $\sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(\cdot, \lambda, \nu) = \chi_C$, where χ_C is the characteristic function of the set C . In other words, $\sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(\cdot, \lambda, \nu)$ is no more than a variational formulation of the characteristic function of the set of constraint C .*

Then (in the general case where $f \neq 0$), it becomes obvious that (P9) can be reformulated as the minimization of $f + \chi_C$, i.e. of

$$f + \sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} \sum_{i=1}^n \lambda_i g_i + \sum_{i=1}^m \nu_i h_i.$$

Definition 1.4.7 (Dual function). *Let us consider (P9) along with its Lagrangian L . The dual function of (P9) is*

$$D: (\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m \mapsto \inf_{x \in \mathbb{R}^d} L(x, \lambda, \nu).$$

Let us remark that D is concave and may take value $-\infty$ for some (λ, ν) .

Property 33 (Weak duality).

$$\sup_{(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m} D(\lambda, \nu) \leq \inf_{x \in C} f(x).$$

Proof. For all $(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m$ and for all $x \in C \subseteq \mathbb{R}^d$,

$$D(\lambda, \nu) = \inf_{x' \in \mathbb{R}^d} L(x', \lambda, \nu) \leq L(x, \lambda, \nu).$$

Moreover, since $x \in C$:

$$L(x, \lambda, \nu) = f(x) + \sum_{i=1}^n \lambda_i g_i(x) + \sum_{i=1}^m \nu_i h_i(x) \leq f(x).$$

Therefore,

$$D(\lambda, \nu) \leq f(x),$$

which leads to the expected result. □

The optimization problem

$$\begin{aligned} & \underset{(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m}{\text{maximize}} && D(\lambda, \nu) \\ & \text{s. t.} && \lambda \succeq 0 \end{aligned} \tag{P10}$$

is called the *dual problem* to (P9), itself called the *primal problem*.

Definition 1.4.8 (Convex problem). Problem (P9) is said convex if

1. $f: \mathbb{R}^d \rightarrow \mathbb{R}$ and $(g_i: \mathbb{R}^d \rightarrow \mathbb{R})_{1 \leq i \leq n}$ are $n + 1$ convex functions;
2. $(h_i: \mathbb{R}^d \rightarrow \mathbb{R})_{1 \leq i \leq m}$ are affine functions.

In this case, \mathcal{C} is a convex set.

Theorem 34 (Strong duality). Let us assume that (P9) is convex. If (Slater's constraint qualification)

$$\exists x \in \mathbb{R}^d : \forall i \in [n], g_i(x) < 0 \text{ and } \forall i \in [m], h_i(x) = 0,$$

then

1. $\sup_{(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m} D(\lambda, \nu) = \inf_{x \in \mathcal{C}} f(x)$ (zero duality gap);
2. $\exists (\lambda^*, \nu^*) \in \mathbb{R}_+^n \times \mathbb{R}^m : \sup_{(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m} D(\lambda, \nu) = D(\lambda^*, \nu^*)$ (dual attainment).

Theorem 34 states that the minimal value of f in \mathcal{C} can be recovered by maximizing the dual function D . However, our main interest is more about linking solutions to these two optimization problems, rather than the optimal values: in practice, the estimator we want to build is solution to the primal problem and we would like to recover it from a solution to the dual problem. Theorem 37 makes this link explicit.

Definition 1.4.9 (Saddle point). $(x^*, \lambda^*, \nu^*) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m$ is a saddle point of L if

$$\forall (x, \lambda, \nu) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m : L(x^*, \lambda, \nu) \leq L(x^*, \lambda^*, \nu^*) \leq L(x, \lambda^*, \nu^*).$$

Property 35. Let $(x^*, \lambda^*, \nu^*) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m$. Then, $(x^*, \lambda^*, \nu^*) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m$ is a saddle point of L if and only if

$$\sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^d} L(x, \lambda, \nu) = \sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(x^*, \lambda, \nu) = L(x^*, \lambda^*, \nu^*),$$

and

$$\inf_{x \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+^n, \nu \in \mathbb{R}^m} L(x, \lambda, \nu) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*).$$

Proof. Without loss of generality, we omit the variable ν . Let assume that (x^*, λ^*) is a saddle point of L . By weak duality, we have

$$\sup_{\lambda \in \mathbb{R}_+^n} \inf_{x \in \mathbb{R}^d} L(x, \lambda) \leq \inf_{x \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+^n} L(x, \lambda).$$

Besides, by definition,

$$\inf_{x \in \mathbb{R}^d} \sup_{\lambda \in \mathbb{R}_+^n} L(x, \lambda) \leq \sup_{\lambda \in \mathbb{R}_+^n} L(x^*, \lambda),$$

and since (x^*, λ^*) is a saddle point of L ,

$$\sup_{\lambda \in \mathbb{R}_+^n} L(x^*, \lambda) = L(x^*, \lambda^*).$$

Then, by the same argument

$$L(x^*, \lambda^*) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*),$$

and by definition,

$$\inf_{x \in \mathbb{R}^d} L(x, \lambda^*) \leq \sup_{\lambda \in \mathbb{R}_+^n} \inf_{x \in \mathbb{R}^d} L(x, \lambda).$$

Chaining everything implies that all inequalities are in fact equalities. That is what we had to prove. The converse implication is straightforward. \square

Corollary 36. *Let us assume that (P9) is convex and that Slater's constraint qualification hold. Let $(\lambda^*, \nu^*) \in \mathbb{R}_+^n \times \mathbb{R}^m$ be the point where the dual is attained. If there exists $x^* \in \mathcal{C}$ such that the primal is attained in x^* , then $(x^*, \lambda^*, \nu^*) \in \mathbb{R}^d \times \mathbb{R}_+^n \times \mathbb{R}^m$ is a saddle point of L and*

$$D(\lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*) = f(x^*).$$

Theorem 37 (Karush–Kuhn–Tucker (KKT) conditions). *Let us assume that (P9) is convex and that Slater's constraint qualification hold. $x^* \in \mathbb{R}^d$ and $(\lambda^*, \nu^*) \in \mathbb{R}^n \times \mathbb{R}^m$ are respectively solutions to (P9) and (P10) if and only if*

1. *primal feasibility:* $\forall i \in [n], g_i(x^*) \leq 0$ and $\forall i \in [m], h_i(x^*) = 0$;
2. *dual feasibility:* $\lambda^* \succeq 0$
3. *complementary slackness:* $\forall i \in [n], \lambda_i^* g_i(x^*) = 0$;
4. *stationarity:* $\nabla_x L(x^*, \lambda^*, \nu^*) = 0$.

Proof. On the first hand, let us assume that $x^* \in \mathbb{R}^d$ and $(\lambda^*, \nu^*) \in \mathbb{R}^n \times \mathbb{R}^m$ are respectively solutions to (P9) and (P10).

1. primal feasibility comes from x^* being solution to (P9);
2. dual feasibility comes from $(\lambda^*, \nu^*) \in \mathbb{R}^n \times \mathbb{R}^m$ being solution to (P10);
3. one has

$$\begin{aligned} D(\lambda^*, \nu^*) &= \inf_{x \in \mathbb{R}^d} L(x, \lambda^*, \nu^*) && \text{(definition)} \\ &\leq L(x^*, \lambda^*, \nu^*) \\ &= f(x^*) + \sum_{i=1}^n \lambda_i^* g_i(x^*) + \sum_{i=1}^m \nu_i^* h_i(x^*) \\ &\leq f(x^*) && (\lambda_i^* g_i(x^*) \geq 0 \text{ and } h_i(x^*) = 0). \end{aligned}$$

- But, by Slater's constraint qualification, strong duality holds, that is $D(\lambda^*, \nu^*) = f(x^*)$, so both previous inequalities are in fact equalities. We deduce that x^* is a minimizer of $L(\cdot, \lambda^*, \nu^*)$ and that $\sum_{i=1}^n \lambda_i g_i(x^*) = 0$. This last inequality is exactly the complementary slackness;
4. from x^* being a minimizer of $L(\cdot, \lambda^*, \nu^*)$ (which is a convex function), from Fermat's rule, we have the stationarity condition.

On the second hand, if $x^* \in \mathbb{R}^d$ and $(\lambda^*, \nu^*) \in \mathbb{R}^n \times \mathbb{R}^m$ fulfill the KKT conditions, then

1. from 1. and 3., $f(x^*) = L(x^*, \lambda^*, \nu^*)$;
2. from 2. and 4., convexity of $L(\cdot, \lambda^*, \nu^*)$ and Fermat's rule, one has $D(\lambda^*, \nu^*) = \inf_{x \in \mathbb{R}^d} L(x, \lambda^*, \nu^*) = L(x^*, \lambda^*, \nu^*)$.

Consequently $f(x^*) = D(\lambda^*, \nu^*)$. Moreover, by weak duality

$$f(x^*) \geq \inf_{x \in \mathcal{C}} f(x) \geq \sup_{(\lambda, \nu) \in \mathbb{R}_+^n \times \mathbb{R}^m} D(\lambda, \nu) \geq D(\lambda^*, \nu^*) = f(x^*),$$

so all the inequalities are equalities and x^* and (λ^*, ν^*) are respectively solutions to (P9) and (P10). \square

1.4.6 Dual problem and support vectors

Let $Q = (k(X_i, X_j)Y_i Y_j)_{1 \leq i, j \leq n}$ be the labeled kernel matrix. By Lagrange duality, a dual to (P7) is:

$$\begin{aligned} & \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} && -\frac{1}{2} \alpha^\top Q \alpha + \mathbf{1}^\top \alpha \\ & \text{s. t.} && \begin{cases} \forall i \in [n]: 0 \leq \alpha_i \leq C \\ y^\top \alpha = 0. \end{cases} \end{aligned} \tag{P11}$$

Problem (P11) is generally solved by sequential minimal optimization (SMO), for which a simplified version is described in Algorithm 3.

Algorithm 3 Sequential minimal optimization.

Input: $C > 0$ (tradeoff parameter), $k: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ (kernel function), $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ (training sample).

$Q \leftarrow (k(X_i, X_j)Y_i Y_j)_{1 \leq i, j \leq n}$ (labeled kernel matrix)

while not converged **do**

 find α_i for which KKT conditions are violated

 pick $\alpha_j \neq \alpha_i$ at random

 solve Problem (P11) with respect to (α_i, α_j) with all other variables fixed

end while

Output: $(\alpha_1, \dots, \alpha_n)$.

Property 38 (KKT conditions for SVM). *Let (h_n^*, b_n^*) be an SVM defined accordingly to (P7). Then, (P11) has a solution, denoted $\alpha^* \in \mathbb{R}^n$, and*

$$\diamond h_n^* = \sum_{i=1}^n \alpha_i^* Y_i k(\cdot, X_i);$$

◇ for all $i \in [n]$,

$$\begin{aligned} Y_i(h_n^*(X_i) + b_n^*) &> 1 \implies \alpha_i^* = 0 \\ Y_i(h_n^*(X_i) + b_n^*) &< 1 \implies \alpha_i^* = C; \end{aligned}$$

◇ for all $i \in [n]$,

$$\begin{aligned} \alpha_i^* = 0 &\implies Y_i(h_n^*(X_i) + b_n^*) \geq 1 \\ \alpha_i^* = C &\implies Y_i(h_n^*(X_i) + b_n^*) \leq 1 \\ 0 < \alpha_i^* < C &\implies Y_i(h_n^*(X_i) + b_n^*) = 1; \end{aligned}$$

◇ denoting

$$\begin{aligned} \mathcal{M} &= \{i \in [n] : 0 < \alpha_i^* < C\} \\ \mathcal{I} &= \{i \in [n] : \alpha_i^* = C\} \\ \mathcal{O} &= \{i \in [n] : \alpha_i^* = 0\} \end{aligned}$$

and respectively \mathcal{I}_+ , \mathcal{I}_- , \mathcal{O}_+ , \mathcal{O}_- the intersection of \mathcal{I} and \mathcal{O} with $\{i \in [n] : Y_i = 1\}$ and $\{i \in [n] : Y_i = -1\}$, one has

$$b_n^* \in \left[\max_{i \in \mathcal{I}_- \cup \mathcal{O}_+} Y_i - h_n^*(X_i), \min_{i \in \mathcal{I}_+ \cup \mathcal{O}_-} Y_i - h_n^*(X_i) \right];$$

◇ $\forall i \in \mathcal{M}, b_n^* = Y_i - h_n^*(X_i)$.

Proof. First, by setting

$$\begin{cases} h = 0 \in \mathcal{H} \\ b = 1 \\ \forall i \in [n], \xi_i = 3, \end{cases}$$

we have

$$\begin{cases} \forall i \in [n], Y_i(h(X_i) + b) = Y_i > -2 = 1 - \xi_i \\ \forall i \in [n], \xi_i = 3 > 0, \end{cases}$$

so Slater's constraint qualification holds and KKT conditions indicate that (P11) has a solution α^* .

Moreover, denoting $\alpha \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^n$ the Lagrangian multipliers for the two constraints of (P7), the Lagrangian reads, for all admissible (h, b, ξ) , $\alpha \succcurlyeq 0$ and $\beta \succcurlyeq 0$:

$$L(h, b, \xi, \alpha) = \frac{1}{2} \|h\|_{\mathcal{H}}^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i^\top (1 - \xi_i - Y_i(\langle h, k(\cdot, X_i) \rangle_{\mathcal{H}} + b)) - \sum_{i=1}^n \beta_i \xi_i.$$

Denoting ξ^* the primal solution compaignon to (h_n^*, b_n^*) and β^* the dual solution compaignon to α^* , by stationarity, we have

$$\nabla_h L(h_n^*, b_n^*, \xi^*) = 0 \iff h_n^* = \sum_{i=1}^n \alpha_i^* Y_i k(\cdot, X_i),$$

and

$$\nabla_{\xi} L(h_n^*, b_n^*, \xi^*) = 0 \iff \beta^* = C\mathbb{1} - \alpha^*.$$

In addition, by complementarity slackness, for all $i \in [n]$:

$$\alpha_i^* = 0 \text{ or } Y_i(h_n^*(X_i) + b_n^*) = 1 - \xi_i^* \quad (1.1)$$

$$\alpha_i^* = C \text{ (} \beta_i^* = 0 \text{) or } \xi_i^* = 0. \quad (1.2)$$

As a consequence, for all $i \in [n]$, if $Y_i(h_n^*(X_i) + b_n^*) > 1$, we necessarily have $\xi_i^* = 0$ (if $\xi_i^* > 0$, we can replace it by 0 without violating the constraint $Y_i(h_n^*(X_i) + b_n^*) \geq 1 - \xi_i^*$ and with a strictly smallest $C \sum_{i=1}^n \xi_i^*$ in the objective value, meaning that ξ_i^* was not solution, which is contradictory). Thus, $Y_i(h_n^*(X_i) + b_n^*) > 1 - \xi_i^*$ and by 1.1, $\alpha_i^* = 0$.

On the other hand, if $Y_i(h_n^*(X_i) + b_n^*) < 1$, then necessarily $\xi_i^* > 0$. Thus, by 1.2, $\alpha_i^* = C$.

Conversely,

$$\alpha_i^* = 0 \implies \xi_i^* = 0 \implies Y_i(h_n^*(X_i) + b_n^*) \geq 1 - \xi_i^* = 1$$

$$\alpha_i^* = C \implies Y_i(h_n^*(X_i) + b_n^*) = 1 - \xi_i^* \implies Y_i(h_n^*(X_i) + b_n^*) \leq 1 \quad (\xi_i^* \geq 0)$$

$$0 < \alpha_i^* < C \implies Y_i(h_n^*(X_i) + b_n^*) = 1 - \xi_i^* \text{ and } \xi_i^* = 0 \implies Y_i(h_n^*(X_i) + b_n^*) = 1.$$

Concerning the intercept, one has for all $i \in [n]$,

$$i \in \mathcal{O}_+ \implies Y_i(h_n^*(X_i) + b_n^*) \geq 1 \implies b_n^* \geq Y_i - h_n^*(X_i)$$

$$i \in \mathcal{O}_- \implies Y_i(h_n^*(X_i) + b_n^*) \geq 1 \implies b_n^* \leq Y_i - h_n^*(X_i)$$

$$i \in \mathcal{I}_+ \implies Y_i(h_n^*(X_i) + b_n^*) \leq 1 \implies b_n^* \leq Y_i - h_n^*(X_i)$$

$$i \in \mathcal{I}_- \implies Y_i(h_n^*(X_i) + b_n^*) \leq 1 \implies b_n^* \geq Y_i - h_n^*(X_i).$$

Thus $Y_i - h_n^*(X_i) \leq b_n^* \leq Y_{i'} - h_n^*(X_{i'})$ for all $i \in \mathcal{I}_- \cup \mathcal{O}_+$ and for all $i' \in \mathcal{I}_+ \cup \mathcal{O}_-$.

Finally, for all $i \in \mathcal{M}$, one has $Y_i(h_n^*(X_i) + b_n^*) = 1$, which implies $b_n^* = Y_i - h_n^*(X_i)$. \square

Remark 1.4.8. When $\mathcal{M} \neq \emptyset$ (that is, there exists a point on the border of the margin), the intercept b_n^* can be obtained easily. Otherwise, b_n^* can be chosen in the prescribed interval (for instance as the mean).

Remark 1.4.9. In practice a good choice of C (with respect to a cross-validation score) makes it sufficiently large so that many points X_i are well classified and out of the margin. In other words, $\mathcal{I} \cap \mathcal{M}$ is relatively small and

$$h_n^* = \sum_{i=1}^n \alpha_i^* Y_i k(\cdot, X_i) = \sum_{\substack{1 \leq i \leq n \\ \alpha_i^* > 0}} \alpha_i^* Y_i k(\cdot, X_i)$$

is supported only by few vectors X_i ($i \in \mathcal{I} \cap \mathcal{M}$). The latter are called the support vectors of the

SVM (h_n^*, b_n^*) .

1.4.7 Statistical perspective

From a statistical perspective, an SVM is not a large margin classifier but an estimator $f_n^* = h_n^* + b_n^*$ (where b_n^* is called an intercept) of the Bayes classifier defined by

$$(h_n^*, b_n^*) \in \arg \min_{\substack{h \in \mathcal{H}: \|h\|_{\mathcal{H}} \leq c \\ b \in \mathbb{R}, f = h + b}} \frac{1}{n} \sum_{i=1}^n \max(0, 1 - Y_i f(X_i)), \quad (1.3)$$

where $c > 0$.

The optimization problem of interest possesses a capacity constraint $\|h\|_{\mathcal{H}} \leq c$, which makes it possible to derive generalization guarantees.

In fact, Equation (1.3) is equivalent to (P7).

Proposition 39 (Tikhonov regularization). *There exists $\lambda \geq 0$ such that the SVM defined by (1.3) is a minimizer of*

$$\underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \quad \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n \max(0, 1 - Y_i(h(X_i) + b)). \quad (\text{P12})$$

Respectively, let $\lambda > 0$ and (h^, b^*) be solutions to (P12). Then, there exists $c > 0$ such that (h^*, b^*) are also minimizers of Equation (1.3).*

Proof. Equation (1.3) can be reformulated as an optimization problem:

$$\begin{aligned} & \underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \quad \ell(h, b) \\ & \text{s. t.} \quad \|h\|_{\mathcal{H}}^2 \leq c^2, \end{aligned}$$

where $\ell(h, b) = \frac{1}{n} \sum_{i=1}^n \max(0, 1 - Y_i(h(X_i) + b))$ and the objective as well as the inequality constraint functions are convex. Let $D : \lambda \geq 0 \mapsto \inf_{h \in \mathcal{H}, b \in \mathbb{R}} \ell(h, b) + \lambda(\|h\|_{\mathcal{H}}^2 - c^2)$ be the dual function. Since $c > 0$, Slater's constraint qualification is fulfilled and by Theorem 34, strong duality holds. Consequently, there exists $\lambda^* \geq 0$ such that

$$\inf_{h \in \mathcal{H}: \|h\|_{\mathcal{H}}^2 \leq c^2, b \in \mathbb{R}} \ell(h, b) = \sup_{\lambda \geq 0} D(\lambda) = D(\lambda^*) = \inf_{h \in \mathcal{H}, b \in \mathbb{R}} \ell(h, b) + \lambda^*(\|h\|_{\mathcal{H}}^2 - c^2).$$

Therefore, for any $(h, b) \in \mathcal{H} \times \mathbb{R}$,

$$\begin{aligned} \ell(h_n^*, b_n^*) + \lambda^*(\|h_n^*\|_{\mathcal{H}}^2 - c^2) &\leq \ell(h_n^*, b_n^*) & (\lambda^*(\|h_n^*\|_{\mathcal{H}}^2 - c^2) \leq 0) \\ &= \inf_{h' \in \mathcal{H}: \|h'\|_{\mathcal{H}}^2 \leq c^2, b' \in \mathbb{R}} \ell(h', b') \\ &= \inf_{h' \in \mathcal{H}, b' \in \mathbb{R}} \ell(h', b') + \lambda^*(\|h'\|_{\mathcal{H}}^2 - c^2) & (\text{strong duality}) \\ &\leq \ell(h, b) + \lambda^*(\|h\|_{\mathcal{H}}^2 - c^2). \end{aligned}$$

Consequently,

$$(h_n^*, b_n^*) \in \arg \min_{\substack{h \in \mathcal{H}, \\ b \in \mathbb{R}}} \frac{2\lambda^*}{2} \|h\|_{\mathcal{H}}^2 + \ell(h, b).$$

The converse implication is more direct by choosing $c = \|h^*\|_{\mathcal{H}}$: let $\lambda > 0$ and $(h^*, b^*) \in \mathcal{H} \times \mathbb{R}$ be solution to

$$\underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \ell(h, b) + \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2.$$

For all $(h, b) \in \mathcal{H} \times \mathbb{R}$ such that $\|h\|_{\mathcal{H}} \leq c$,

$$\begin{aligned} \ell(h^*, b^*) &= \ell(h^*, b^*) + \frac{\lambda}{2} \|h^*\|_{\mathcal{H}}^2 - \frac{\lambda}{2} \|h^*\|_{\mathcal{H}}^2 \\ &\leq \ell(h, b) + \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 - \frac{\lambda}{2} c^2 \\ &\leq \ell(h, b) \quad (\|h\|_{\mathcal{H}}^2 - c^2 \leq 0). \end{aligned}$$

□

1.5 A detour to nonparametric regression

1.5.1 Least mean squares

Let $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function and let us consider the model:

$$Y = g(X) + \epsilon,$$

where

- ◇ $X \in \mathbb{R}^d$ is a random vector;
- ◇ $\epsilon \in \mathbb{R}$ is a random variable, such that $\epsilon \in L^2$ and $\mathbb{E}[\epsilon|X] = 0$;
- ◇ $g(X) \in L^2$.

Let $\mathcal{F} = \{f: \mathbb{R}^d \rightarrow \mathbb{R}, f(X) \in L^2\}$.

Theorem 40. *The function g is a minimizer of the least mean squares risk functional $f \in \mathcal{F} \mapsto \mathbb{E}((Y - f(X))^2)$ over \mathcal{F} .*

Proof. Let $f \in \mathcal{F}$. For all $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$, we have

$$(y - f(x))^2 - (y - g(x))^2 = (f(x) - g(x))(f(x) + g(x) - 2y).$$

Moreover, since $\mathbb{E}(Y|X) = g(X)$, we obtain

$$\mathbb{E}((Y - f(X))^2 - (Y - g(X))^2|X) = (f(X) - g(X))^2 \geq 0.$$

Thus

$$\mathbb{E}((Y - f(X))^2) \geq \mathbb{E}((Y - g(X))^2).$$

□

Given an RKHS, regression can be performed in the same manner than SVM, which is usually called kernel ridge regression (KRR):

$$\underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \quad \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n (Y_i - (h(X_i) + b))^2. \quad (\text{P13})$$

1.5.2 Least absolute deviations

Let $g: \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function and let us consider the model:

$$Y = g(X) + \epsilon,$$

where

- ◇ $X \in \mathbb{R}^d$ is a random vector;
- ◇ $\epsilon \in \mathbb{R}$ is a random variable, such that $\epsilon \in L^1$ and $\mathbb{P}(\epsilon \geq 0|X) = \frac{1}{2}$;
- ◇ $g(X) \in L^1$.

Let $\mathcal{F} = \{f: \mathbb{R}^d \rightarrow \mathbb{R}, f(X) \in L^1\}$.

Theorem 41. *The function g is a minimizer of the least absolute deviations risk functional $f \in \mathcal{F} \mapsto \mathbb{E}(|Y - f(X)|)$ over \mathcal{F} .*

The proof is a good exercise.

1.5.3 Support vector regression

Even though quite natural, regression with support vector machines was originally introduced thanks to the so called ϵ -insensitive:

$$\ell_{\epsilon}: x \in \mathbb{R} \mapsto \max(0, |x| - \epsilon).$$

The parameter ϵ enforces the notion of support vectors.

The resulting estimator $f_n^* = h_n^* + b_n^*$ is called support vector regression (SVR) and defined by

$$(h_n^*, b_n^*) \in \arg \min_{\substack{h \in \mathcal{H}, \|h\|_{\mathcal{H}} \leq c \\ b \in \mathbb{R}, f = h + b}} \frac{1}{n} \sum_{i=1}^n \ell_{\epsilon}(Y_i - f(X_i)). \quad (1.4)$$

where $c > 0$. Numerically, ones solves

$$\underset{h \in \mathcal{H}, b \in \mathbb{R}}{\text{minimize}} \quad \frac{\lambda}{2} \|h\|_{\mathcal{H}}^2 + \frac{1}{n} \sum_{i=1}^n \ell_{\epsilon}(Y_i - (h(X_i) + b)), \quad (\text{P14})$$

where $\lambda \geq 0$.

Let us consider $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq c\}$ be the class of hypotheses.

Theorem 42. Let $\phi: \mathbb{R} \rightarrow \mathbb{R}$ be a convex and L_ϕ -Lipschitz continuous loss function and f_n^* be defined by:

$$f_n^* \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \phi(Y_i - f(X_i)).$$

Assume that there exists $B > 0$ such that $\sup_{f \in \mathcal{F}} \phi(Y - f(X)) \leq B$ almost surely and let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$,

$$\mathbb{E}[\phi(Y - f_n^*(X)) | X_1, \dots, X_n] - \inf_{f \in \mathcal{F}} \{\mathbb{E}(\phi(Y - f(X)))\} \leq 8L_\phi \mathbb{E} R_n(\mathcal{F}(X_1^n)) + 2B\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. Refer to Gérard Biau's class:

$$\begin{aligned} & \mathbb{E}(\phi(Y - f_n^*(X)) | X_1, \dots, X_n) - \inf_{f \in \mathcal{F}} \{\mathbb{E}(\phi(Y - f(X)))\} \\ & \leq 2 \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \phi(Y_i - f(X_i)) - \mathbb{E}(\phi(Y - f(X))) \right| \\ & \leq 4 \mathbb{E} R_n((\phi \circ \mathcal{F})(X, Y)_1^n) + 2B\sqrt{\frac{\log(1/\delta)}{2n}} \\ & \leq 8L_\phi \mathbb{E} R_n(\mathcal{F}(X_1^n)) + 2B\sqrt{\frac{\log(1/\delta)}{2n}}. \end{aligned}$$

□

Corollary 43 (Generalization bound). Let f_n^* be defined by:

$$f_n^* \in \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \ell_\epsilon(Y_i - f(X_i)).$$

Let us assume that the kernel k is bounded and let $\kappa > 0$ be an upper bound of it : $\sup_{x \in \mathcal{X}} k(x, x) \leq \kappa^2$. Assume also that Y is almost surely bound by $B > 0$ and let $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$,

$$\mathbb{E}[\ell_\epsilon(Y - f_n^*(X)) | X_1, \dots, X_n] - \inf_{f \in \mathcal{F}} \{\mathbb{E}(\ell_\epsilon(Y - f(X)))\} \leq 8\frac{c\kappa}{\sqrt{n}} + 2(B + c\kappa)\sqrt{\frac{\log(1/\delta)}{2n}}.$$

Proof. We apply the previous theorem: $\phi = \ell_\epsilon$, so $L_\phi = 1$. In addition

$$\begin{aligned}
\forall (x, y) \in \mathcal{X} \times [-B, B], \forall f \in \mathcal{F}: \ell_\epsilon(y - f(x)) &\leq |y| + |f(x)| \\
&\leq B + |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \\
&\leq B + \|f\|_{\mathcal{H}} \|k(\cdot, x)\|_{\mathcal{H}} \quad \text{by Cauchy-Schwarz} \\
&\leq B + c\sqrt{\langle k(\cdot, x), k(\cdot, x) \rangle_{\mathcal{H}}} \\
&\leq B + c\sqrt{k(x, x)} \\
&\leq B + c\kappa.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
R_n(\mathcal{F}(X_1^n)) &\leq \frac{c}{n} \sqrt{\sum_{i=1}^n k(X_i, X_i)} \\
&\leq \frac{c}{n} \sqrt{n\kappa^2} \\
&\leq \frac{c\kappa}{\sqrt{n}}.
\end{aligned}$$

□

Now, we focus on deriving a dual to (P14).

Lemma 44. *One has*

$$\forall x \in \mathbb{R}: \quad \max(0, |x| - \epsilon) = \inf_{(\xi^+, \xi^-) \in \mathbb{R}_+ \times \mathbb{R}_+ : -\xi^- - \epsilon \leq x \leq \xi^+ + \epsilon} \xi^+ + \xi^-.$$

The proof is a good exercise.

Now, let $C = 1/(\lambda n)$ and $K = (k(X_i, X_j))_{1 \leq i, j \leq n}$ be the kernel matrix. Then, (P14) can be rewritten equivalently with slack variables (rescaling the objective function):

$$\begin{aligned}
&\underset{\substack{h \in \mathcal{H}, b \in \mathbb{R} \\ \xi \in \mathbb{R}^n}}{\text{minimize}} \quad \frac{1}{2} \|h\|_{\mathcal{H}}^2 + C \sum_{i=1}^n (\xi_i^+ + \xi_i^-) \\
&\text{s. t.} \quad \begin{cases} Y_i - (h(X_i) + b) \leq \xi_i^+ + \epsilon \\ (h(X_i) + b) - Y_i \leq \xi_i^- + \epsilon \\ \forall i \in [n], \xi_i^+ \geq 0, \xi_i^- \geq 0. \end{cases} \quad (\text{P15})
\end{aligned}$$

A dual of (P15) is:

$$\begin{aligned}
&\underset{\alpha^+ \in \mathbb{R}^n, \alpha^- \in \mathbb{R}^n}{\text{maximize}} \quad -\frac{1}{2} (\alpha^+ - \alpha^-)^\top K (\alpha^+ - \alpha^-) + y^\top (\alpha^+ - \alpha^-) - \epsilon \mathbf{1}^\top (\alpha^+ + \alpha^-) \\
&\text{s. t.} \quad \begin{cases} \forall i \in [n]: 0 \leq \alpha_i^+ \leq C, 0 \leq \alpha_i^- \leq C \\ \mathbf{1}^\top (\alpha^+ - \alpha^-) = 0. \end{cases} \quad (\text{P16})
\end{aligned}$$

1.6 Other methods

1.6.1 k-nearest neighbors

The principle of the k-nearest neighbor method is to estimate directly the Bayes classifier thanks to estimations of $\mathbb{P}(Y = j|X)$. There are several manners to do so, the simplest one is

$$\mathbb{P}(Y = j|X = x) \approx \frac{nV}{\sum_{i=1}^n \mathbf{1}_{X_i \in \mathcal{V}_x}} \frac{\sum_{i=1}^n \mathbf{1}_{Y_i=j, X_i \in \mathcal{V}_x}}{nV} = \frac{1}{\sum_{i=1}^n \mathbf{1}_{X_i \in \mathcal{V}_x}} \sum_{i=1}^n \mathbf{1}_{Y_i=j, X_i \in \mathcal{V}_x},$$

where \mathcal{V}_x is a neighborhood of x of volume V . Unfortunately, a usual ball of prescribed radius is useless since there may be regions of the space where this neighborhood is empty, thus giving an infinite estimator of the probability.

To circumvent this problem, the neighborhood is chosen as the k-nearest neighbors of x .

There exist of course several other manners to define a suitable neighborhood:

smoothing : the naive choice is to choose for \mathcal{V}_x an ℓ_2 -ball of radius ϵ . Then, $\mathbf{1}_{X_i \in \mathcal{V}_x} = \mathbf{1}_{\|X_i - x\|_{\ell_2} \leq \epsilon}$,

where $\mathbf{1}_{\leq \epsilon}$ can be approximated by a smooth function: $\alpha \mapsto e^{-\alpha^2/2}$. This boils down to use a kernel method to estimate $\mathbb{P}(Y = j|X = x)$ and leads to the weighted version of the k-nearest neighbor method.

partitioning : \mathcal{V}_x can be chosen such that $\cup_{i=1}^n \mathcal{V}_{X_i} = \mathbb{R}^d$, in other words such that neighborhoods make a partition of the entire space. This is what is done by decision trees, the cells of which consist in hypercubes of \mathbb{R}^d .

At the end of the day, the k-nearest neighbors rule consists in predicting, for $x \in \mathbb{R}^d$, the majority vote (for classification, or the mean for regression) of the k-nearest neighbors of x . Formally, the predicted class is:

$$g(x) \in \arg \max_{y \in \mathcal{Y}} \sum_{i=1}^k \mathbf{1}_{Y_{(i)}=y},$$

where the ranked labeled $\{Y_{(1)}, \dots, Y_{(n)}\}$ are such that $X_{(1)} - x \leq \dots \leq X_{(n)} - x$.

For the smoothed (or weighted) version of k-nearest neighbors, the vote of each neighbor is considered in the prediction, but weighted (generally) by $e^{-\gamma X_{(i)} - x^2}$. The new classification rule becomes:

$$g_\gamma(x) \in \arg \max_{y \in \mathcal{Y}} \sum_{i=1}^k e^{-\gamma X_{(i)} - x^2} \mathbf{1}_{Y_{(i)}=y}.$$

1.6.2 Decision trees

Decision trees, and in particular classification and regression trees (CART), are supervised estimators introduced by Leo Breiman et al. The paradigm of a binary decision tree is to recursively split the space \mathcal{X} with simple rules such that: is the explicative variable x_j greater than the threshold τ or not? Doing so, a decision tree is built, for which each node corresponds to a simple rule (and secondarily to a partition cell of \mathcal{X}). The final result is a partition of \mathcal{X} by hypercubes.

At each step of the learning algorithm,

1. consider the partition $\mathcal{P} = \{\mathcal{X}\}$;
2. for each cell \mathcal{A} of \mathcal{P} , define the two-cell partition $\mathcal{A} = \mathcal{L}_{j,\tau} \cup \mathcal{R}_{j,\tau}$, where $j \in [d]$ is a feature index and $\tau \in \mathbb{R}$ is a threshold, and

$$\begin{cases} \mathcal{L}_{j,\tau} = \{x \in \mathcal{A} : x_j \leq \tau\} \\ \mathcal{R}_{j,\tau} = \{x \in \mathcal{A} : x_j > \tau\} \end{cases}$$

are the "left" and "right" parts of \mathcal{A} . Then, find the best pair (feature, threshold) for splitting:

$$(j, \tau) \in \arg \min_{\substack{1 \leq j \leq d \\ \tau \in \mathbb{R}}} \frac{|\mathcal{L}_{j,\tau}|}{|\mathcal{A}|} D(\mathcal{L}_{j,\tau}) + \frac{|\mathcal{R}_{j,\tau}|}{|\mathcal{A}|} D(\mathcal{R}_{j,\tau})$$

where D is a distortion measure for a cell (see below);

3. replace \mathcal{A} by $\mathcal{L}_{j,\tau}$ and $\mathcal{R}_{j,\tau}$ in the partition \mathcal{P} ;
4. go to 2.

Given a cell \mathcal{A} , one may define the ratio of observations of \mathcal{A} of class $y \in \mathcal{Y}$:

$$p_y(\mathcal{A}) = \frac{|\{i \in [n] : X_i \in \mathcal{A}, Y_i = y\}|}{|\mathcal{A}|}.$$

Then, the distortion of the cell \mathcal{A} may be:

- ◇ Gini impurity: $D(\mathcal{A}) = \sum_{y \in \mathcal{Y}} p_y(\mathcal{A})(1 - p_y(\mathcal{A}))$ (classification);
- ◇ entropy: $D(\mathcal{A}) = - \sum_{y \in \mathcal{Y}} p_y(\mathcal{A}) \log(p_y(\mathcal{A}))$ (classification);
- ◇ mean squared error: $D(\mathcal{A}) = \frac{1}{|\mathcal{A}|} \sum_{\substack{1 \leq i \leq n \\ X_i \in \mathcal{A}}} \left(Y_i - \bar{Y}_{\mathcal{A}} \right)^2$, with $\bar{Y}_{\mathcal{A}} = \frac{1}{|\mathcal{A}|} \sum_{\substack{1 \leq i \leq n \\ X_i \in \mathcal{A}}} Y_i$ (regression).

Remark 1.6.1. *Gini impurity and random prediction* The Gini impurity corresponds the error obtained when producing a random label according to the empirical distribution of labels in the given cell \mathcal{A} .

Consider a binary classification problem with proportion of labels 1 $\pi = \mathbb{P}(Y = 1 | X \in \mathcal{A})$ in the cell \mathcal{A} of interest. Let g be a classifier with $g(X) | X \in \mathcal{A} \stackrel{d}{=} Z$, where $Z \sim \mathcal{B}(\pi)$. Then

$$\begin{aligned} \mathbb{P}(Y \neq g(X) | X \in \mathcal{A}) &= \mathbb{P}(Y \neq Z | X \in \mathcal{A}) \\ &= \mathbb{P}(Y = 1 \& Z \neq 1 | X \in \mathcal{A}) + \mathbb{P}(Y \neq 1 \& Z = 1 | X \in \mathcal{A}) \\ &= \mathbb{P}(Y = 1 | X \in \mathcal{A}) \mathbb{P}(Z \neq 1) + \mathbb{P}(Y \neq 1 | X \in \mathcal{A}) \mathbb{P}(Z = 1) \\ &= \pi(1 - \pi) + (1 - \pi)\pi \\ &= 2\pi(1 - \pi). \end{aligned}$$

For regression, Jerome Friedman suggested an improved criterion (in its original paper tackling gradient boosting), referred to as Friedman's mean squared error:

$$(j, \tau) \in \arg \min_{\substack{1 \leq j \leq d \\ \tau \in \mathbb{R}}} \frac{|\mathcal{L}_{j,\tau}| |\mathcal{R}_{j,\tau}|}{|\mathcal{L}_{j,\tau}| + |\mathcal{R}_{j,\tau}|} \left(\bar{Y}_{\mathcal{L}_{j,\tau}} - \bar{Y}_{\mathcal{R}_{j,\tau}} \right)^2.$$

Last but not least, several stopping rules are of interests:

- ◇ maximal depth of the tree;
- ◇ minimal number of observations required to split an internal node;
- ◇ minimal number of observations required to be at a leaf node;
- ◇ maximal number of leaf nodes.

1.6.3 Bagging

Bagging is a portmanteau word for *bootstrap aggregating*. The paradigm of bagging is to train independently several base classifiers (g_1, \dots, g_T) , with $g_t: \mathbb{R}^d \rightarrow \{\pm 1\}$, and to build a new classifier by averaging the predictions of the base classifiers:

$$g_n^T(x) = \text{sign} \left(\frac{1}{T} \sum_{t=1}^T g_t(x) \right).$$

Doing so, the variance of the prediction is reduced and so it is for the global error. The requirements for such a result are:

- ◇ base classifiers should be more accurate than chance;
- ◇ base classifiers should be estimated independently from each other.

In practice, base classifiers are trained *quasi-independently* by bootstrapping the training set.

Bagging is also valid for multiclass problems: for C classes, the prediction is:

$$g_n^T(x) = \arg \max_{1 \leq j \leq C} \frac{1}{T} \sum_{t=1}^T g_t(x) \mathbf{1}_{g_t(x)=j} = \arg \max_{1 \leq j \leq C} \text{card} \left(\{t \in [T] : g_t(x) \mathbf{1}_{g_t(x)=j}\} \right),$$

where $g_t: \mathbb{R}^d \rightarrow [C]$, which corresponds to the majority vote since base classifiers are equally weighted.

Finally, one may also bag regressors $g_t: \mathbb{R}^d \rightarrow \mathbb{R}$ by a simple averaging:

$$g_n^T(x) = \frac{1}{T} \sum_{t=1}^T g_t(x).$$

1.6.4 Random forests

Random forests are bagged trees: for binary classification, a random forest is

$$g_n^T(x) = \text{sign} \left(\frac{1}{T} \sum_{t=1}^T g_t(x) \right),$$

where the base classifiers (g_1, \dots, g_T) , with $g_t: \mathbb{R}^d \rightarrow \{\pm 1\}$, are learned quasi-independently by bootstrap.

However, in order to enforce the independent learning, each decision tree g_t owns an additional randomization step in its learning procedure:

1. at each cell, select a subset of features at random;
2. find the best pair (feature, threshold) for splitting.

1.7 Exercises

1.7.1 Discriminant analysis

Exercise 1.1 (MLE in the Gaussian model (proof of Proposition 1)). Let $\mu^* \in \mathbb{R}^d$, Σ^* be a PD matrix and $\{X_1, \dots, X_n\}$ be a sample *iid* according to $\mathcal{N}(\mu^*, \Sigma^*)$.

1. Show that $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ is the MLE of μ^* .
2. Keeping in mind that for any square matrices A and B , $\frac{\partial}{\partial A} \text{tr}(AB) = B^\top$ and $\frac{\partial}{\partial A} \log(|A|) = (A^{-1})^\top$, show that $\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^\top$ is the MLE of Σ^* .

Exercise 1.2 (LDA (proof of Proposition 4)). Consider LDA with means μ_1 and μ_{-1} , covariance denoted Σ and prior probability $\pi = \mathbb{P}(Y = 1)$. Show that $g^*: x \in \mathbb{R}^d \mapsto \text{sign}(w^\top x + b)$, where

$$\begin{cases} w &= \Sigma^{-1}(\mu_1 - \mu_{-1}) \\ b &= \frac{1}{2}(\mu_{-1}^\top \Sigma^{-1} \mu_{-1} - \mu_1^\top \Sigma^{-1} \mu_1) + \log\left(\frac{\pi}{1-\pi}\right) \end{cases}$$

is the Bayes classifier.

Exercise 1.3 (QDA (proof of Proposition 5)). Consider QDA with means μ_1 and μ_{-1} , covariances Σ_1 and Σ_{-1} , and prior probability $\pi = \mathbb{P}(Y = 1)$. Show that $g^*: x \in \mathbb{R}^d \mapsto \text{sign}(\frac{1}{2}x^\top Qx + w^\top x + b)$, where

$$\begin{cases} Q &= \Sigma_{-1}^{-1} - \Sigma_1^{-1} \\ w &= \Sigma_1^{-1} \mu_1 - \Sigma_{-1}^{-1} \mu_{-1} \\ b &= \frac{1}{2}(\mu_{-1}^\top \Sigma_{-1}^{-1} \mu_{-1} - \mu_1^\top \Sigma_1^{-1} \mu_1) - \frac{1}{2} \log\left(\frac{|\Sigma_1|}{|\Sigma_{-1}|}\right) + \log\left(\frac{\pi}{1-\pi}\right) \end{cases}$$

is the Bayes classifier.

1.7.2 Boosting

Exercise 1.4 (Adaboost (proof of Lemma 12)). In Adaboost, assume that there exists $\gamma \in (0, 1/2)$ such that $\forall t \in [T]$, $\epsilon_t \leq \frac{1}{2} - \gamma$ almost surely and let $f_T: \mathcal{X} \rightarrow \mathbb{R}$ be the classifier at the last iteration.

1. Show that:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \frac{f_T(X_i)}{\|w\|_{\ell_1}} < \gamma} \leq 2^T \prod_{t=1}^T \sqrt{\epsilon_t^{1-\gamma} (1 - \epsilon_t)^{1+\gamma}}.$$

2. Analyze the behavior of $x \in [0, \frac{1}{2}] \mapsto \log(x^{1-\nu}(1+x)^{1+\nu})$. Deduce that:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \frac{f_T(X_i)}{\|w\|_{\ell_1}} < \nu} \leq 2^T \left(\left(\frac{1}{2} - \nu \right)^{1-\nu} \left(\frac{1}{2} + \nu \right)^{1+\nu} \right)^{T/2}.$$

3. Conclude that:

$$\frac{1}{n} \sum_{i=1}^n \mathbf{1}_{Y_i \frac{f_T(X_i)}{\|w\|_{\ell_1}} < \nu} \leq \left[\left((1-2\nu)^{1-\nu} (1+2\nu)^{1+\nu} \right) \right]^{T/2}.$$

Exercise 1.5 (Adaboost). In Adaboost, show that the error made by a weak learner at the future iteration is $\frac{1}{2}$, i.e. at each iteration $t > 0$:

$$\sum_{i=1}^n D_{t+1}(i) \mathbf{1}_{Y_i \neq g_t(X_i)} = \frac{1}{2}.$$

Exercise 1.6 (Convergence of gradient boosting (proof of Theorem 14)). Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$. We consider the optimization problem

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \ f(x),$$

and the iterative algorithm with iteration:

$$\begin{cases} d_t = P(\nabla f(x_t)) \\ x_{t+1} = x_t - \eta d_t, \end{cases}$$

where $P : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is an operator and $\eta > 0$ is a step size defined later.

Let us assume that:

(H1) f is differentiable and ∇f is L -Lipschitz continuous ($L > 0$);

(H2) f is μ -strongly convex ($\mu > 0$). This implies that

$$\forall x, x' \in \mathbb{R}^d : \quad f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle_{\ell_2} + \frac{\mu}{2} \|x' - x\|_{\ell_2}^2;$$

(H3) f has a minimizer denoted x^* ;

(H4) there exists $\gamma \in [0, 1]$ such that:

$$\forall y \in \mathbb{R}^d : \quad \|y - P(y)\|_{\ell_2}^2 \leq (1 - \gamma) \|y\|_{\ell_2}^2.$$

1. Knowing that a differentiable function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is convex if and only if

$$\forall x, x' \in \mathbb{R}^d : \quad (\nabla f(x) - \nabla f(x'))^\top (x - x') \geq 0,$$

show that $g = \frac{1}{2} \|\cdot\|_{\ell_2}^2 - f$ is convex.

2. Show that

$$\forall x, x' \in \mathbb{R}^d : \quad f(x') \leq f(x) + \langle \nabla f(x), x' - x \rangle_{\ell_2} + \frac{L}{2} \|x' - x\|_{\ell_2}^2.$$

3. Show that

$$\forall x \in \mathbb{R}^d : \quad \|\nabla f(x)\|_{\ell_2}^2 \geq 2\mu(f(x) - f(x^*)).$$

4. Show that for each iteration $t \geq 0$:

$$f(x_{t+1}) \leq f(x_t) - \frac{\gamma\eta}{2} \|\nabla f(x_t)\|_{\ell_2}^2 - \frac{\eta}{2}(1 - L\eta) \|d_t\|_{\ell_2}^2.$$

5. Choosing $\eta = \frac{1}{2L}$ and defining $\Delta_t = f(x_t) - f(x^*)$, show that:

$$\Delta_{t+1} \leq \left(1 - \frac{\gamma\mu}{2L}\right) \Delta_t.$$

6. Conclude on the linear convergence of the iterate:

$$f(x_t) - f(x^*) \leq \left(1 - \frac{\gamma\mu}{2L}\right)^t (f(x_0) - f(x^*)).$$

1.7.3 SVM

Exercise 1.7 (Distance to a hyperplane (proof of Proposition 16)). Let $(w, b) \in \mathbb{R}^d \setminus \{0\} \times \mathbb{R}$ and $\mathbb{H} = \{z \in \mathbb{R}^d : w^\top z + b = 0\}$. Show that the distance between \mathbb{H} and any point $x \in \mathbb{R}^d$ is

$$d(\mathbb{H}, x) = \frac{|w^\top x + b|}{\|w\|_{\ell_2}}.$$

Exercise 1.8 (Kernel trick). Let $\{(X_i, Y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \{\pm 1\}$ be an *iid* sample and for $j \in \{\pm 1\}$, $\hat{\mu}_j = \frac{1}{\sum_{i=1}^n \mathbf{1}_{Y_i=j}} \sum_{i=1}^n \mathbf{1}_{Y_i=j} X_i$ be the center of class j . Show that the kernel trick can be applied to the classification rule:

$$\forall x \in \mathbb{R}^d : \quad g(x) = \begin{cases} 1 & \text{if } \|x - \hat{\mu}_1\|_{\ell_2} < \|x - \hat{\mu}_{-1}\|_{\ell_2} \\ -1 & \text{otherwise.} \end{cases}$$

Exercise 1.9 (Techniques for constructing kernels). Given valid kernels k_1 and k_2 on $\mathbb{R}^d \times \mathbb{R}^d$, show that the functions k defined below are still kernels:

1. $\forall x, x' \in \mathbb{R}^d : k(x, x') = ck_1(x, x')$, where $c > 0$.
2. $\forall x, x' \in \mathbb{R}^d : k(x, x') = k_1(x, x')f(x)f(x')$, where $f \in \mathbb{R}^{\mathbb{R}^d}$.
3. $\forall x, x' \in \mathbb{R}^d : k(x, x') = \exp(k_1(x, x'))$.

4. $\forall x, x' \in \mathbb{R}^d : k(x, x') = k_1(x, x') + k_2(x, x')$.
5. $\forall x, x' \in \mathbb{R}^d : k(x, x') = k_1(x, x')k_2(x, x')$.
6. $\forall x, x' \in \mathbb{R}^d : k(x, x') = q(k_1(x, x'))$, where $q \in \mathbb{R}^{\mathbb{R}}$ is a polynomial with nonnegative coefficients.
7. $\forall x, x' \in \mathbb{R}^d : k(x, x') = k_1(\varphi(x), \varphi(x'))$, where $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^p$.
8. $\forall x, x' \in \mathbb{R}^d : k(x, x') = x^\top A x'$, where $A \in \mathbb{R}^{d \times d}$ is positive semi-definite (PSD).
9. $\forall x, x' \in \mathbb{R}^d : k(x, x') = \exp \left(-\|x - x'\|_{\ell_2}^2 \right)$.

Exercise 1.10 (Kernelized regression). Let $\{(X_i, Y_i)\}_{1 \leq i \leq n} \subset \mathbb{R}^d \times \mathbb{R}$ be an *iid* sample, \mathcal{H} be the RKHS associated to a kernel k on $\mathbb{R}^d \times \mathbb{R}^d$. Let us consider the optimization problem:

$$\begin{aligned} & \underset{h \in \mathcal{H}, \xi \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|h\|_{\mathcal{H}}^2 + \frac{C}{2} \|\xi\|_{\ell_2}^2 \\ & \text{s. t.} \quad \forall i \in [n] : Y_i - h(X_i) - \xi_i = 0, \end{aligned} \tag{P17}$$

where $C > 0$.

1. Exhibit the dual problem to (P17).
2. Let (h^*, ξ^*) and α^* be solutions respectively to (P17) and its dual. Justify and exhibit the link between these quantities.
3. Determine the value of α^* .

Now, we focus on the unconstrained version on (P17):

$$\underset{h \in \mathcal{H}, \xi \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|h\|_{\mathcal{H}}^2 + \frac{C}{2} \sum_{i=1}^n (Y_i - h(X_i))^2. \tag{P18}$$

4. Let h^* be a solution to (P18). Justify that there exists $\beta^* \in \mathbb{R}^n$ such that $h^* = \sum_{i=1}^n \beta_i^* k(\cdot, X_i)$. Deduce a parametric version of (P18).
5. Determine the value of β^* .

1.7.4 Regression

Exercise 1.11 (Robust regression (proof of Theorem 41)). Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be a measurable function and let us consider the model $Y = g(X) + \epsilon$, where $X \in \mathbb{R}^d$ is a random vector, $g(X) \in L^1$ and $\epsilon \in \mathbb{R}$ is a random variable, such that $\epsilon \in L^1$ and $\mathbb{P}(\epsilon \geq 0 | X) = \frac{1}{2}$. Let also $\mathcal{F} = \{f : \mathbb{R}^d \rightarrow \mathbb{R}, f(X) \in L^1\}$.

1. Show that $\forall x \in \mathbb{R}, |x| = (1 - 2\mathbf{1}_{x < 0})x$ and deduce that for all $f \in \mathcal{F}$,

$$|Y - g(X)| = (1 - 2\mathbf{1}_{Y - g(X) < 0})(f(X) - g(X)) + (1 - 2\mathbf{1}_{Y - g(X) < 0})(Y - f(X)).$$

2. Deduce that for all $f \in \mathcal{F}$,

$$|Y - f(X)| - |Y - g(X)| = 2(Y - f(X))(\mathbf{1}_{Y - g(X) < 0} - \mathbf{1}_{Y - f(X) < 0}) - (1 - 2\mathbf{1}_{Y - g(X) < 0})(f(X) - g(X)).$$

3. Show that $(Y - f(X))(\mathbf{1}_{Y-g(X) < 0} - \mathbf{1}_{Y-f(X) < 0}) \geq 0$.
4. Deduce that $\mathbb{E}[|Y - f(X)| - |Y - g(X)| | X] \geq 0$ and then that g is a minimizer of $f \in \mathcal{F} \mapsto \mathbb{E}(|Y - f(X)|)$ over \mathcal{F} .

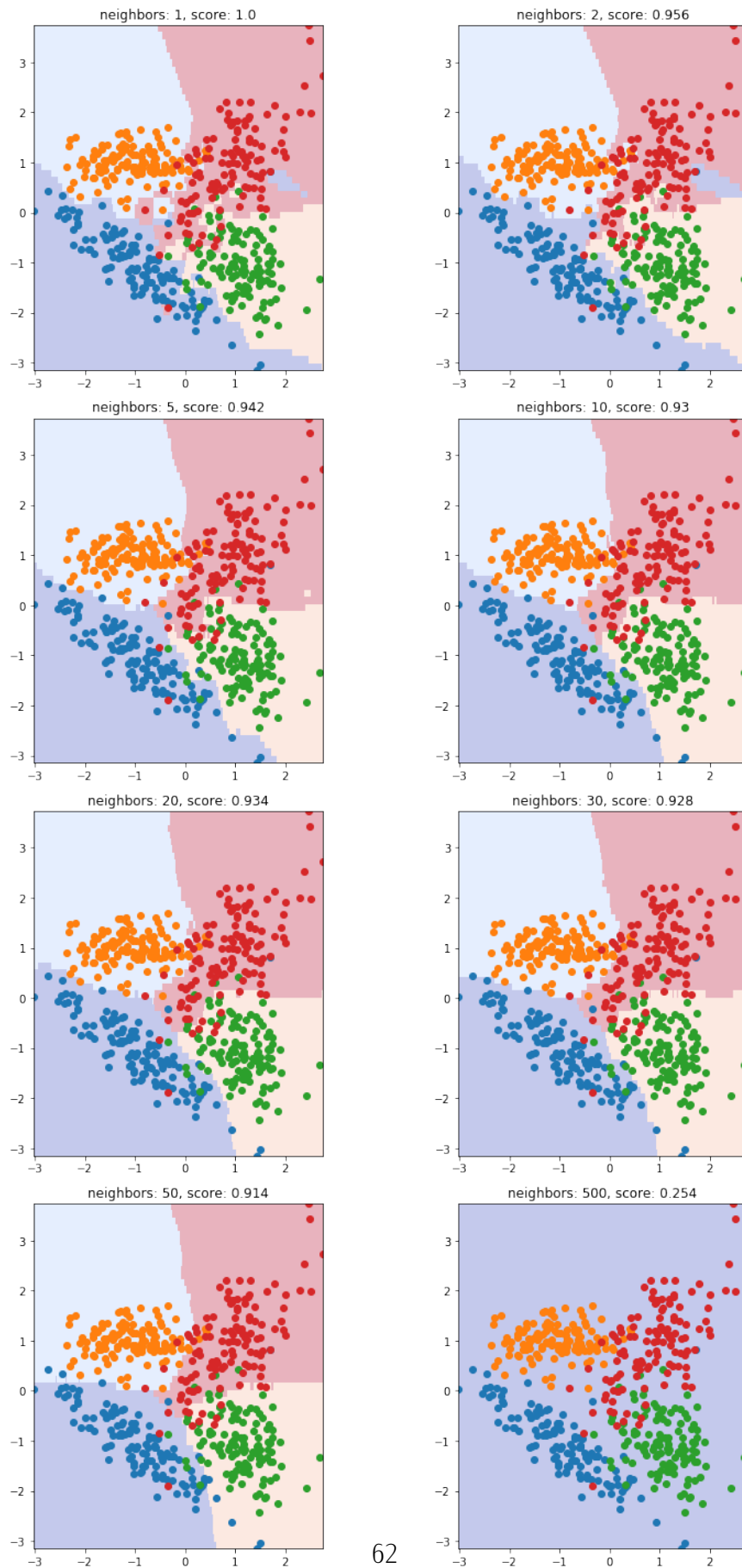


Figure 1.9: Example of classification frontier with a nearest neighbors.

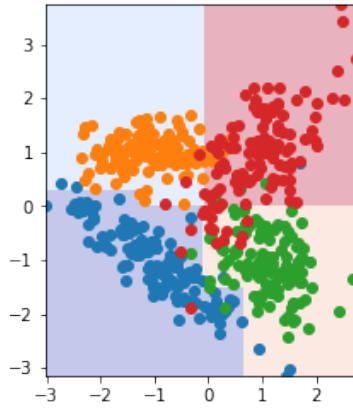


Figure 1.10: Example of classification frontier with a decision tree.

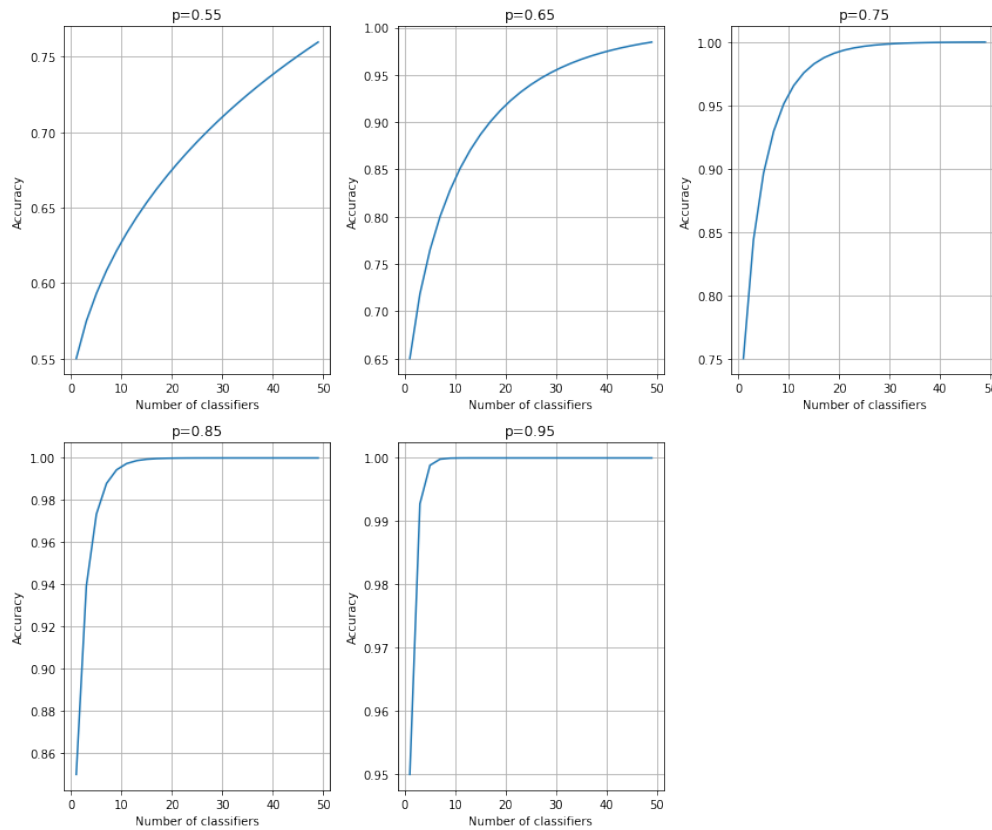


Figure 1.11: Classification accuracy when bagging independent weak classifiers with same error probability $1 - p$.

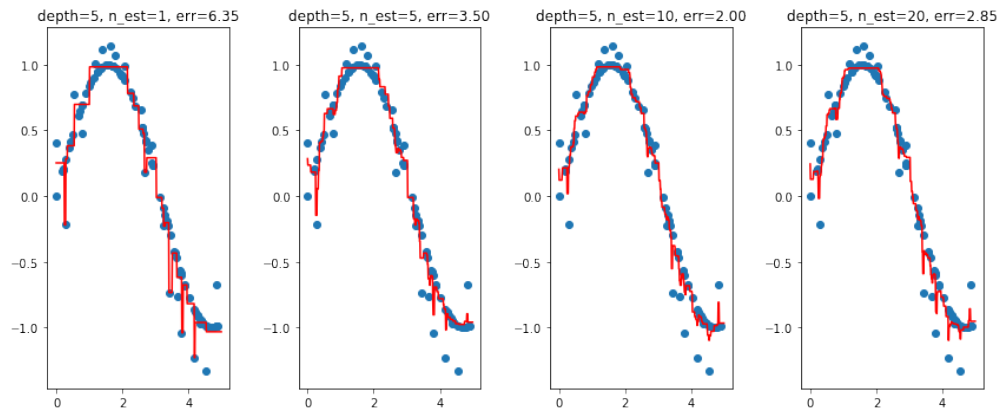


Figure 1.12: Example of regression with bagging trees.

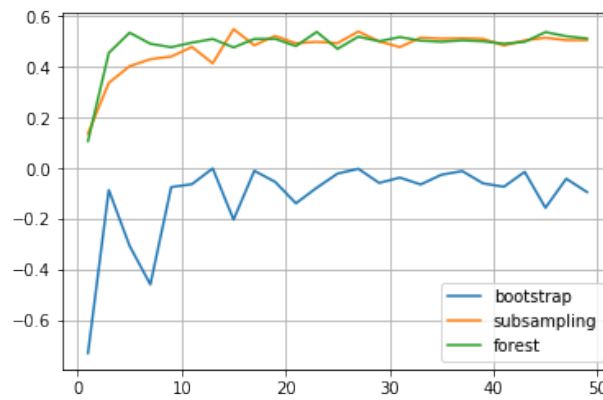


Figure 1.13: Regression accuracy for the diabetes dataset.

Chapter 2

Clustering

Similarly to classification, we are interested in a pair of random variables $(X, Y) \in \mathbb{R}^d \times [k]$, where k is a positive integer, and we wish to predict Y given X . Yet, in clustering, we only observe X and we have no prior information concerning Y . Therefore, there is no reason not to consider permutations of its values. As a consequence, clustering is more interested in partitioning the space \mathbb{R}^d than predicting a prescribed label.

Then, the aim of clustering would be to produce a partition $\{C_1, \dots, C_k\}$, such that there exists a permutation $\sigma: [k] \rightarrow [k]$ for which, for all $j \in [k]$,

$$C_j = \{x \in \mathbb{R}^d : \mathbb{P}(Y = \sigma(j)|X = x) \geq \mathbb{P}(Y = \ell|X = x), \forall \ell \in [k]\}.$$

Nevertheless, the difficulty of clustering is two-fold: we do not observe Y and we do not know the number of classes k . In fact, it is easy to understand that the problem of clustering, as defined just above, is ill-posed. Indeed, many joint distributions on (X, Y) may produce the marginal distribution of X , that we observe. This is true for several number of classes and several meanings of them.

Example 2.0.1. Consider two pairs of random variables (X, Y) and (\tilde{X}, \tilde{Y}) in $[-1, 1] \times \{\pm 1\}$ such that

$$\begin{cases} X|Y = 1 \sim \mathcal{U}([0, 1]) \\ X|Y = -1 \sim \mathcal{U}([-1, 0]) \\ \mathbb{P}(Y = 1) = \frac{1}{2} \end{cases} \quad \text{and} \quad \begin{cases} \tilde{X}|\tilde{Y} = 1 \sim \mathcal{U}([-\frac{1}{2}, 1]) \\ \tilde{X}|\tilde{Y} = -1 \sim \mathcal{U}([-1, -\frac{1}{2}]) \\ \mathbb{P}(\tilde{Y} = 1) = \frac{3}{4}. \end{cases}$$

Then, it is straightforward that $X \sim \mathcal{U}([-1, 1])$ and $\tilde{X} \sim \mathcal{U}([-1, 1])$.

As a consequence, a concrete, but vague, definition of clustering is to *organize the data in some meaningful way*. This often means creating clusters (or a partition) such that:

1. points inside a cluster are similar (this corresponds to a mode in the marginal distribution of X);
2. points in separated clusters are dissimilar (this corresponds to the existence of a frontier of low density).

In this context, k describes the number of clusters to discover. It may be fixed beforehand or chosen automatically by a low-density assumption.

In this chapter, we describe several methods of clustering, from a statistical modeling point of view to heuristic approaches.

2.1 Gaussian mixtures

2.1.1 Mixture model

Definition 2.1.1 (Mixture model). Let $\{P_\lambda = f_\lambda \cdot \mu : \lambda \in \Lambda\}$ be a statistical model dominated by a measure μ , m be a positive integer, $(\lambda_1, \dots, \lambda_m) \in \Lambda^m$ and (π_1, \dots, π_m) be a probability vector:

$$\forall j \in [m]: \pi_j \geq 0 \quad \sum_{j=1}^m \pi_j = 1.$$

Then, the distribution

$$\left(\sum_{j=1}^m \pi_j f_{\lambda_j} \right) \cdot \mu = \sum_{j=1}^m \pi_j P_{\lambda_j}$$

is called a mixture model.

Proposition 45 (Latent variable). Let $\{P_\lambda = f_\lambda \cdot \mu : \lambda \in \Lambda\}$ be a statistical model (for random variables in \mathbb{R}^d) dominated by a measure μ , m be a positive integer, $(\lambda_1, \dots, \lambda_m) \in \Lambda^m$ and (π_1, \dots, π_m) be a probability vector.

Let now $Z \sim \mathcal{M}(1, \pi_1, \dots, \pi_m)$ be a multinomial variable^a and $Y = \sum_{j=1}^m j \mathbf{1}_{Z_j=1}$. Then

$$\forall j \in [m]: \mathbb{P}(Y = j) = \pi_j.$$

In addition, let X be a random variable such that $X|Y \sim P_{\lambda_Y}$. Then $X \sim \sum_{j=1}^m \pi_j P_{\lambda_j}$.

^a $\forall (z_1, \dots, z_m) \in \{0, 1\}^m : \sum_{j=1}^m z_j = 1, \mathbb{P}(Z = (z_1, \dots, z_m)) = \prod_{j=1}^m \pi_j^{z_j}$.

Proof. By definition of Z , $\exists! j \in [m] : Z_j = 1$. Therefore, $Y \in [m]$ almost surely and $\mathbb{P}(Y = j) = \prod_{\ell=1}^m \pi_\ell^{Z_\ell}$, with $Z_\ell = 1$ if $\ell = j$ and 0 otherwise, meaning that $\mathbb{P}(Y = j) = \pi_j$.

Moreover, let f be the probability density function (pdf) of X , then

$$\forall x \in \mathbb{R}^d: f(x) = \sum_{j=1}^m f_{\lambda_j}(x) \mathbb{P}(Y = j) = \sum_{j=1}^m \pi_j f_{\lambda_j}(x).$$

Thus, X is distributed according to a mixture model defined by $(P_{\lambda_1}, \dots, P_{\lambda_m})$ and (π_1, \dots, π_m) . \square

Proposition 45 bridges the gap between mixture models and clustering: when X is distributed according to a mixture model with k components, we can describe it using k clusters defined by a latent variable $Y \in [k]$. Then, the distribution of $X|Y$ is given by the mixture components.

Conversely, clustering is naturally modeled by a mixture model: clusters are conditional variables $X|Y$ but we only observe X . Thus, we focus on the marginal density of X , which is, by Bayes' theorem:

$$\forall x \in \mathbb{R}^d: f(x) = \sum_{j=1}^k \pi_j f_{\lambda_j}(x),$$

where $\pi_j = \mathbb{P}(Y = j)$ is the prior probability of a cluster. Then, the Bayes rule for clustering is given by

$$g^*: x \in \mathbb{R}^d \mapsto \arg \max_{1 \leq j \leq k} \mathbb{P}(Y = j | X = x) = \arg \max_{1 \leq j \leq k} \pi_j f_{\lambda_j}(x).$$

The final partitioning $\{C_1, \dots, C_k\}$ is such that for all $j \in [k]$, $C_j = \{x \in \mathbb{R}^d : g^*(x) = j\}$.

Remark 2.1.1. Proposition 45 explains how to sample X according to a mixture model. For clarity, this is detailed in Algorithm 4.

Algorithm 4 Sampling of a mixture model.

Input: $(P_{\theta_1}, \dots, P_{\theta_m})$ (mixture components) and (π_1, \dots, π_m) (probability vector).

$z \leftarrow \text{sample from } \mathcal{M}(1, \pi_1, \dots, \pi_m)$ (multinomial variable)

$y \leftarrow \sum_{j=1}^m j \mathbf{1}_{z_j=1}$ (cluster label)

$x \leftarrow \text{sample from } P_{\theta_y}$.

Output: x .

The next step when considering such a probabilistic model for clustering is estimating the parameter $\theta = (\pi_1, \dots, \pi_k, \lambda_1, \dots, \lambda_k) \in \Theta$ of the mixture model, where

$$\Theta = \left\{ \theta = (\pi_1, \dots, \pi_k, \lambda_1, \dots, \lambda_k), \pi \in [0, 1]^k, \mathbf{1}^\top \pi = 1, (\lambda_1, \dots, \lambda_k) \in \Lambda^k \right\}.$$

A good choice would be by MLE: given n iid copies of X , denoted $\{X_i\}_{1 \leq i \leq n}$, we consider the estimator $\hat{\theta} = (\hat{\pi}_1, \dots, \hat{\pi}_k, \hat{\lambda}_1, \dots, \hat{\lambda}_k)$ maximizing the empirical log-likelihood of the statistical model

$$\mathcal{P} = \left\{ \sum_{j=1}^k \pi_j f_{\lambda_j} : \theta = (\pi_1, \dots, \pi_k, \lambda_1, \dots, \lambda_k) \in \Theta \right\}$$

associated to X :

$$\ell_{X_1^n}(\theta) = \ell_{X_1^n}(\pi_1, \dots, \pi_k, \lambda_1, \dots, \lambda_k) = \sum_{i=1}^n \log \left(\sum_{j=1}^k \pi_j f_{\lambda_j}(X_i) \right).$$

A procedure for maximizing ℓ_n is described in the forthcoming section.

2.1.2 A toy example

Let us describe a very simple (clustering) example in order to motivate the need for the expectation-maximization (EM) algorithm. Let $(X, Y) \in \mathbb{R} \times \{\pm 1\}$ be a pair of random variables such that

$$X|Y \sim \mathcal{N}(\mu_Y^*, 1),$$

$\mu_Y^* \in \mathbb{R}$. In other words, $X|Y$ has density $p_{\mu_Y^*} : x \in \mathbb{R} \mapsto \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu_Y^*)^2}{2}}$.

We assume that we observe a sample (X_1, \dots, X_n) , the variables of which are *iid* copies of X , and we would like to estimate the distribution of X , based on the parametric model induced by the previous assumption.

For this purpose, we observe that the marginal density of X is given, for all $x \in \mathbb{R}$, by:

$$\begin{aligned} m_{\theta^*}(x) &= \mathbb{P}(Y = 1)p_{\mu_1^*}(x) + \mathbb{P}(Y = -1)p_{\mu_{-1}^*}(x) \\ &= \pi_1^* p_{\mu_1^*}(x) + (1 - \pi_1^*) p_{\mu_{-1}^*}(x), \end{aligned}$$

where $\pi_1^* = \mathbb{P}(Y = 1)$ and $\theta^* = (\pi_1^*, \mu_1^*, \mu_{-1}^*)$. Thus, we should consider the statistical model $\{m_\theta : \theta = (\pi_1, \mu_1, \mu_{-1}) \in (0, 1) \times \mathbb{R}^2\}$ to estimate the density of X .

Considering that we would like to estimate θ^* by maximum likelihood, we should maximize, with respect to θ , the marginal log-likelihood

$$\begin{aligned} \ell_{X_1^n}(\theta) &= \sum_{i=1}^n \log(m_\theta(X_i)) \\ &= \sum_{i=1}^n \log(\pi_1 p_{\mu_1}(X_i) + (1 - \pi_1) p_{\mu_{-1}}(X_i)) \\ &= \sum_{i=1}^n \log\left(\frac{\pi_1}{\sqrt{2\pi}} e^{-\frac{(X_i - \mu_1)^2}{2}} + \frac{1 - \pi_1}{\sqrt{2\pi}} e^{-\frac{(X_i - \mu_{-1})^2}{2}}\right), \end{aligned}$$

where X_1^n is notation for (X_1, \dots, X_n) .

This maximization problem has no closed-form solution. For this reason, we have to resort to an iterative algorithm in order to estimate θ^* . The Newton-Raphson method is an available option. However, we describe here another option: the EM algorithm.

To describe this algorithm, let us remark that the joint density of (X, Y) is given, for all $(x, y) \in \mathbb{R} \times \{\pm 1\}$, by:

$$\begin{aligned} g_{\theta^*}(x, y) &= \mathbb{P}(Y = y) p_{\mu_y^*}(x) \\ &= [\pi_1^* p_{\mu_1^*}(x)]^{1_{y=1}} [(1 - \pi_1^*) p_{\mu_{-1}^*}(x)]^{1_{y=-1}}. \end{aligned}$$

Consequently, the induced statistical model to estimate the density of (X, Y) is $\mathcal{P}_m = \{g_\theta : \theta = (\pi_1, \mu_1, \mu_{-1}) \in (0, 1) \times \mathbb{R}^2\}$. In addition, the statistical model for $Y|X$ induced by our first assumption is $\mathcal{P}_c = \{x \in \mathbb{R} \mapsto q_{\theta, x} \in (0, 1) : \theta = (\pi_1, \mu_1, \mu_{-1})\}$.

such that for all $x \in \mathbb{R}$:

$$\begin{aligned} q_{\theta^*, x} &= \mathbb{P}(Y = 1 | X = x) \\ &= \frac{g_{\theta^*}(x, 1)}{m_{\theta^*}(x)} \\ &= \frac{\pi_1^* p_{\mu_1^*}(x)}{m_{\theta^*}(x)}. \end{aligned}$$

Both \mathcal{P}_m and \mathcal{P}_c are sampling model in that they assume being provided with an *iid* sample of observations.

Now, let (Z_1, \dots, Z_n) be a vector of random variables living in $\{\pm 1\}$, such that $(X_1, Z_1), \dots, (X_n, Z_n)$ are *iid*. Z_1, \dots, Z_n are supposed to represent the unknown labels Y_1, \dots, Y_n associated to X_1, \dots, X_n . The joint log-likelihood of θ based on the statistical model \mathcal{P}_m and the sample $\{(X_1, Z_1), \dots, (X_n, Z_n)\}$ is:

$$\begin{aligned} \ell_{(X, Z)_1^n}(\theta) &= \sum_{i=1}^n \log(g_{\theta}(X_i, Z_i)) \\ &= \sum_{i=1}^n [\mathbf{1}_{Z_i=1} \log(\pi_1 p_{y_1}(X_i)) + \mathbf{1}_{Z_i=-1} \log((1 - \pi_1) p_{y_{-1}}(X_i))] \\ &= \sum_{i=1}^n \mathbf{1}_{Z_i=1} \left[\log(\pi_1) - \frac{1}{2} \log(2\pi) - \frac{1}{2} (X_i - \mu_1)^2 \right] \\ &\quad + \sum_{i=1}^n \mathbf{1}_{Z_i=-1} \left[\log(1 - \pi_1) - \frac{1}{2} \log(2\pi) - \frac{1}{2} (X_i - \mu_{-1})^2 \right]. \end{aligned}$$

Defining $\tilde{m} = \sum_{i=1}^n \mathbf{1}_{Z_i=1}$, maximizing this quantity is straightforward and provides the solutions:

$$\begin{cases} \tilde{\pi} &= \frac{\tilde{m}}{n} \\ \tilde{\mu}_1 &= \frac{1}{\tilde{m}} \sum_{\substack{1 \leq i \leq n \\ Z_i=1}} X_i \\ \tilde{\mu}_{-1} &= \frac{1}{n - \tilde{m}} \sum_{\substack{1 \leq i \leq n \\ Z_i=-1}} X_i. \end{cases}$$

Unfortunately,

$$\ell_{(X, Z)_1^n}(\theta) = \sum_{i=1}^n [\mathbf{1}_{Z_i=1} \log(\pi_1 p_{y_1}(X_i)) + \mathbf{1}_{Z_i=-1} \log((1 - \pi_1) p_{y_{-1}}(X_i))]$$

cannot be computed because the random variables $\mathbf{1}_{Z_i=1}$ and $\mathbf{1}_{Z_i=-1}$ are unknown. Thus, $(\tilde{\pi}, \tilde{\mu}_1, \tilde{\mu}_{-1})$ is not an estimator of θ^* .

To circumvent that pitfall, let us assume that a candidate $\hat{\theta} = (\hat{\pi}_1, \hat{\mu}_1, \hat{\mu}_{-1})$ (a function of (X_1, \dots, X_n)) is available to estimate θ^* . The underlying idea of the EM algorithm is to choose the distribution of (Z_1, \dots, Z_n) that suits the best that of (Y_1, \dots, Y_n) given that we have have observed (X_1, \dots, X_n) and that we know $\hat{\theta}$; and then, to replace the unknown quantities $\mathbf{1}_{Z_i=1}$ and $\mathbf{1}_{Z_i=-1}$ by their expected values $\mathbb{E}[\mathbf{1}_{Z_i=1} | X_1^n]$ given that we have observed (X_1, \dots, X_n) . More formally, let us build the Z_i 's such that,

$(X_1, Z_1)|\hat{\theta}, \dots, (X_n, Z_n)|\hat{\theta}$ are *iid* pairs of random variables with, for all $i \in [n]$, $Z_i|X_i^n = Z_i|(\hat{\theta}, X_i)$ ruled by $q_{\hat{\theta}, X_i}$.

As a consequence, for all $i \in [n]$, we can compute the required expected value given the current knowledge:

$$\begin{aligned} p_i &= \mathbb{E}[\mathbf{1}_{Z_i=1}|X_i^n] \\ &= \mathbb{P}(Z_i = 1|X_i^n) \\ &= q_{\hat{\theta}, X_i} \\ &= \frac{\hat{\pi}_1 p_{\hat{\theta}_1}(X_i)}{m_{\hat{\theta}}(X_i)}. \end{aligned}$$

The novel criterion to maximize, with respect to θ , instead of $\ell_{(X, Z)_1^n}$ is

$$F_{X_1^n}(\theta|\hat{\theta}) = \mathbb{E}[\ell_{(X, Z)_1^n}(\theta)|X_1^n] = \sum_{i=1}^n [p_i \log(\pi_1 p_{y_1}(X_i)) + (1 - p_i) \log((1 - \pi_1) p_{y_{-1}}(X_i))],$$

where the vertical bar in $F_{X_1^n}$ is a notation to emphasize that $F_{X_1^n}$ is computed given $\hat{\theta}$.

The EM algorithm can be summarized by the iterative process, given an initialization $\hat{\theta}_0$:

1. Expectation: compute, for all $i \in [n]$,

$$p_i^{(t)} = \mathbb{E}[\mathbf{1}_{Z_i^{(t)}=1}|X_1^n],$$

with $Z_i^{(t)}|X_1^n$ distributed according to $q_{\hat{\theta}_t, X_i}$.

2. Maximization: set $\hat{\theta}_{t+1} \in \arg \max_{\theta \in \Theta} F_{X_1^n}(\theta|\hat{\theta}_t)$, where

$$F_{X_1^n}(\theta|\hat{\theta}_t) = \sum_{i=1}^n [p_i^{(t)} \log(\pi_1 p_{y_1}(X_i)) + (1 - p_i^{(t)}) \log((1 - \pi_1) p_{y_{-1}}(X_i))].$$

Theorem 50 justifies that the EM algorithm builds a reasonable estimator of $\hat{\theta}^*$. Before, let us see in details what happens for a mixture of multivariate normal distributions.

2.1.3 EM for Gaussian mixtures (soft k-means)

In this section, we apply the EM algorithm to a mixture of k Gaussian distributions in \mathbb{R}^d . In other words, we are tackling a clustering problem with $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^d \times [k]$. In addition, this means that:

1. the set of parameters is:

$$\Theta = \{\theta = (\pi_1, \dots, \pi_k, \lambda_1, \dots, \lambda_k) : \pi \in [0, 1]^k, \mathbf{1}^\top \pi = 1, \lambda \in \Lambda\},$$

where Λ is the set of Gaussian parameters defined by:

$$\Lambda = \{\lambda = (\lambda_1, \dots, \lambda_k) : \forall j \in [k], \lambda_j = (\mu_j, \Sigma_j) \in \mathbb{R}^d \times \mathbb{R}^{d \times d}, \Sigma_j \text{ PD matrix}\};$$

2. $X|Y \sim \mathcal{N}(\mu_Y^*, \Sigma_Y^*)$ and we denote $p_{\lambda_Y^*} = p_{(\mu_Y^*, \Sigma_Y^*)} : x \in \mathbb{R}^d \mapsto |2\pi\Sigma_Y^*|^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\mu_Y^*)^\top \Sigma_Y^{*-1}(x-\mu_Y^*)}$ its pdf;

3. Y has a density defined by $\pi^* \in [0, 1]^k$: for all $j \in [k]$, $\mathbb{P}(Y = j) = \pi_j^*$;
4. $(X, Y) \sim G_{\theta^*}$, with density $g_{\theta^*}: (x, y) \in \mathbb{R}^d \times [k] \mapsto \pi_y^* p_{\lambda_y^*}(x)$;
5. $X \sim M_{\theta^*}$ with density $m_{\theta^*}: x \in \mathbb{R}^d \mapsto \sum_{j=1}^k \pi_j^* p_{\lambda_j^*}(x)$;
6. $Y|X \sim Q_{\theta^*, X}$, defined by the probability vector $q_{\theta^*, X} \in [0, 1]^k$, such that for all $j \in [k]$:

$$(q_{\theta^*, X})_j = \mathbb{P}(Y = j|X) = \frac{g_{\theta^*}(X, j)}{m_{\theta^*}(X)} = \frac{\pi_j^* p_{\lambda_j^*}(X)}{\sum_{\ell=1}^k \pi_\ell^* p_{\lambda_\ell^*}(X)}.$$

Let $(X_1, Y_1), \dots, (X_n, Y_n)$ be n iid copies of (X, Y) . In clustering, we only observe (X_1, \dots, X_n) and we would like to infer (Y_1, \dots, Y_n) . We now detail the two steps of the EM algorithm applied to this problem in order to estimate $Q_{\theta^*, X}$.

Expectation step

Let $\hat{\theta}_t = (\hat{\pi}_1^{(t)}, \dots, \hat{\pi}_k^{(t)}, \hat{\lambda}_1^{(t)}, \dots, \hat{\lambda}_k^{(t)}) \in \Theta$ be a current estimate of θ^* , and $(Z_1^{(t)}, \dots, Z_n^{(t)})$ be a sample such that

1. $\{(X_1, Z_1^{(t)})|\hat{\theta}_t, \dots, (X_n, Z_n^{(t)})|\hat{\theta}_t\}$ is iid;
2. for each $i \in [n]$, $Z_i^{(t)}|X_i^n \sim Q_{\hat{\theta}_t, X_i}$, that is, for all $j \in [k]$:

$$\mathbb{P}(Z_i^{(t)} = j|X_i^n) = \frac{\hat{\pi}_j^{(t)} p_{\lambda_j^{(t)}}(X_i)}{\sum_{\ell=1}^k \hat{\pi}_\ell^{(t)} p_{\lambda_\ell^{(t)}}(X_i)}. \quad (2.1)$$

The first step of EM is to compute the conditional expectation of the joint log-likelihood, which is, for any $\theta \in \Theta$:

$$\begin{aligned} F_{X_1^n}(\theta|\hat{\theta}_t) &= \mathbb{E} \left[\sum_{i=1}^n \log \left(g_\theta(X_i, Z_i^{(t)}) \right) | X_1^n \right] \\ &= \sum_{i=1}^n \sum_{j=1}^k p_{ij}^{(t)} [\log(\pi_j) + \log(p_{\lambda_j}(X_i))], \end{aligned}$$

where, for all $i \in [n]$ and $j \in [k]$,

$$p_{ij}^{(t)} = \mathbb{P}(Z_i^{(t)} = j|X_i^n) = \frac{\hat{\pi}_j^{(t)} p_{\lambda_j^{(t)}}(X_i)}{\sum_{\ell=1}^k \hat{\pi}_\ell^{(t)} p_{\lambda_\ell^{(t)}}(X_i)}. \quad (2.2)$$

Maximization step

Given the computation of $F_{X_1^n}(\theta|\hat{\theta}_t)$, the goal of this second step is to maximize $F_{X_1^n}(\theta|\hat{\theta}_t)$ with respect to θ , that is to solve

$$\begin{aligned} & \underset{\theta \in \Theta}{\text{maximize}} \quad \sum_{i=1}^n \sum_{j=1}^k p_{ij}^{(t)} \left[\log(\pi_j) - \frac{1}{2}(X_i - \mu_j)^\top \Sigma_j^{-1} (X_i - \mu_j) - \frac{1}{2} \log(|\Sigma_j|) \right] \\ & \text{s. t.} \quad \begin{cases} \theta = (\pi_1, \dots, \pi_k, (\mu_1, \Sigma_1), \dots, (\mu_k, \Sigma_k)) \\ \sum_{j=1}^k \pi_j = 1 \\ \forall j \in [k]: \pi_j \geq 0 \\ \mu_j \in \mathbb{R}^d \\ \Sigma_j \in \mathbb{R}^{d \times d}, PD. \end{cases} \end{aligned} \quad (\text{P19})$$

Property 46. *Solution to Problem (P19) is*

$$\hat{\theta}_{t+1} = (\hat{\pi}_1^{(t+1)}, \dots, \hat{\pi}_k^{(t+1)}, (\hat{\mu}_1^{(t+1)}, \hat{\Sigma}_1^{(t+1)}), \dots, (\hat{\mu}_k^{(t+1)}, \hat{\Sigma}_k^{(t+1)})),$$

where for all $j \in [k]$:

$$\begin{cases} \hat{\pi}_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^n p_{ij}^{(t)} \\ \hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} X_i}{\sum_{i=1}^n p_{ij}^{(t)}} \\ \hat{\Sigma}_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} [(X_i - \hat{\mu}_j^{(t+1)})(X_i - \hat{\mu}_j^{(t+1)})^\top]}{\sum_{i=1}^n p_{ij}^{(t)}}. \end{cases}$$

The proof is a good exercise.

Algorithm 5 EM for Gaussian mixtures (soft k-means).

Input: $\{X_i\}_{1 \leq i \leq n}$ (training sample).

$\pi_j \leftarrow \frac{1}{k}$, for all $j \in [k]$ (*initialization*)

$\mu_j \leftarrow$ random point, for all $j \in [k]$

$\Sigma_j \leftarrow$ overall sample covariance, for all $j \in [k]$

while not converged **do**

$p_{ij} \leftarrow \frac{\pi_j \phi(\mu_j, \Sigma_j)(X_i)}{\sum_{\ell=1}^k \pi_\ell \phi(\mu_\ell, \Sigma_\ell)(X_i)} \approx \mathbb{P}(Y_i = j | X_i)$ (*expectation*)

$\pi_j \leftarrow \frac{1}{n} \sum_{i=1}^n p_{ij}$ (*maximization*)

$\mu_j \leftarrow \frac{\sum_{i=1}^n p_{ij} X_i}{\sum_{i=1}^n p_{ij}}$

$\Sigma_j \leftarrow \frac{\sum_{i=1}^n p_{ij} [(X_i - \mu_j)(X_i - \mu_j)^\top]}{\sum_{i=1}^n p_{ij}}$

end while

The steps described above are summed up in Algorithm 5. This algorithm (EM for Gaussian mixtures) is often called soft k-means because of its similarity to the k-means algorithm (Algorithm 8, see Remark 2.2.4).

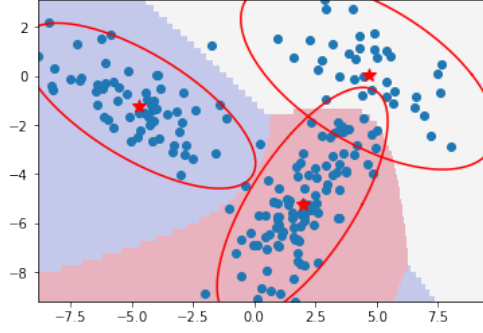


Figure 2.1: Example of soft k-means and clustering frontier.

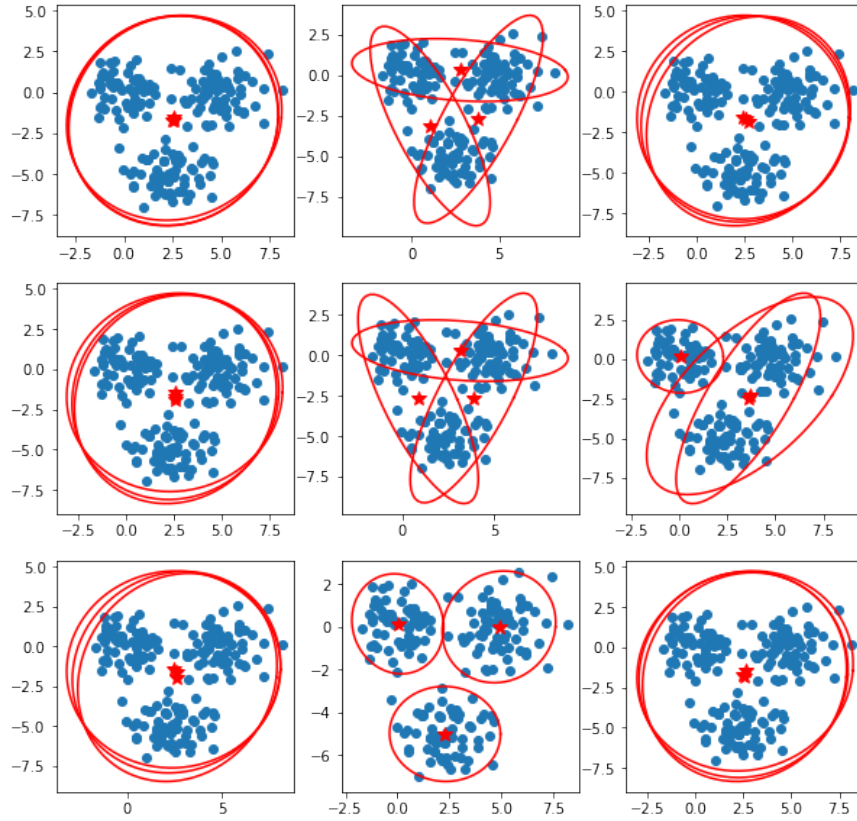


Figure 2.2: Soft k-means can produce very different results (including unexpected ones) according to the random initialization of means.

2.1.4 The general EM algorithm

We consider the statistical model $\{G_\theta : \theta \in \Theta\}$, where G_θ is a distribution over $\mathcal{X} \times \mathcal{Y}$ (where \mathcal{X} and \mathcal{Y} are any probabilistic spaces), defined by its marginal in X , denoted M_θ , and by the conditional distribution $Q_{\theta,X}$ of $Y|X$. We assume the model to be dominated by a product measure $\mu \times \nu$. The

densities of interest are denoted:

$$\begin{cases} g_\theta &= \frac{dG_\theta}{d(\mu \times \nu)} \\ m_\theta &= \frac{dM_\theta}{d\mu} \\ q_{\theta,x} &= \frac{dQ_{\theta,x}}{d\nu}, \quad x \in \mathcal{X}. \end{cases}$$

Let $\theta^* \in \Theta$ and $(X, Y) \sim G_{\theta^*}$. We aim at computing an MLE $\hat{\theta}_{MLE}$ of θ^* , assuming that we only observe X (Y is a latent variable). Thus, $\hat{\theta}_{MLE}$ may naturally be a marginal MLE:

$$\hat{\theta}_{MLE} \in \arg \max_{\theta \in \Theta} \log(m_\theta(X)).$$

In many situations, determining $\hat{\theta}_{MLE}$ is intractable, because of the form of the marginal density of X (for instance, as seen in the previous section, the marginal density of a mixture model is a sum of several contributions, making the numerical estimation difficult). However, if the joint density is more enjoyable, it is quite tempting to consider maximizing the joint log-likelihood

$$\log(g_\theta(X, Z)),$$

where $Z \in \mathcal{Y}$ is a latent variable, supposed to represent the unobserved Y and chosen adequately. A good choice is to choose $Z|X \sim Q_{\hat{\theta},X}$, where $\hat{\theta}$ is a candidate to estimate θ^* .

However, since we are not able to observe Z in practice, we resort to maximizing a proxy: the expected value of the joint log-likelihood, given by:

$$F_X(\theta|\hat{\theta}) = \mathbb{E}[\log(g_\theta(X, Z)) | X].$$

The EM aims at computing $\hat{\theta}_{MLE}$ by proceeding alternatively in two steps, as described in Algorithm 6.

Algorithm 6 EM algorithm.

Input: $T \in \mathbb{N}$ (number of iterations), X (observed sample).

$\hat{\theta}_0 \leftarrow$ random initialization

for $t = 0$ **to** $T - 1$ **do**

set $Z^{(t)}|X \sim Q_{\hat{\theta}_t,X}$

E step: compute $F(\theta|\hat{\theta}_t) = \mathbb{E}[\log(g_\theta(X, Z^{(t)})) | X]$

M step: set $\hat{\theta}_{t+1} \in \arg \max_{\theta \in \Theta} F(\theta|\hat{\theta}_t)$

end for

Output: $\hat{\theta}_T$.

Remark 2.1.2 (The iid sample case). In many situations, $(X, Y) = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is an iid sample. Let us denote G'_{θ^*} the distribution of (X_1, Y_1) and Q'_{θ^*,X_1} the distribution of $Y_1|X_1$, then $G_{\theta^*} = (G'_{\theta^*})^{\otimes n}$ and $Q_{\theta^*,X} = Q'_{\theta^*,X_1} \otimes \dots \otimes Q'_{\theta^*,X_n}$. In other words, $Y_1|X_1, \dots, Y_n|X_n$ are independent (but not identically distributed).

Thus, $(Z^{(t)}|X) \sim (Q'_{\hat{\theta}_t,X_1} \otimes \dots \otimes Q'_{\hat{\theta}_t,X_n})$, which means that for all $i \in [n]$, $Z_i^{(t)}|X_1^n = Z_i^{(t)}|(\hat{\theta}_t, X_i) \sim Q_{\hat{\theta}_t,X_i}$ and $Z_1^{(t)}|(\hat{\theta}_t, X_1), \dots, Z_n^{(t)}|(\hat{\theta}_t, X_n)$ are independent. It comes that $Z_1^{(t)}|\hat{\theta}_t, \dots, Z_n^{(t)}|\hat{\theta}_t$ are iid.

In addition, denoting g'_θ the density of G'_θ , the criterion to maximize becomes:

$$F_{X_1^n}(\theta|\hat{\theta}_t) = \mathbb{E} \left[\sum_{i=1}^n \log \left(g'_\theta(X_i, Z_i^{(t)}) \right) | X_1^n \right].$$

In the forthcoming paragraphs, the reasons of the success of EM are explained. It will be shown that the expected joint log-likelihood is a lower bound of the marginal log-likelihood and that EM, for lack of converging to a maximizer of $\log(m_\theta(X))$, produces a monotonically increasing sequence.

Lemma 47. Let $Q_x = q_x$ be any distribution on \mathcal{Y} parameterized by $x \in \mathcal{X}$, with density q_x with respect to ν , and $Z \in \mathcal{Y}$ a random variable such that $Z|X \sim Q_x$. Then, for all $\theta \in \Theta$, one has:

$$\log(m_\theta(X)) = \mathbb{E}[\log(g_\theta(X, Z)) | X] + D_{KL}(Q_x || Q_{\theta, X}) + H(Q_x),$$

where D_{KL} and H are respectively the Kullback-Leibler divergence and the entropy defined by:

$$D_{KL}(Q_x || Q_{\theta, X}) = \mathbb{E} \left[\log \left(\frac{q_x(Z)}{q_{\theta, X}(Z)} \right) | X \right] \quad \text{and} \quad H(Q_x) = -\mathbb{E}[\log(q_x(Z)) | X].$$

In particular, for any $\theta \in \Theta$ and $\theta' \in \Theta$, setting $Z|X \sim Q_{\theta', X}$, one has

$$\log(m_\theta(X)) = F(\theta|\theta') + D_{KL}(\theta' || \theta) + H(\theta'),$$

where $F(\theta|\theta') = \mathbb{E}[\log(g_\theta(X, Z)) | X]$ and, with a slight abuse of notation, $D_{KL}(\theta' || \theta) = D_{KL}(Q_{\theta', X} || Q_{\theta, X})$ and $H(\theta') = H(Q_{\theta', X})$.

Proof. Let $Z \in \mathcal{Y}$ be a random variable such that $Z|X \sim Q_x$ and remember that, by the Bayes rule: $g_\theta(X, Z) = q_{\theta, X}(Z)m_\theta(X)$. Thus,

$$\begin{aligned} \log(m_\theta(X)) &= \mathbb{E}[\log(m_\theta(X)) | X] \\ &= \mathbb{E}[\log(g_\theta(X, Z)) | X] - \mathbb{E}[\log(q_{\theta, X}(Z)) | X] \\ &= \mathbb{E}[\log(g_\theta(X, Z)) | X] - \mathbb{E}[\log(q_{\theta, X}(Z)) | X] + \mathbb{E}[\log(q_x(Z)) | X] - \mathbb{E}[\log(q_x(Z)) | X] \\ &= \mathbb{E}[\log(g_\theta(X, Z)) | X] + \mathbb{E} \left[\log \left(\frac{q_x(Z)}{q_{\theta, X}(Z)} \right) | X \right] - \mathbb{E}[\log(q_x(Z)) | X] \\ &= \mathbb{E}[\log(g_\theta(X, Z)) | X] + D_{KL}(Q_x || Q_{\theta, X}) + H(Q_x). \end{aligned}$$

The final statement is obtained by setting $Q_x = Q_{\theta', X}$. □

Property 48 (Kullback-Leibler divergence). For any distributions $P = p \cdot \mu$ and $Q = q \cdot \mu$ absolutely continuous with respect to the same measure μ ,

$$D_{KL}(P || Q) \geq 0,$$

and

$$D_{KL}(P||Q) = 0 \iff p = q \quad \mu - \text{almost surely.}$$

Proof. Let $Z \sim P$. By convexity of $-\log$ and Jensen's inequality, one has:

$$\begin{aligned} D_{KL}(P||Q) &= \mathbb{E} \left[\log \left(\frac{p(Z)}{q(Z)} \right) \right] \\ &= \mathbb{E} \left[-\log \left(\frac{q(Z)}{p(Z)} \right) \right] \\ &\geq -\log \left(\mathbb{E} \left[\frac{q(Z)}{p(Z)} \right] \right) \\ &= -\log \left(\int \frac{q(z)}{p(z)} p(z) d\mu \right) \\ &= 0. \end{aligned}$$

The second part follows from the conditions for equality in Jensen's inequality: by strict convexity of $-\log$, equality holds if and only if $\frac{q}{p}$ is a constant μ -almost surely. By summation to 1, this constant is necessarily 1, meaning that $p = q$ μ -almost surely. \square

Proposition 49. For any $\theta \in \Theta$ and $\theta' \in \Theta$ (function of X), one has

$$\log(m_\theta(X)) \geq F(\theta|\theta') + H(\theta'),$$

and

$$\log(m_\theta(X)) = F(\theta|\theta) + H(\theta).$$

Proof. This is a consequence of Lemma 47: for the inequalities, it is enough to observe that $D_{KL}(\theta'|\theta) \geq 0$ and $H(\theta') \geq 0$, whereas for the equality, we set $\theta' = \theta$. \square

Proposition 49 tells us that the marginal log-likelihood is always bounded from below by $F(\theta|\hat{\theta}) + H(\hat{\theta})$. Therefore, EM can be viewed as the two maximization steps described in Algorithm 7 and illustrated in Figure 2.3.

Theorem 50. Let $(\hat{\theta}_t)_{t \geq 0}$ be the sequence defined by $\hat{\theta}_0 \in \Theta$ and for all integer t ,

$$\hat{\theta}_{t+1} \in \arg \max_{\theta \in \Theta} F(\theta|\hat{\theta}_t).$$

Then the sequence $(\log(m_{\hat{\theta}_t}(X)))_{t \geq 0}$ is nondecreasing.

Algorithm 7 EM algorithm (maximization–maximization).

Input: $T \in \mathbb{N}$ (number of iterations), X (observed sample).

$\hat{\theta}_0 \leftarrow$ random initialization

for $t = 0$ **to** $T - 1$ **do**

E step: set $\hat{\theta}'_t \in \arg \max_{\theta \in \Theta} F(\hat{\theta}_t | \theta) + H(\theta)$, that is $\hat{\theta}'_t = \hat{\theta}_t$ (best lower bound of $\log(m_\theta(X))$ knowing $\hat{\theta}_t$)

M step: set $\hat{\theta}_{t+1} \in \arg \max_{\theta \in \Theta} F(\theta | \hat{\theta}_t)$ (maximize the lower bound given $\hat{\theta}_t$)

end for

Output: $\hat{\theta}_T$.

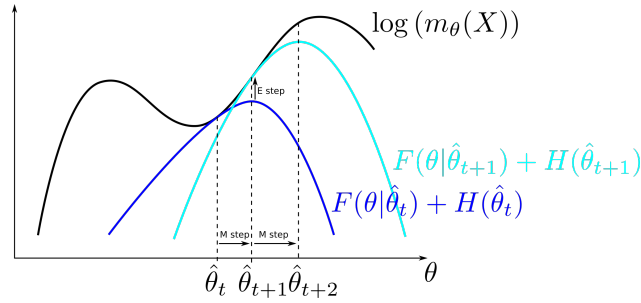


Figure 2.3: Illustration of the E step (finding the best lower bound) and the M step (maximizing the lower bound) of the EM algorithm.

Proof. For all integer t ,

$$\begin{aligned}
 \log(m_{\hat{\theta}_{t+1}}(X)) &\geq F(\hat{\theta}_{t+1} | \hat{\theta}_t) + H(\hat{\theta}_t) && \text{(Proposition 49)} \\
 &\geq F(\hat{\theta}_t | \hat{\theta}_t) + H(\hat{\theta}_t) && \text{(Definition of } \hat{\theta}_{t+1}) \\
 &= \log(m_{\hat{\theta}_t}(X)) .
 \end{aligned}$$

□

2.1.5 Model selection

Model selection for clustering generally lies in choosing the number of clusters k . Some empirical methods will be presented in Section 2.5, nevertheless, we quickly introduce here two criteria specific to likelihood maximization.

When computing an MLE, it is possible to increase the likelihood by adding parameters. For instance, with Gaussian mixtures, considering $k = n$, $\mu_j = X_j$ for all $j \in [k]$ and $\Sigma_j = \sigma^2 I$, with $\sigma^2 \rightarrow 0$ leads to a likelihood increasing to 1. This situation is typical from overfitting the training sample.

The Bayesian information criterion (BIC) and Akaike information criterion (AIC) criteria help in choosing the number of clusters by adding a penalty term growing with the number of free parameters in the model. Given an *iid* sample $\{X_1, \dots, X_n\}$ and an MLE $\hat{\theta}$, BIC and AIC are defined by

$$BIC = -2 \log(m_{\hat{\theta}}(X_1, \dots, X_n)) + m \log(n),$$

and

$$AIC = -2 \log (m_{\hat{\theta}}(X_1, \dots, X_n)) + 2m,$$

where m is the number of free parameters (for Gaussian mixtures, $m = (k-1) + kd + k \frac{d(d+1)}{2}$). The number of clusters can be chosen as the one minimizing either BIC or AIC (note that BIC is more conservative than AIC in that its penalty term is larger than the one in AIC as soon as $n \geq 8$).

Criteria BIC and AIC come respectively from Bayesian and information theory. It can be shown that, under different assumptions and for n large, the log-likelihood of a model can be approximated by either $-\frac{BIC}{2}$ or $-\frac{AIC}{2}$. As a consequence, minimizing one of these two criteria, tends to maximize the likelihood of the model.

2.2 Cost minimization methods

Similarly to what has been seen for supervised learning, clustering can be tackled either under Gaussian assumptions or by minimization of a cost. Clustering is then a partitioning of minimal cost.

In order to define this cost, we need a dissymmetry measure $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$, that is a non-negative symmetric function d such that $d(x, x) = 0$. Note that a distance is a dissymmetry measure with extra properties (separation and triangle inequality).

Based on that, the cost of interest can be defined by two different manners: based on the centers of the clusters or based on the points inside the clusters.

2.2.1 Center-based objectives

Let $\varphi: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ be a monotonically increasing function. Center-based cost functions are based on cluster centers $(\mu_1, \dots, \mu_k) \in \mathcal{X}^k$, also called centroids or representer. For a given subset $C \subset X$, its center is defined by:

$$\mu(C) = \arg \min_{\mu \in \mathcal{X}} \mathbb{E}(\varphi(d(X, \mu)) \mathbf{1}_{X \in C}),$$

and, given an *iid* sample (X_1, \dots, X_n) along with the empirical cluster $\hat{C} = \{X_i : i \in [n]\} \cap C$, the empirical twin of $\mu(C)$ is

$$\hat{\mu}(C) = \arg \min_{\mu \in \mathcal{X}} \frac{1}{n} \sum_{i=1}^n \varphi(d(X_i, \mu)) \mathbf{1}_{X_i \in C} = \arg \min_{\mu \in \mathcal{X}} \frac{1}{n} \sum_{X \in \hat{C}} \varphi(d(X, \mu)).$$

Let $P(\mathcal{X})$ be the complete set partition of size k of \mathcal{X} . Then, we are interested in minimizing the distortion over all possible partitioning:

$$\underset{(C_1, \dots, C_k) \in P(\mathcal{X})}{\text{minimize}} \quad D(C_1, \dots, C_k), \quad (\text{P20})$$

where

$$D(C_1, \dots, C_k) = \mathbb{E} \left(\sum_{j=1}^k \varphi(d(X, \mu(C_j))) \mathbf{1}_{X \in C_j} \right).$$

Remark 2.2.1. Defining the quantizer $q: x \in \mathcal{X} \mapsto \sum_{j=1}^k \mu(C_j) \mathbf{1}_{x \in C_j}$, we obtain for D the usual distortion:

$$D(C_1, \dots, C_k) = \mathbb{E}(\varphi(d(X, q(X)))).$$

k-means objective

Let us consider $\varphi: x \in \mathbb{R}_+ \mapsto x^2$. k-means is mostly known with the Euclidean distance d . Then one has for all $j \in [k]$:

$$\mu(C_j) = \arg \min_{\mu \in \mathcal{X}} \mathbb{E} \left(\|X - \mu\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) = \frac{1}{\mathbb{P}(X \in C_j)} \mathbb{E} (X \mathbf{1}_{X \in C_j}) ,$$

and

$$\hat{\mu}(C_j) = \frac{1}{|\hat{C}_j|} \sum_{X \in \hat{C}_j} X.$$

Also

$$D(C_1, \dots, C_k) = \mathbb{E} \left(\sum_{j=1}^k \|X - \mu(C_j)\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right)$$

and

$$\hat{D}(C_1, \dots, C_k) = \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j} \|X - \hat{\mu}(C_j)\|_{\ell_2}^2 .$$

Proposition 51. For any partition (C_1, \dots, C_k) of \mathcal{X} , we have:

$$\mathbb{E} \left(\|X - \mathbb{E} X\|_{\ell_2}^2 \right) = \mathbb{E} \left(\sum_{j=1}^k \|X - \mu(C_j)\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mathbb{E} X - \mu(C_j)\|_{\ell_2}^2$$

Proof. Let us expand the first, then the second term of the right hand side, keeping in mind that for all $j \in [k]$:

$$\mu(C_j) \mathbb{P}(X \in C_j) = \mathbb{E} (X \mathbf{1}_{X \in C_j}) .$$

First,

$$\begin{aligned}
\mathbb{E} \left(\sum_{j=1}^k \|X - \mu(C_j)\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) &= \mathbb{E} \left(\sum_{j=1}^k \|X\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) + \mathbb{E} \left(\sum_{j=1}^k \|\mu(C_j)\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) \\
&\quad - 2 \mathbb{E} \left(\sum_{j=1}^k \langle X, \mu(C_j) \rangle_{\ell_2} \mathbf{1}_{X \in C_j} \right) \\
&= \mathbb{E}(\|X\|_{\ell_2}^2) + \sum_{j=1}^k \|\mu(C_j)\|_{\ell_2}^2 \mathbb{P}(X \in C_j) - 2 \sum_{j=1}^k \|\mu(C_j)\|_{\ell_2}^2 \mathbb{P}(X \in C_j) \\
&= \mathbb{E}(\|X\|_{\ell_2}^2) - \sum_{j=1}^k \|\mu(C_j)\|_{\ell_2}^2 \mathbb{P}(X \in C_j).
\end{aligned}$$

Then,

$$\begin{aligned}
\sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mathbb{E} X - \mu(C_j)\|_{\ell_2}^2 &= \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mathbb{E} X\|_{\ell_2}^2 + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mu(C_j)\|_{\ell_2}^2 \\
&\quad - 2 \sum_{j=1}^k \mathbb{P}(X \in C_j) \langle \mathbb{E} X, \mu(C_j) \rangle_{\ell_2} \\
&= \|\mathbb{E} X\|_{\ell_2}^2 + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mu(C_j)\|_{\ell_2}^2 - 2 \left\langle \mathbb{E} X, \sum_{j=1}^k \mathbb{E}(X \mathbf{1}_{X \in C_j}) \right\rangle_{\ell_2} \\
&= \|\mathbb{E} X\|_{\ell_2}^2 + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mu(C_j)\|_{\ell_2}^2 - 2 \|\mathbb{E} X\|_{\ell_2}^2 \\
&= -\|\mathbb{E} X\|_{\ell_2}^2 + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mu(C_j)\|_{\ell_2}^2.
\end{aligned}$$

By summation,

$$\begin{aligned}
\mathbb{E} \left(\sum_{j=1}^k \|X - \mu(C_j)\|_{\ell_2}^2 \mathbf{1}_{X \in C_j} \right) + \sum_{j=1}^k \mathbb{P}(X \in C_j) \|\mathbb{E} X - \mu(C_j)\|_{\ell_2}^2 &= \mathbb{E}(\|X\|_{\ell_2}^2) - \|\mathbb{E} X\|_{\ell_2}^2 \\
&= \mathbb{E} \left(\|X - \mathbb{E} X\|_{\ell_2}^2 \right).
\end{aligned}$$

□

This proposition is sometimes referred to as “Huygens property”. The three terms are respectively:

1. the total inertia;

2. the intraclass inertia;
3. the interclass inertia.

The previous proposition highlights that minimizing the intraclass inertia ($D(C_1, \dots, C_k)$) also maximizes the interclass inertia.

k-medoids objective

This is similar to k-means but it requires the cluster centers to be members of the input set (X_1, \dots, X_n) . Thus, for all $j \in [k]$:

$$\hat{\mu}(C_j) = X_t \quad \text{with} \quad t \in \arg \min_{i \in [n]} \sum_{X \in C_j} d(X, X_i)^2.$$

k-median objective

This is similar to k-medoids except that φ is the identity function:

$$\hat{\mu}(C_j) = X_t \quad \text{with} \quad t \in \arg \min_{i \in [n]} \sum_{X \in C_j} d(X, X_i),$$

for all $j \in [k]$ and

$$\hat{D}(C_1, \dots, C_k) = \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j} d(X, \hat{\mu}(C_j)).$$

2.2.2 k-means algorithm

As a recall, the (empirical) optimization problem of k-means is:

$$\underset{(C_1, \dots, C_k) \in P(\mathcal{X})}{\text{minimize}} \quad \sum_{j=1}^k \sum_{X \in \hat{C}_j} \|X - \hat{\mu}(C_j)\|_{\ell_2}^2,$$

which can be reformulated as:

$$\begin{aligned} & \underset{\substack{(C_1, \dots, C_k) \in P(\mathcal{X}) \\ (\hat{\mu}_1, \dots, \hat{\mu}_k) \in \mathcal{X}}}{\text{minimize}} \quad \sum_{j=1}^k \sum_{X \in \hat{C}_j} \|X - \hat{\mu}_j\|_{\ell_2}^2 \\ & \text{s. t.} \quad \hat{\mu}_j = \hat{\mu}(C_j) = \frac{1}{|\hat{C}_j|} \sum_{X \in \hat{C}_j} X, \quad \forall j \in [k]. \end{aligned}$$

Minimizing the k-means objective function turns out to be often computationally infeasible (it is NP-hard and even NP-hard to approximate to within some constant). For this reason, we resort to alternating the

computation of $(\hat{\mu}(C_j))_{1 \leq j \leq k}$ (for a fixed partition) and of a Voronoi partitioning (C_1, \dots, C_k) corresponding to the precomputed cluster centers $(\hat{\mu}_1, \dots, \hat{\mu}_k)$ (see Algorithm 8):

$$C_1 = \{x \in \mathcal{X} : d(x, \hat{\mu}_1) \leq d(x, \hat{\mu}_\ell), \forall \ell \in [k]\},$$

and for all $j \in [k], j > 1$:

$$C_j = \{x \in \mathcal{X} : d(x, \hat{\mu}_j) \leq d(x, \hat{\mu}_\ell), \forall \ell \in [k]\} \setminus \bigcup_{\ell=1}^{j-1} C_\ell.$$

Remark 2.2.2. Since k -means computes a Voronoi partitioning, it implicitly assumes convex clusters, that are uniquely defined by their centroids (no notion of variance). In this sense, k -means is weaker than Gaussian mixtures (see Figure 2.4 versus Figure 2.1, and Figure 2.5).

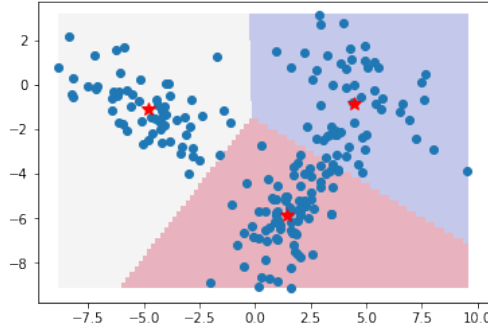


Figure 2.4: Example of k -means clustering.

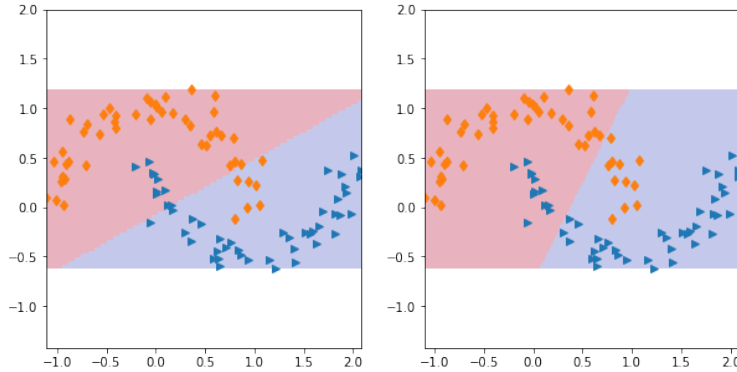


Figure 2.5: Comparison of k -means and soft k -means on non-Gaussian clusters.

Remark 2.2.3. The k -means algorithm is also known as Lloyd's algorithm, after Stuart Lloyd, who proposed the method in 1957 in the context of vector quantization for image compression.

Algorithm 8 k-means.

Input: $T \in \mathbb{N}$ (number of iterations), $\{X_i\}_{1 \leq i \leq n}$ (training sample).

$\hat{\mu}_i \leftarrow$ random point from \mathcal{X} for all $i \in [n]$ (*initialization*)

for $t = 1$ **to** T **do**

 compute a Voronoi partitioning (C_1, \dots, C_k) corresponding to cluster centers $(\hat{\mu}_1, \dots, \hat{\mu}_k)$

$\hat{C}_j \leftarrow \{X_1, \dots, X_n\} \cap C_j$ for all $j \in [k]$

$\hat{\mu}_j \leftarrow \frac{1}{|\hat{C}_j|} \sum_{X \in \hat{C}_j} X$

end for

Output: (C_1, \dots, C_k) .

Remark 2.2.4 (k-means and soft k-means). *Algorithms 5 and 8 share many similarities. Indeed, when computing a Voronoi partitioning, the k-means algorithm assigns each point X_i to a cluster \hat{C}_j and then update the cluster centroid $\hat{\mu}_j$ by averaging the members of the cluster \hat{C}_j .*

However, the soft k-means algorithm first estimates the probability that each example X_i belongs to each cluster \hat{C}_j (based on a Mahalanobis distance between X_i and $\hat{\mu}_j$) and then updates the centroids with a weighted average over the entire sample $\{X_1, \dots, X_n\}$.

If we fix the covariance matrices Σ_j of Algorithm 5 to $\sigma^2 I$, then the probability of assigning X_i to \hat{C}_j becomes a monotone function of the Euclidean distance between the data point X_i and the centroid $\hat{\mu}_j$. Moreover, as $\sigma^2 \rightarrow 0$, these probabilities become 0 and 1, and the two algorithms coincide.

Remark 2.2.5 (k-means and weighted k-nearest neighbors). *The update of the centers in Algorithm 5 can be rewritten, for all $j \in [k]$,*

$$\mu_j = \frac{1}{n} \sum_{i=1}^n w_i X_i,$$

where $w_i \propto p_{ij} \propto e^{-X_i - \mu_j^2_{\Sigma_j}}$, with \cdot_{Σ_j} being a Mahalanobis distance. In other words, the contribution of each point is considered weighted by an exponential function of the distance, which is the same spirit as weighted k-nearest neighbors.

Proposition 52. *The empirical distortion $\hat{D}: (C_1, \dots, C_k) \mapsto \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j} \|X - \hat{\mu}(C_j)\|_{\ell_2}^2$ is monotonically decreasing along the iterations of Algorithm 8.*

Proof. Let $t \in [T - 1]$ and let us denote with the superscript t the values at iteration t . As a recall, we have

$$\hat{D}(C_1, \dots, C_k) = \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k \|X_i - \hat{\mu}(C_j)\|_{\ell_2}^2 \mathbf{1}_{X_i \in \hat{C}_j}.$$

We will show that

$$\begin{aligned}
\hat{D}(C_1^t, \dots, C_k^t) &= \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^t} \|X - \hat{\mu}_j^t\|_{\ell_2}^2 \\
&\geq \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \hat{\mu}_j^t\|_{\ell_2}^2 && \text{Step 1: Voronoi partitioning} \\
&\geq \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \hat{\mu}_j^{t+1}\|_{\ell_2}^2 && \text{Step 2: computation of the means} \\
&= \hat{D}(C_1^{t+1}, \dots, C_k^{t+1}).
\end{aligned}$$

First step

At the first step of iteration $t + 1$ of Algorithm 8, a Voronoi partitioning $(C_1^{t+1}, \dots, C_k^{t+1})$ based on $(\hat{\mu}_1^t, \dots, \hat{\mu}_k^t)$ is computed. By definition of a Voronoi partitioning, we have for all $j \in [k]$ and $\forall x \in C_j^{t+1}$:

$$\|x - \mu_j^t\|_{\ell_2}^2 \leq \|x - \mu_\ell^t\|_{\ell_2}^2, \quad \forall \ell \in [k].$$

So $\forall x \in \mathbb{R}^d$:

$$\begin{aligned}
\sum_{j=1}^k \|x - \mu_j^t\|_{\ell_2}^2 \mathbf{1}_{x \in C_j^{t+1}} &= \|x - \mu_j^t\|_{\ell_2}^2 && \text{for } j \text{ such that } x \in C_j^{t+1} \\
&\leq \|x - \mu_\ell^t\|_{\ell_2}^2 && \forall \ell \in [k] \\
&= \sum_{\ell=1}^k \|x - \mu_\ell^t\|_{\ell_2}^2 \mathbf{1}_{x \in C_\ell^t}.
\end{aligned}$$

Summing over all X_i ($i \in [n]$), we obtain:

$$\sum_{i=1}^n \sum_{j=1}^k \|X_i - \mu_j^t\|_{\ell_2}^2 \mathbf{1}_{X_i \in \hat{C}_j^{t+1}} \leq \sum_{i=1}^n \sum_{j=1}^k \|X_i - \mu_j^t\|_{\ell_2}^2 \mathbf{1}_{X_i \in \hat{C}_j^t},$$

that is, when reorganizing the sums:

$$\sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_j^t\|_{\ell_2}^2 \leq \sum_{j=1}^k \sum_{X \in \hat{C}_j^t} \|X - \mu_j^t\|_{\ell_2}^2.$$

Second step

The second step of Algorithm 8 updates the cluster centers based on the partition $(C_1^{t+1}, \dots, C_k^{t+1})$: for all $j \in [k]$,

$$\hat{\mu}_j^{t+1} \in \arg \min_{\mu \in \mathcal{X}} \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu\|_{\ell_2}^2.$$

Thus, in particular we have:

$$\sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_j^{t+1}\|_{\ell_2}^2 \leq \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_t^t\|_{\ell_2}^2.$$

Summing over all clusters $j \in [k]$, we obtain:

$$\sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_j^{t+1}\|_{\ell_2}^2 \leq \sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_t^t\|_{\ell_2}^2.$$

Outcome

Gathering the two steps, we get:

$$\begin{aligned} \hat{D}(C_1^{t+1}, \dots, C_k^{t+1}) &= \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_j^{t+1}\|_{\ell_2}^2 \\ &\leq \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^{t+1}} \|X - \mu_t^t\|_{\ell_2}^2 \\ &\leq \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j^t} \|X - \mu_j^t\|_{\ell_2}^2 \\ &= \hat{D}(C_1^t, \dots, C_k^t). \end{aligned}$$

□

Remark 2.2.6. First, we have no guarantee concerning the number of iterations the k -means algorithm needs in order to reach convergence. In fact, k -means might stop at a point which is not even a local minimum.

Second, there is no nontrivial lower bound on the gap between the value of the k -means objective for the partition returned by Algorithm 8 and the optimal k -means objective value.

The algorithm named k -means++ is an attempt to answer the first caveat of wrong initialization in k -means. To describe it (see Algorithm 9), let, for all $j \in [k]$, $j > 1$, Δ_j be the dissimilarity defined by:

$$\Delta_j: x \in \mathcal{X} \mapsto \min_{1 \leq \ell \leq j-1} \|x - \hat{\mu}_\ell\|_{\ell_2},$$

and let us denote δ_x the Dirac measure in $x \in \mathcal{X}$.

Algorithm 9 k-means++.

Input: $T \in \mathbb{N}$ (number of iterations), $\{X_i\}_{1 \leq i \leq n}$ (training sample).

$\hat{\mu}_1 \leftarrow$ random point from $\{X_i\}_{1 \leq i \leq n}$ (*initialization*)

for $j = 2$ **to** k **do**

$\hat{\mu}_j \leftarrow$ random point from $\{X_i\}_{1 \leq i \leq n}$ with density $\sum_{i=1}^n \frac{\Delta_j(\cdot)^2}{\sum_{\ell=1}^n \Delta_j(X_\ell)^2} \delta_{X_i}(\cdot)$

end for

$(C_1, \dots, C_k) \leftarrow$ output of k-means algorithm based on $(\hat{\mu}_1, \dots, \hat{\mu}_k)$

Output: (C_1, \dots, C_k) .

Remark 2.2.7. *Despite the lack of theoretical guarantees concerning k-means, another drawback is that clusters are not hierarchically built when k increases. A possible strategy for answering this point is hierarchical clustering, described later. In addition, when using Ward's cluster linkage, hierarchical clustering tries effectively to minimize the k-means objective.*

2.2.3 Point-based objectives

Contrarily to center-based objectives, point-based objectives do not require to compute a cluster center $\mu(C)$. The distortion $D(C_1, \dots, C_k)$ to minimize is computed based on pair of points belonging to the clusters. For example, the *sum of in-cluster distances* is

$$D(C_1, \dots, C_k) = \mathbb{E} \left(\sum_{j=1}^k d(X, Y) \mathbf{1}_{X \in C_j \cap Y \in C_j} \right),$$

and its empirical twin is

$$\hat{D}(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{X, Y \in \hat{C}_j} d(X, Y).$$

Let $s: \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ be a similarity measure. Another example lies in the distortion defined by the sum of interclass similarities:

$$D(C_1, \dots, C_k) = \mathbb{E} \left(\sum_{j=1}^k s(X, Y) \mathbf{1}_{X \in C_j \cap Y \notin C_j} \right).$$

With center-based objectives, the clustering approach focuses on one side of the intuitive definition of clustering: making sure that points in the same cluster are similar. On the contrary, the previous objective shed light on the other requirement: points separated into different clusters should be dissimilar.

When it comes to estimation, it is convenient to represent the relationships between training points by a similarity graph, in which each vertex represents a data point X_i and vertices are connected by an edge whose weight is their similarity. Such a graph can be defined by the similarity (or adjacency) matrix

$W = (s(X_i, X_j))_{1 \leq i, j \leq n}$. Given the index sets I_j of each empirical cluster \hat{C}_j , the previous point-based objective function has an empirical twin given by:

$$\hat{D}(C_1, \dots, C_k) = \sum_{j=1}^k \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}.$$

Minimizing $\hat{D}(C_1, \dots, C_k)$ is often referred as the *graph cut problem*.

2.2.4 Similarity graphs

To be a bit more formal, we consider a *similarity graph* $G = (V, E)$, for which the vertices $V = (v_1, \dots, v_n)$ represent the points (X_1, \dots, X_n) . Two vertices v_i and v_j are connected if the similarity $s(X_i, X_j) > 0$ (or greater than a prescribed threshold) and the edge between these two vertices is weighted by their similarity $s(X_i, X_j)$. The weighted adjacency matrix is $W = (s(X_i, X_j))_{1 \leq i, j \leq n}$.

The graph G is assumed undirected, which is equivalent to W being symmetric. In practice, this comes from considering a symmetric similarity measure s .

Definition 2.2.1. The degree of a vertex $v_i \in V$ is $d_i = \sum_{\ell=1}^n W_{i,\ell}$.

Given $A \subset V$, we call size of A the number of its vertices $|A|$ and volume of A $\text{vol}(A) = \sum_{i \in [n]; v_i \in A} d_i$.

A is said connected if any two vertices of A can be joined by a path such that all intermediate points also lie in A .

A is called a connected component if it is connected and if there are no connections between vertices in A and $V \setminus A$.

When constructing a similarity graph, the goal is to model the local neighborhood relationships between data points. In the forthcoming paragraphs, we describe four popular similarity graphs based on a given distance $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$.

The ϵ -neighborhood graph Two points X_i and X_j are connected if and only if their distance is less than a positive threshold ϵ : $d(X_i, X_j) \leq \epsilon$. If ϵ is small enough, all connected points are roughly at the same distance. Therefore, the weights assigned are $W_{i,j} = 1$ (if $d(X_i, X_j) \leq \epsilon$ and 0 otherwise). Given this rule, the ϵ -neighborhood graph is usually considered as an unweighted graph.

k-nearest neighbor graph Two points X_i and X_j are connected if and only if X_i is among the k -nearest neighbors of X_j or the other way around. Similarly to the ϵ -neighborhood graph, the weights are $W_{i,j} = 1$ if X_i and X_j are connected and 0 otherwise.

Mutual k-nearest neighbor graph Two points X_i and X_j are connected if and only if X_i is among the k -nearest neighbors of X_j and X_j is among the k -nearest neighbors of X_i . The weights are assigned similarly to the k-nearest neighbor graph.

The fully connected graph Points are connected if they have a positive similarity $s(X_i, X_j)$ and the edges are weighted by $s(X_i, X_j)$. It is important to note that since a similarity graph is supposed to reflect the local neighborhood relationships, the similarity should be defined accordingly. In practice, it is asked to s to be fulfill $s(x, x') = 1, \forall (x, x') \in X^2 : x = x'$ and to decrease quickly to 0 when x and x' differ. A popular choice is the Gaussian similarity: $s(x, x') = e^{-\frac{d(x, x')^2}{2\sigma^2}}$, in which σ^2 plays a role similar to ϵ and k for the similarity graphs introduced previously.

2.2.5 Spectral clustering

For $k = 2$, finding a minimal cut of a graph is a relatively easy problem and can be solved efficiently (for instance thanks to the Stoer-Wagner algorithm). However, it often results in separating an individual vertex from the rest of the graph, which is not satisfactory. Several solutions to this problem have been suggested but the most common ones are to normalize the empirical distortion either by the size of the clusters:

$$\hat{D}_r(C_1, \dots, C_k) = \sum_{j=1}^k \frac{1}{|\hat{C}_j|} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell},$$

(let us remind that $|\hat{C}_j| = |I_j|$) or by their volume:

$$\hat{D}_n(C_1, \dots, C_k) = \sum_{j=1}^k \frac{1}{\text{vol}(\hat{C}_j)} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}.$$

These objectives are respectively called *ratio cut* and *normalized cut*. Unfortunately, the balancing introduced by the cluster importance makes the minimization problem computationally hard to solve. Therefore, from now on, we describe a relaxation procedure resulting in the so called *spectral clustering* algorithm.

Definition 2.2.2 (Unnormalized graph Laplacian). Let $W \in \mathbb{R}^{n \times n}$ be a symmetric matrix. The diagonal matrix $D \in \mathbb{R}^{n \times n}$ such that $D_{i,i} = \sum_{j=1}^n W_{i,j}$, $\forall i \in [n]$ and $L = D - W$ are respectively called the degree matrix and the Laplacian of the graph defined by W .

Proposition 53. Let W and L be respectively the adjacency matrix and the Laplacian of the similarity graph of (X_1, \dots, X_n) . For any positive integer k and for all partitioning (C_1, \dots, C_k) of (X_1, \dots, X_n) , we have

$$\hat{D}_r(C_1, \dots, C_k) = \text{tr}(H^\top L H),$$

$$\text{where } H = \left(\frac{1}{\sqrt{|I_j|}} \mathbf{1}_{i \in I_j} \right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq k}}.$$

In addition, the columns of H are orthonormal to each other ($H^\top H = I$).

Proof. First, denoting $h_j \in \mathbb{R}^n$ the columns of H (for $j \in [k]$), we have

$$\text{tr}(H^\top LH) = \text{tr}((L^{1/2}H)^\top (L^{1/2}H)) = \sum_{j=1}^k (L^{1/2}h_j)^\top (L^{1/2}h_j) = \sum_{j=1}^k h_j^\top Lh_j.$$

In addition, for all $u \in \mathbb{R}^n$,

$$\begin{aligned} u^\top Lu &= u^\top Du - u^\top Wu \\ &= \sum_{1 \leq i \leq n} D_{i,i} u_i^2 - \sum_{1 \leq i, \ell \leq n} W_{i,\ell} u_i u_\ell \\ &= \frac{1}{2} \left(\sum_{1 \leq i \leq n} D_{i,i} u_i^2 + \sum_{1 \leq \ell \leq n} D_{\ell,\ell} u_\ell^2 - 2 \sum_{1 \leq i, \ell \leq n} W_{i,\ell} u_i u_\ell \right) \\ &= \frac{1}{2} \left(\sum_{1 \leq i, \ell \leq n} W_{i,\ell} u_i^2 + \sum_{1 \leq i, \ell \leq n} W_{i,\ell} u_\ell^2 - 2 \sum_{1 \leq i, \ell \leq n} W_{i,\ell} u_i u_\ell \right) \quad (W_{i,\ell} \text{ symmetric}) \\ &= \frac{1}{2} \sum_{1 \leq i, \ell \leq n} W_{i,\ell} (u_i - u_\ell)^2. \end{aligned}$$

Therefore, for all $j \in [k]$,

$$\begin{aligned} h_j^\top Lh_j &= \frac{1}{2} \sum_{1 \leq i, \ell \leq n} W_{i,\ell} (H_{i,j} - H_{\ell,j})^2 \\ &= \frac{1}{2} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} \frac{W_{i,\ell}}{|I_j|} + \frac{1}{2} \sum_{\substack{i \notin I_j \\ \ell \in I_j}} \frac{W_{i,\ell}}{|I_j|} \\ &= \frac{1}{|I_j|} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}, \end{aligned}$$

since $H_{i,j} - H_{\ell,j}$ is nonzero only if $i \in I_j$ and $\ell \notin I_j$ or the other way around.

Gathering everything, we have:

$$\text{tr}(H^\top LH) = \sum_{j=1}^k h_j^\top Lh_j = \sum_{j=1}^k \frac{1}{|I_j|} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell} = \hat{D}_r(C_1, \dots, C_k).$$

□

Remark 2.2.8. Up to normalization, H represents the one-hot-encoding of the clusters. For example, for $k = 3$, if we reorganize the sample (X_1, \dots, X_n) such that \hat{C}_1 appears first, then \hat{C}_2 and so on,

we get

$$H = \begin{pmatrix} \frac{1}{|\hat{C}_1|} & 0 & 0 \\ \vdots & \vdots & \vdots \\ \frac{1}{|\hat{C}_1|} & 0 & 0 \\ 0 & \frac{1}{|\hat{C}_2|} & 0 \\ \vdots & \vdots & \vdots \\ 0 & \frac{1}{|\hat{C}_2|} & 0 \\ 0 & 0 & \frac{1}{|\hat{C}_3|} \\ \vdots & \vdots & \vdots \\ 0 & 0 & \frac{1}{|\hat{C}_3|} \end{pmatrix}.$$

Owing to Proposition 53, the ratio cut problem

$$\underset{(\hat{C}_1, \dots, \hat{C}_k) \in P(\{X_1, \dots, X_n\})}{\text{minimize}} \sum_{j=1}^k \frac{1}{|I_j|} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}$$

is equivalent to

$$\begin{aligned} & \underset{H \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad \text{tr}(H^\top L H) \\ & \text{s. t.} \quad \begin{cases} H^\top H = I \\ \forall j \in [k], \forall i \in [n]: H_{i,j} \in \left\{ 0, \frac{1}{\sqrt{|I_j|}} \right\}. \end{cases} \end{aligned}$$

Unfortunately, two difficulties arise: first, this is an integer programming problem which we may not be able to solve efficiently. Second, the values $(|I_1|, \dots, |I_k|)$ are not known in advance. Therefore, we relax the problem by discarding the last constraint. Unnormalized spectral clustering boils down to solving

$$\begin{aligned} & \underset{H \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad \text{tr}(H^\top L H) \\ & \text{s. t.} \quad H^\top H = I. \end{aligned} \tag{P21}$$

As we will see in Chapter 3 (Theorem 59), (P21) is solved by the matrix H for which the columns are the minor eigenvectors of L . The resulting algorithm (see Algorithm 10) is called *Unnormalized spectral clustering*. It proceeds by mapping the data (X_1, \dots, X_n) to the rows of the k minor eigenvectors of L and then by performing a vanilla k-means.

Proposition 54. *Let W and L be respectively the adjacency matrix and the Laplacian of the similarity graph of (X_1, \dots, X_n) . For any positive integer k and for all partitioning (C_1, \dots, C_k) of (X_1, \dots, X_n) , we have*

$$\hat{D}_n(C_1, \dots, C_k) = \text{tr}(H^\top L H),$$

$$\text{where } H = \left(\frac{1}{\sqrt{\text{vol}(\hat{C}_j)}} \mathbf{1}_{i \in I_j} \right)_{\substack{1 \leq i \leq n \\ 1 \leq j \leq k}}.$$

Algorithm 10 Unnormalized spectral clustering.

Input: $W \in \mathbb{R}^{n \times n}$ (adjacency matrix).

$L \leftarrow$ Laplacian of W

$H \leftarrow k$ minor eigenvectors of L as columns

$Y_i \leftarrow i^{\text{th}}$ row of H (for all $i \in [n]$) ($Y_i \in \mathbb{R}^k$)

$(\hat{C}_1, \dots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on (Y_1, \dots, Y_n)

Output: $(\hat{C}_1, \dots, \hat{C}_k)$.

In addition, the columns of $D^{\frac{1}{2}}H$ are orthonormal to each other ($H^\top DH = I$).

The proof is a good exercise.

Owing to Proposition 54, the normalized cut problem

$$\underset{(\hat{C}_1, \dots, \hat{C}_k) \in P(\{X_1, \dots, X_n\})}{\text{minimize}} \sum_{j=1}^k \frac{1}{\text{vol}(\hat{C}_j)} \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}$$

is equivalent to

$$\begin{aligned} & \underset{H \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad \text{tr}(H^\top LH) \\ & \text{s. t.} \quad \begin{cases} H^\top DH = I \\ \forall j \in [k], \forall i \in [n]: H_{i,j} \in \left\{ 0, \frac{1}{\sqrt{\text{vol}(\hat{C}_j)}} \right\}, \end{cases} \end{aligned}$$

and can be relaxed to

$$\begin{aligned} & \underset{H \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad \text{tr}(H^\top LH) \\ & \text{s. t.} \quad H^\top DH = I. \end{aligned} \tag{P22}$$

Assuming that all points are connected, D is invertible and (P22) can be reformulated

$$\begin{aligned} & \underset{H \in \mathbb{R}^{n \times k}}{\text{minimize}} \quad \text{tr}(U^\top L_s U) \\ & \text{s. t.} \quad \begin{cases} H = D^{-\frac{1}{2}} U \\ U^\top U = I, \end{cases} \end{aligned}$$

where $L_s = D^{-\frac{1}{2}} L D^{-\frac{1}{2}}$. Therefore (P22) is solved by the matrix U for which the columns are the minor eigenvectors of L_s , which corresponds to H for which the columns are the minor eigenvectors of $L_w = D^{-1} L$ (see below). The resulting algorithm (see Algorithm 11) is called *Normalized spectral clustering*.

Remark 2.2.9. $\lambda \in \mathbb{R}_+$ is eigenvalue of L_w with eigenvector u if and only if λ and u solve the generalized eigenvalue problem $Lu = \lambda Du$.

Algorithm 11 Normalized spectral clustering (with L_w).

Input: $W \in \mathbb{R}^{n \times n}$ (adjacency matrix).

$L_w \leftarrow$ Laplacian of W

$H \leftarrow k$ minor eigenvectors of L_w as columns (similar to the generalized eigenproblem $Lu = \lambda Du$)

$Y_i \leftarrow i^{th}$ row of H (for all $i \in [n]$) ($Y_i \in \mathbb{R}^k$)

$(\hat{C}_1, \dots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on (Y_1, \dots, Y_n)

Output: $(\hat{C}_1, \dots, \hat{C}_k)$.

Remark 2.2.10. Comparing ratio cut and normalized cut leads to a very interesting discovery. First, let us remark that, as already mentioned, both objective functions encodes the second part of the intuitive definition of clustering: points separated into different clusters should be dissimilar.

In addition, the balancing introduced takes into account the importance of the clusters, either through their size or their volume. This is so since minimizing the min cut objectives leads to minimizing the cuts between the clusters while maximizing their importance. However, ratio cut and normalized cut behave differently concerning cluster importance. Indeed, it is easy to see that, for all $j \in [k]$:

$$\sum_{\substack{i \in I_j \\ \ell \in I_j}} W_{i,\ell} = \text{vol}(\hat{C}_j) - \sum_{\substack{i \in I_j \\ \ell \notin I_j}} W_{i,\ell}.$$

In other words, the intra-cluster similarity is maximized as soon as the volume of the cluster is maximized and the cut with the rest of the vertices is minimized; which is what is achieved by normalized cut minimization. On the other hand, the size $|\hat{C}_j|$ of a cluster is not necessarily related to the intra-cluster similarity.

In this sense, normalized cut minimization addresses both parts of the clustering definition.

Moreover, it can be shown that, L_w behaves as expected when $n \rightarrow \infty$ and so it is for the resulting partitioning provided by normalized spectral clustering. On the contrary, L can lead to completely unreliable results, even for small sample size [von Luxburg, 2007].

There exists another popular normalized spectral clustering algorithm (see Algorithm 12) based on the third Laplacian that popped up during this analysis: L_s .

Algorithm 12 Normalized spectral clustering (with L_s).

Input: $W \in \mathbb{R}^{n \times n}$ (adjacency matrix).

$L_s \leftarrow$ Laplacian of W

$H \leftarrow k$ minor eigenvectors of L_s as columns

$Y_i \leftarrow i^{th}$ row of H normalized to 1 (for all $i \in [n]$) ($Y_i \in \mathbb{R}^k$, $\sum_{j=1}^k (Y_i)_j^2 = 1$)

$(\hat{C}_1, \dots, \hat{C}_k) \leftarrow$ output of k-means algorithm based on (Y_1, \dots, Y_n)

Output: $(\hat{C}_1, \dots, \hat{C}_k)$.

Remark 2.2.11. First, there is no theoretical guarantees concerning the “quality” of these two relaxations.

Second, there exist many other relaxations. Some of them rely on semidefinite programming.

Last but not least, spectral relaxations are not appealing for the quality of the solutions they provide but for the simplicity of the problem in which they results (standard linear algebra – eigenvalue – problems).

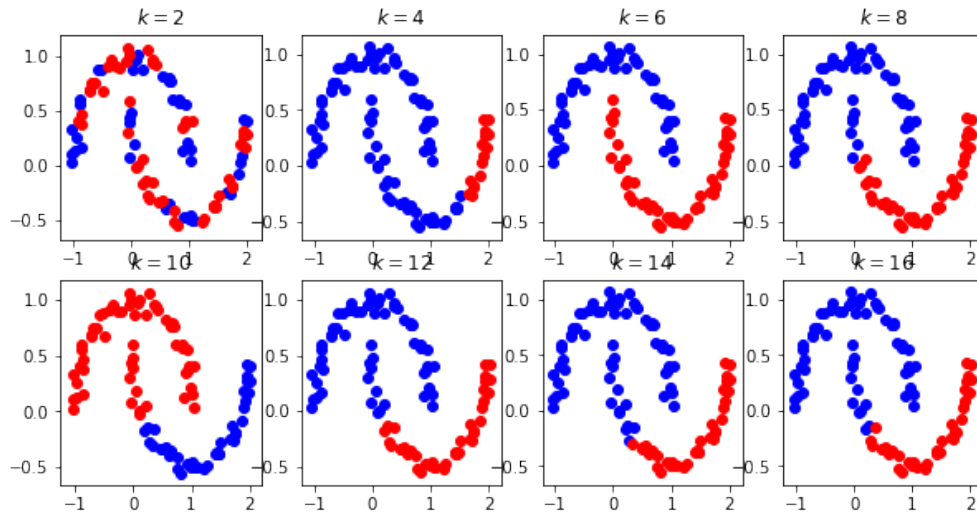


Figure 2.6: Example of spectral clustering (k-nearest neighbor graph).

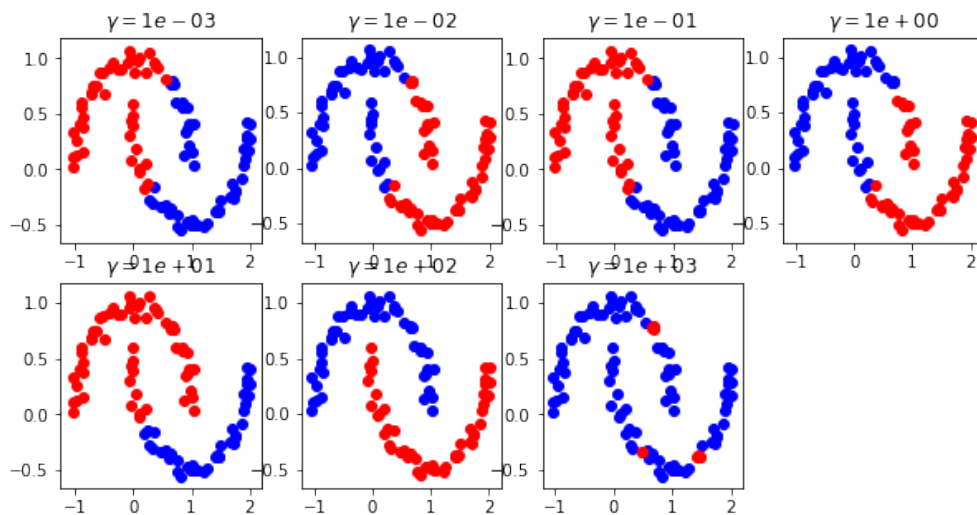


Figure 2.7: Example of spectral clustering (fully connected Gaussian graph).

2.2.6 Properties of graph Laplacians

Let us consider $W \in \mathbb{R}_+^{n \times n}$ a symmetric adjacency matrix and $D \in \mathbb{R}^{n \times n}$ its degree matrix. So far, we have seen three Laplacians, summed up in the following definition.

Definition 2.2.3.

Unnormalized Laplacian: $L = D - W$;

Normalized Laplacian 1: $L_s = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$;

Normalized Laplacian 2: $L_w = D^{-1} L = I - D^{-1} W$.

Normalized Laplacians are subscripted by s and w because they are respectively symmetrically normalized by $D^{-\frac{1}{2}}$ (on left and right) and whitened by D .

Graph Laplacians have many properties. In the next proposition, we recover in particular the two properties bridging the gap between graph cut and eigenvalue decomposition (Item 1) and discover that 0 is an eigenvalue of L and L_w with eigenvector $\mathbf{1}$.

Proposition 55.

1. One has, $\forall u \in \mathbb{R}^n$:

$$u^\top L u = \frac{1}{2} \sum_{1 \leq i, \ell \leq n} W_{i, \ell} (u_i - u_\ell)^2$$

$$u^\top L_s u = \frac{1}{2} \sum_{1 \leq i, \ell \leq n} W_{i, \ell} \left(\frac{u_i}{\sqrt{D_{i, i}}} - \frac{u_\ell}{\sqrt{D_{\ell, \ell}}} \right)^2.$$

2. 0 is eigenvalue of L and L_w with eigenvector $\mathbf{1}$. 0 is eigenvalue of L_s with eigenvector $D^{\frac{1}{2}} \mathbf{1}$.
3. $\lambda \in \mathbb{R}_+$ is eigenvalue of L_w with eigenvector u if and only if λ is eigenvalue of L_s with eigenvector $D^{\frac{1}{2}} u$.
4. L , L_s and L_w are symmetric PSD matrices.

The proof is a good exercise.

Proposition 56. Let G be an undirected graph with non-negative weights. Then, the multiplicities of the eigenvalue 0 of L , L_s and L_w are the same and equal the number k of connected components (A_1, \dots, A_k) in G .

In addition, the eigenspace of 0 for both L and L_w is spanned by $\{\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_k}\}$ and the eigenspace of 0 for L_s is spanned by $\{D^{-\frac{1}{2}} \mathbf{1}_{A_1}, \dots, D^{-\frac{1}{2}} \mathbf{1}_{A_k}\}$.

We refer to [von Luxburg, 2007, Fig. 1] for an illustration of eigenvector properties.

2.2.7 Practical details

Similarity graph ϵ -neighborhood cannot handle *different scales* (different distances between data points) in different regions of the space. k -nearest neighbor graph can and connects regions of high and low densities. On the contrary, mutual k -nearest neighbor graph does not connect regions of high and low densities. It can be used to detect clusters of different densities.

k -nearest neighbor graph is a good starting point.

Connectivity parameter For k -nearest neighbor graph, choose k such that the graph is connected or has significantly fewer connected components than clusters to detect. Otherwise, spectral clustering will trivially return connected components as clusters. Some asymptotic connectivity results suggest to choose k in the order of $\log(n)$.

Very generally, we can observe that the mutual k -nearest neighbor graph has much fewer edges than the k -nearest neighbor graph for the same parameter k . This suggest to choose k larger for the mutual k -nearest neighbor graph.

For the ϵ -neighborhood graph, ϵ should be chosen such that the graph is connected. The smallest value of ϵ for which the graph is connected can be estimated by the length of the longest edge in a minimal spanning tree covering the fully connected graph of the data. However this method is very sensitive to outliers and isolated tight clusters.

For a fully connected graph, σ can be chose as the mean distance of a point to its k -nearest neighbors, where k is chosen similarly as above (k of the order $\log(n)$).

Number of clusters The gap heuristic: choose k such that all eigenvalues $\lambda_1, \dots, \lambda_k$ are very small and λ_{k+1} is relatively large (see Figures 2.8 and 2.9). A justification of this procedure, coming from perturbation theory, is that in the ideal case of k completely disconnected clusters, the eigenvalue 0 has multiplicity k and $\lambda_{k+1} > 0$.

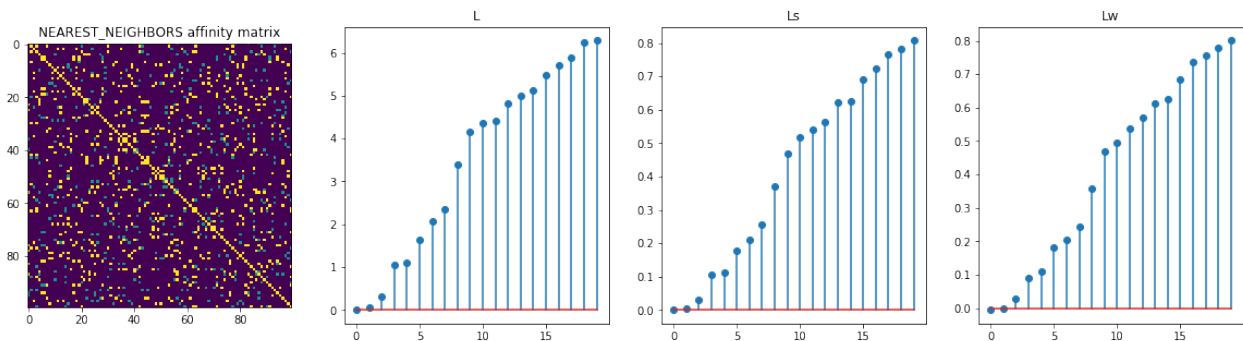


Figure 2.8: Example of Laplacian eigenvalues (k -nearest neighbor graph).

Graph Laplacian One may look at the degree distribution of the similarity graph. If most vertices have approximately the same degree, then all graph Laplacians will perform equally. However, if the degrees

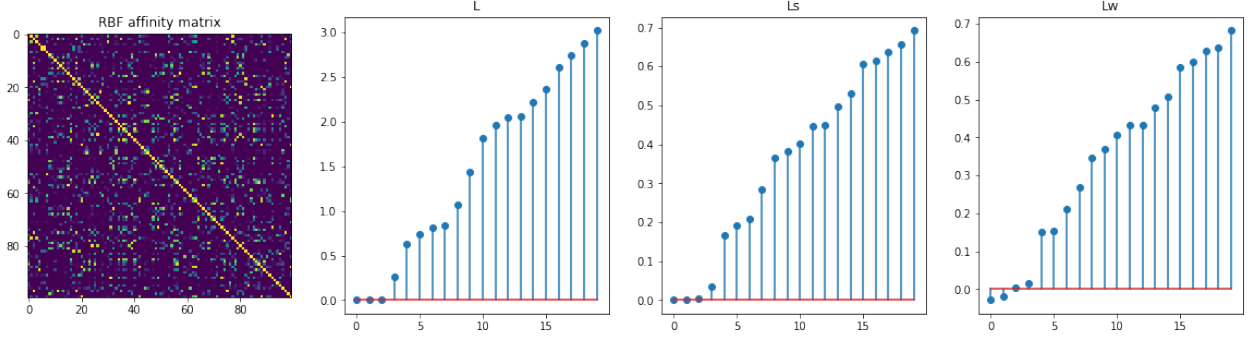


Figure 2.9: Example of Laplacian eigenvalues (fully connected Gaussian graph).

of the graph are very broadly distributed, then Laplacians differ considerably and we suggest to choose unnormalized Laplacians, and particularly L_w rather than L_s .

This choice is justified first by Remark 2.2.10.

2.3 Hierarchical clustering

Hierarchical methods for clustering aim at answering a major drawback of k-means: the lack of hierarchy in clusters (*i.e.* decreasing k does not lead to merging clusters). This section introduces very simple methods based on measuring the similarity (or linkage) between clusters. We focus on *agglomerative* approaches (which are based on merging clusters) and put *divisive* ones aside (based on splitting clusters).

2.3.1 Agglomerative approaches

Linkage-based methods are probably the simplest and most intuitive paradigm of clustering. In their agglomerative version, they start from the partitioning of the training set (X_1, \dots, X_n) in which each cluster is a unit set $\{X_i\}$ (for $i \in [n]$) and merge successively the *closest* clusters. Straightforwardly, the number of clusters decreases at each iteration and clusters are nested: each cluster \hat{C}^t at iteration t is either the same as at iteration $t-1$ ($\hat{C}^t = \hat{C}^{t-1}$) or the union of two previous clusters ($\hat{C}^t = \hat{C}_1^{t-1} \cup \hat{C}_2^{t-1}$).

Two parameters need to be defined in such a procedure: the (dis)similarity (or linkage) between two clusters and the merging stopping rule. To make the first point precise, let $d: \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_+$ be a dissimilarity and consider two subsets A and B of (X_1, \dots, X_n) . We now give common examples of cluster dissimilarities $D: P(\{X_1, \dots, X_n\})^2 \rightarrow \mathbb{R}_+$.

Single linkage

$$D(A, B) = \min_{x \in A, y \in B} d(x, y).$$

Complete linkage

$$D(A, B) = \max_{x \in A, y \in B} d(x, y).$$

Average linkage

$$D(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y).$$

Ward's minimum variance

Given the intraclass inertia for a generic subset $C \subset (X_1, \dots, X_n)$:

$$I(C) = \sum_{x \in C} d(x, m_C)^2,$$

where $m_C = \frac{1}{|C|} \sum_{y \in C} y$, the cluster distance in Ward's method is

$$D(A, B) = I(A \cup B) - I(A) - I(B),$$

which is the increase of intraclass inertia when merging A and B . For the Euclidean distance,

$$D(A, B) = \frac{|A||B|}{|A| + |B|} \|m_A - m_B\|_{\ell_2}^2.$$

Since Ward's method merges clusters by minimizing the increase in the total intraclass inertia, it is very similar to k-means but approximates a minimizer of the k-means objective with an agglomerative hierarchical procedure.

Remark 2.3.1. *Linkage methods can be used with a variety of distances (or affinities), in particular:*

- ◇ *Euclidean distance (or ℓ_2);*
- ◇ *Manhattan distance (or Cityblock, or ℓ_1);*
- ◇ *cosine distance;*
- ◇ *any precomputed affinity matrix.*

If the agglomerative procedure runs until the end, all points share the same large cluster. The resulting sequence of partitioning can be represented as a tree, called a dendrogram, the root of which is the unique cluster that gathers all points (the final cluster) and the leaves of which are the unit set clusters (algorithm initialization).

If one is more interested in a useful partitioning instead of the clustering dendrogram, one needs to employ a stopping, which may be:

- ◇ a fixed number of clusters;
- ◇ a distance upper bound \bar{D} (or alternatively a *scaled distance upper bound* $\alpha \in \mathbb{R}_+$ such that $\bar{D} = \alpha \max_{1 \leq i, j \leq n} d(X_i, X_j)$ for single, complete and average linkages).

2.3.2 Connection with minimum spanning trees

Given a connected edge-weighted undirected graph $G = (V, E)$, with weight function $d: V \times V \rightarrow \mathbb{R}_+$, the problem of minimum spanning tree (MST) is to find a subgraph $T = (V, E')$ that connects all vertices

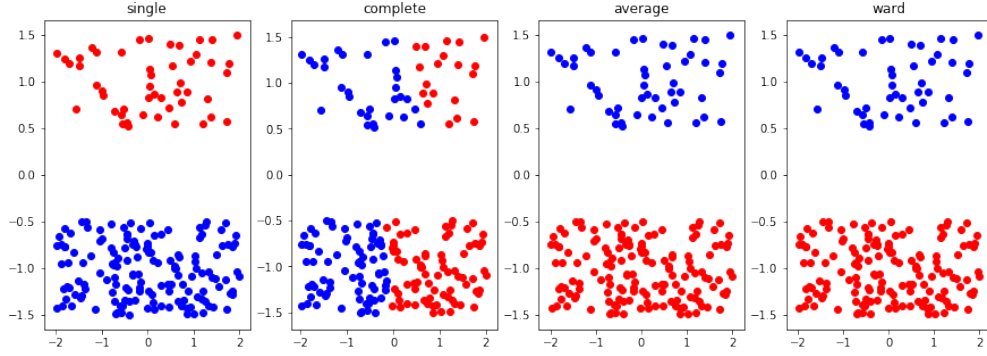


Figure 2.10: Example of agglomerative clustering (two rectangles).

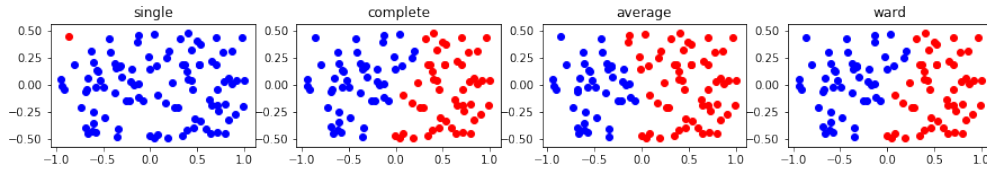


Figure 2.11: Example of agglomerative clustering (single rectangle).

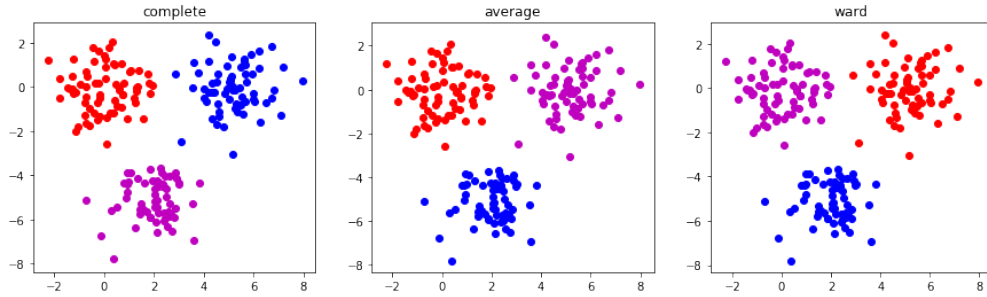


Figure 2.12: Example of agglomerative clustering (Gaussian clusters).

with minimal sum of weights $\sum_{\{u,v\} \in E'} d(u, v)$. It is easy to see that such a subgraph T is necessary a tree. Indeed, if there was a cycle in T , we could remove an edge on the cycle to get a new subgraph connecting all vertices and with fewer sum of weights.

There are several classic techniques for finding an MST: Borůvka's, Kruskal's, Prim's and reverse-delete algorithms. All are greedy methods. In particular, Kruskal's algorithm consists in adding edges in increasing weight (at the initialization, we consider an edge with minimal weight), skipping those whose addition would create a cycle.

Back to clustering and similarly to spectral methods, let us consider the similarity graph, in which each vertex represents a data point and vertices X_i and X_j are all connected by an edge whose weight is their distance $d(X_i, X_j)$. Then, applying Kruskal's algorithm to this graph is exactly performing a single linkage agglomerative clustering on the set (X_1, \dots, X_n) . Indeed, picking edges in increasing weight in Kruskal's algorithm corresponds to merging the closest clusters in single linkage. In addition, discarding edges that would create a cycle is exactly saying that we measure distances and merge two different

clusters in single linkage.

To complete the comparison, once an MST T is created, deleting the $k - 1$ most expensive edges in T produces k connected subgraphs, which are exactly the clusters produced by single linkage.

2.4 Density-based clustering

The algorithm called density-based spatial clustering of applications with noise (DBSCAN) assumes that clusters are dense regions separated by low-density corridors. It is one of the most common clustering algorithms because of its efficiency and its ability to automatically determine the number of clusters.

Given two parameters, a radius $\epsilon > 0$ and a minimal number of neighbors m , DBSCAN considers three types of points:

1. core points are points that have at least m neighbors within a distance ϵ (the ϵ -neighborhood), including themselves (the ϵ -neighborhood is at least a unit set);
2. reachable points are non-core points that fall in the neighborhood of a core point;
3. outliers are other points.

Clusters are formed by core points that fall in the neighborhoods of each other and by their reachable points.

Remark 2.4.1. *With $m = 2$, DBSCAN performs the same clustering as single linkage for which the dendrogram has been cut at height ϵ .*

Advantages

- ◇ DBSCAN makes assumption on cluster density, thus it determines automatically the number of clusters.
- ◇ There is no shape restriction for discovered clusters (while k-means requires convex clusters).
- ◇ DBSCAN prevents the single-link effect (different clusters being connected by a thin line of points).
- ◇ There is a notion of outliers/noise.
- ◇ DBSCAN is mostly insensitive to sample ordering.

Drawbacks

- ◇ DBSCAN is only deterministic on core and noise points. Border points that are reachable from several clusters can be part of either cluster, depending on the order the data is processed.
- ◇ DBSCAN cannot cluster data sets well with large differences in densities, since the parameters ϵ and m cannot be chosen appropriately for all clusters.

Choice of the parameters

1. Because of the curse of dimensionality, m should be chosen of the order of $2d$.
2. The radius ϵ can be chosen at the elbow of the monotonic curve of maximum distances between a point and its $m - 1$ nearest neighbors.

Algorithm 13 DBSCAN.

Input: $\epsilon > 0$ (neighborhood radius), $m \in \mathcal{N}$ (minimal number of neighbors), $\{X_i\}_{1 \leq i \leq n}$ (training sample).

```
 $T \leftarrow \{X_i\}_{1 \leq i \leq n}$  (unlabeled points)
 $k \leftarrow 0$  (current number of clusters)
while  $T \neq \emptyset$  do
  pick  $X$  in  $T$ 
   $N \leftarrow \epsilon$ -neighborhood of  $X$ 
  if  $|N| \geq m$  then
     $k \leftarrow k + 1$ 
    initialize a new cluster  $\hat{C}_k = \emptyset$ 
    move  $X$  from  $T$  to  $\hat{C}_k$ 
     $S \leftarrow (N \setminus \{X\}) \cap T$  (unlabeled neighbors)
    while  $S \neq \emptyset$  do
      pick  $Y$  in  $S$ 
      move  $Y$  from  $S$  to  $\hat{C}_k$  (and remove  $Y$  from  $T$ )
       $N' \leftarrow \epsilon$ -neighborhood of  $Y$ 
      if  $|N'| \geq m$  then
         $S \leftarrow S \cup (N' \cap T)$  (unlabeled neighbors)
      end if
    end while
  end if
end while
Output:  $(\hat{C}_1, \dots, \hat{C}_k, T)$  ( $k$  clusters and a set of outliers)
```

DBSCAN in action

1. [A fancy demo of DBSCAN.](#)

2.5 Clustering evaluation

When the ground truth is known, several criteria can be used for evaluating a clustering performance: adjusted rand index, normalized and adjusted mutual informations, V-measure, Fowlkes–Mallows score. All scores measure the similarity between class labels ignoring permutation. Besides, adjusted indexes and Fowlkes–Mallows score have chance normalization: random uniform label assignment gets a 0 score.

Now, we briefly present some methods for assessing a clustering performance when the ground truth is not known. These methods can be used to select the number of clusters k or other parameters.

2.5.1 Elbow method

The elbow method is a method of validation of a partitioning based on the intraclass inertia. It is often used to choose an appropriate number of clusters and is particularly suited for convex clusters (due to the nature of its criterion).

The elbow method looks at the intraclass inertia (or inversely at the percentage of variance explained, that is the ratio of the interclass inertia to the total inertia) as a function of the number of clusters (see Figures 2.13 and 2.14).

The idea of the elbow method is that such a curve has two regimes:

- ◇ going from 1 to 2 (then 3) clusters will decrease a lot the intraclass inertia since these clusters help discovering groups;
- ◇ once we have more clusters than actual groups, there is a very limited gain (in the intraclass inertia) of adding a new cluster.

Therefore, intuitively, there should be an angle in the graph, separating the two regimes mentioned above. The point where this angle appears is called an *elbow* and precisely indicates the number of clusters to choose. However, in practice this elbow cannot always be unambiguously identified.

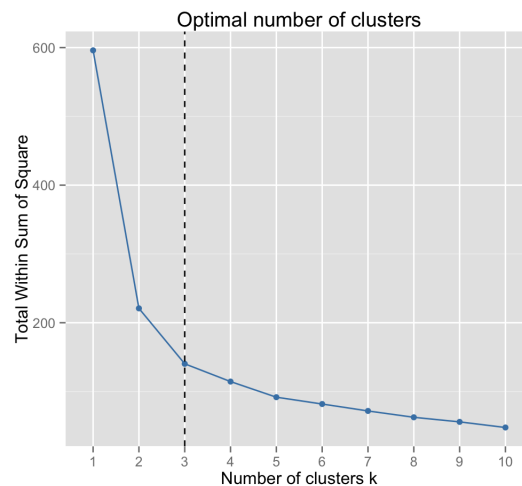


Figure 2.13: The elbow method (in theory). Courtesy of WWW.

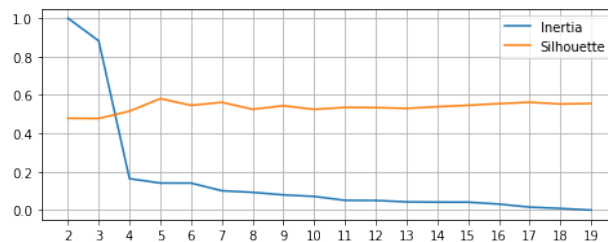


Figure 2.14: The elbow method (in practice).

2.5.2 Silhouette coefficient

The higher, the better.

For all $i \in [i]$, let \hat{C}_j be the cluster associated to X_i and

$$a_i = \frac{1}{|\hat{C}_j| - 1} \sum_{\substack{Y \in \hat{C}_j \\ Y \neq X_i}} d(X_i, Y),$$

as well as

$$b_i = \min_{\substack{1 \leq \ell \leq k \\ \ell \neq j}} \frac{1}{|\hat{C}_\ell|} \sum_{Y \in \hat{C}_\ell} d(X_i, Y)$$

being respectively the average distance of X_i to its companions and to the members of the *neighboring cluster*. These values can be interpreted as how well X_i is compliant with its cluster and different from the neighboring cluster. The silhouette value for X_i is

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)} \in [-1, 1].$$

The average s_i over all data of a cluster $\frac{1}{|\hat{C}_j|} \sum_{\substack{1 \leq i \leq n \\ X_i \in \hat{C}_j}} s_i$ measures how tightly grouped all the data in the cluster are, and how distant from the neighboring cluster they are. In this sense, it is a density-based index, which is close to 1 when \hat{C}_j is a dense group separated by a low-density corridor from its neighbor (similarly to the operating of DBSCAN).

The silhouette coefficient is well suited for choosing the number k of clusters: if there are too many or too few clusters, some of them will typically display much narrower silhouettes than the others, meaning that they should or should not be split (see for instance [this interesting example](#)).

Averaging the silhouette coefficients over all data produces a measure of how appropriate is the partitioning:

$$s = \frac{1}{n} \sum_{i=1}^n s_i.$$

2.5.3 Calinski-Harabasz index

Another density-based index is the Calinski-Harabasz coefficient, which boils down to a normalized ratio of the between-cluster dispersion (or interclass inertia) and the within-cluster dispersion (or intraclass inertia):

$$s = \frac{n - k}{k - 1} \frac{b}{w},$$

where b is the between-cluster dispersion

$$b = \sum_{j=1}^k \frac{|\hat{C}_j|}{n} \|\hat{\mu} - \hat{\mu}_j\|_{\ell_2}^2,$$

and w is the within-cluster dispersion:

$$w = \frac{1}{n} \sum_{j=1}^k \sum_{X \in \hat{C}_j} \|X - \hat{\mu}_j\|_{\ell_2}^2,$$

where μ is the global mean and $\hat{\mu}_j$ is the mean of the cluster \hat{C}_j .

Remark 2.5.1. *Paradoxically, both previous indexes are generally higher for convex clusters than for other concepts of clusters, such as those produced by DBSCAN.*

Chapter 3

Dimensionality reduction

Dimensionality reduction is related to the concept of lossy compression in information theory. It consists in transforming data from a high-dimensional space to new data from a lower-dimensional space, with as few loss of information as possible.

Dimensionality reduction is motivated by:

- ◇ computational challenges;
- ◇ poor generalization ability (for example, for nearest neighbors classifiers, the sample complexity increases exponentially with the dimension);
- ◇ interpretability of the data (finding a meaningful structure, displaying the data).

Theorem 57 ([Shalev-Shwartz and Ben-David, 2014, Theorems 19.4 and 19.5]). *Let $\mathcal{X} = [0, 1]^d$, $\mathcal{Y} = \{\pm 1\}$ and $k \in \mathbb{N}$, $k \geq 10$. For any distribution D over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function given X is L -Lipschitz continuous ($L > 0$) and any sample $\{(X_i, Y_i)\}_{1 \leq i \leq n}$ iid according to D , one has*

$$\mathbb{P}(g_n(X) \neq Y) \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathbb{P}(g^*(X) \neq Y) + \frac{6L\sqrt{d} + k}{n^{\frac{1}{d+1}}},$$

where g_n is the k -nearest neighbors rule and g^* is the Bayes classifier.

In addition, for all $L > 1$ and any integer $k > 0$, there exists a distribution over $\mathcal{X} \times \mathcal{Y}$ for which the conditional probability function given X is L -Lipschitz continuous, $\mathbb{P}(g^*(X) \neq Y) = 0$ and for any sample of size $n \leq \frac{(L+1)^d}{2}$,

$$\mathbb{P}(g_n(X) \neq Y) \geq \frac{1}{4},$$

where g_n and g^* are defined as above.

It is easy to see that for the last term of the generalization bound in Theorem 57 to be smaller than $\epsilon > 0$, we should have $n \geq \left(\frac{6L\sqrt{d} + k}{\epsilon}\right)^{d+1}$. That is, the size of the training set should increase exponentially

with the dimension. Besides, the rest of Theorem 57 tells us that this is not just an artifact of our upper bound, since there exists distributions for which we need exponentially big training samples to get low errors. This phenomenon is often referred to as *curse of dimensionality*.

Other examples of the curse of dimensionality

Sampling

Sampling evenly a unit hypercube with a lattice that has a spacing of $\epsilon \in (0, 1]$, requires ϵ^{-d} points. For instance, for $\epsilon = 10^{-2}$, 100 points are required to sample the segment $[0, 1]$ while 10^{20} points are needed for the 10-dimensional unit hypercube.

Reciprocally, 100 points represent well the segment $[0, 1]$, while they are really insufficient for “covering” $[0, 1]^{10}$.

Distance functions

Given a radius $r > 0$, the ratio of the volumes of an inscribed hypersphere with radius r (in dimension d) to a hypercube with edges of length $2r$ is given by

$$\frac{\frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}}{(2r)^d} = \frac{\pi^{d/2}}{2^{d-1} d\Gamma(d/2)} = \left(\frac{\sqrt{\pi}}{2} \right)^d \frac{2}{d\Gamma(d/2)},$$

where Γ is the gamma function and $\sqrt{\pi}/2 \approx 0.9$. It is easy to see that this quantity goes exponentially fast to 0 when d grows. As a consequence, we are used to say that “points are concentrated in the corners” of the hypercube. This assertion is enforced by the fact that the distance between a corner and the hypercube center is $r\sqrt{d}$. In this sens, nearly all of the high-dimensional space is far away from the centre.

The Volume is in a narrow Annulus

The ratio of the volume of a sphere of radius $(1 - \epsilon)r$ ($\epsilon \in [0, 1]$, $r > 0$) to the volume of a sphere of radius r is

$$\frac{\frac{2((1-\epsilon)r)^d \pi^{d/2}}{d\Gamma(d/2)}}{\frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}} = (1 - \epsilon)^d,$$

which decreases exponentially to 0 as d goes to infinity. In other words, in high dimensions, all of the volume of the sphere is concentrated in a narrow annulus at the surface.

Going back to dimensionality reduction, this chapter mainly focuses on linear methods, that is, finding a matrix $W \in \mathbb{R}^{p \times d}$, where d is the input dimension and p is the desired reduced dimension, that induces the mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$. Without supervised information, a natural criterion for choosing W is that the reduction mapping enables a reasonable recovery of the original data x .

The fundamental hypothesis underlying dimensionality reduction is that data does not fill the entire space but lives in a manifold of small dimension. Informally, a manifold is a (topological) space, that is locally homeomorphic to the Euclidean space of dimension p (which is also the dimension of the manifold). The

locally feature means that for any point, there exists a neighborhood (hence the need of the topology), that is homeomorphic to the Euclidean space.

There are two points of view in dimensionality reduction:

1. there is a manifold $\mathcal{M} \subseteq \mathbb{R}^d$ of dimension p lower than d , in which we aim at “projecting” the data with few distortion of the geometry;
2. there is a manifold $\mathcal{M} = f(\mathcal{N})$, where \mathcal{N} is a manifold from \mathbb{R}^p (the latent manifold) and $f: \mathcal{N} \rightarrow \mathcal{M}$ is a homeomorphism with particular features such as isometry. Then, we look for a latent variable $y \in \mathcal{N}$ such that $x \approx f(y)$.

3.1 Linear methods

3.1.1 Principal component analysis

As explained before, the most straightforward approach of dimensionality reduction is to find a compression matrix $W \in \mathbb{R}^{p \times d}$, where $p < d$ is the reduced dimension, such that the (linear) mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ enables a reasonable recovery of the original data. That is, there exists a recovery matrix $U \in \mathbb{R}^{d \times p}$ such that the mapping $x \in \mathbb{R}^d \mapsto UWx \in \mathbb{R}^d$ is almost the identity.

Given the random variable of interest $X \in \mathbb{R}^d$, a natural variational formulation, of dimensionality reduction is

$$\underset{W \in \mathbb{R}^{p \times d}, U \in \mathbb{R}^{d \times p}}{\text{minimize}} \quad \mathbb{E} \left(\|X - UWX\|_{\ell_2}^2 \right). \quad (\text{P23})$$

Lemma 58. *If $p \leq d$ and (P23) admits a solution, then there exists $V \in \mathbb{R}^{d \times p}$, such that $V^\top V = I_p$ and the pair (V^\top, V) is also solution to (P23).*

Proof. Let $(W, U) \in \mathbb{R}^{d \times p} \times \mathbb{R}^{p \times d}$ be solution to (P23). We will show that there exists $V \in \mathbb{R}^{d \times p}$, such that $V^\top V = I_p$ and

$$\mathbb{E} \left(\|X - UWX\|_{\ell_2}^2 \right) \geq \mathbb{E} \left(\|X - VV^\top X\|_{\ell_2}^2 \right),$$

which proves that (V^\top, V) is also solution to (P23).

Let $\mathcal{R} = \{UWx : x \in \mathbb{R}^d\}$ be the range of $x \in \mathbb{R}^d \mapsto UWx \in \mathbb{R}^d$ and $r = \text{rank}(UW) \leq \min(p, d) = p$. By definition, \mathcal{R} is a r -dimensional linear subspace of \mathbb{R}^d . So, there exists an orthonormal basis $\{v_1, \dots, v_r\}$, $v_j \in \mathbb{R}^d$ for all $j \in [r]$, of \mathcal{R} . Now, since $\mathcal{R} \subseteq \mathbb{R}^d$, we can complete the basis of \mathcal{R} to form an orthonormal basis of \mathbb{R}^d , denoted $\{v_1, \dots, v_r, \dots, v_d\}$. Thus, let $V = [v_1 | \dots | v_p] \in \mathbb{R}^{d \times p}$ be the matrix whose columns are the first p basis vectors of \mathbb{R}^d , and $\bar{\mathcal{R}} = \{Vy : y \in \mathbb{R}^p\}$. Then, by definition of V (and since $p \geq r$), $\mathcal{R} \subseteq \bar{\mathcal{R}}$. In addition, since $\{v_1, \dots, v_p\}$ are orthonormal, $V^\top V = I_p$.

Let $x \in \mathbb{R}^d$. The orthogonal projection of x on $\bar{\mathcal{R}}$ is $V\hat{y}$, where $\hat{y} \in \mathbb{R}^p$ is a minimizer of $\psi: y \in \mathbb{R}^p \mapsto \|x - Vy\|_{\ell_2}^2$. By Fermat's rule, \hat{y} fulfills $\nabla \psi(\hat{y}) = 0$, that is $-2V^\top(x - V\hat{y}) = 0$, which leads to $\hat{y} = V^\top x$ (since $V^\top V = I_p$).

Therefore, since $UWx \in \tilde{\mathcal{R}}$, there exists $y_x \in \mathbb{R}^p$ such that $UWx = Vy_x$ and one has

$$\|x - UWx\|_{\ell_2}^2 = \|x - Vy_x\|_{\ell_2}^2 \geq \|x - V\hat{y}\|_{\ell_2}^2 = \|x - VV^\top x\|_{\ell_2}^2.$$

In other words, since $UWx \in \tilde{\mathcal{R}}$, its distance to x is greater than the distance between x and its orthogonal projection on $\tilde{\mathcal{R}}$ (sic). Then, by integration, we obtain

$$\mathbb{E} \left(\|X - UWx\|_{\ell_2}^2 \right) \geq \mathbb{E} \left(\|X - VV^\top X\|_{\ell_2}^2 \right),$$

which concludes the proof. \square

Owing to the preceding lemma, (P23) boils down to minimizing $\mathbb{E} \left(\|X - UU^\top X\|_{\ell_2}^2 \right)$ with respect to $U \in \mathbb{R}^{d \times p}$ such that $U^\top U = I_p$. In addition, by simple algebra, one has

$$\begin{aligned} \mathbb{E} \left(\|X - UU^\top X\|_{\ell_2}^2 \right) &= \mathbb{E} \left(\|X\|_{\ell_2}^2 + X^\top UU^\top UU^\top X - 2X^\top UU^\top X \right) \\ &= \mathbb{E} \left(\|X\|_{\ell_2}^2 - X^\top UU^\top X \right) && (U^\top U = I_p) \\ &= \mathbb{E} \left(\|X\|_{\ell_2}^2 \right) - \mathbb{E} \left(X^\top UU^\top X \right) \\ &= \mathbb{E} \left(\|X\|_{\ell_2}^2 \right) - \mathbb{E} \left(\text{tr} \left(X^\top UU^\top X \right) \right) && (X^\top UU^\top X \text{ is a scalar}) \\ &= \mathbb{E} \left(\|X\|_{\ell_2}^2 \right) - \mathbb{E} \left(\text{tr} \left(U^\top XX^\top U \right) \right) && (*) \\ &= \mathbb{E} \left(\|X\|_{\ell_2}^2 \right) - \text{tr} \left(U^\top \mathbb{E} \left(XX^\top \right) U \right), \end{aligned}$$

where we have used that the trace is invariant under cyclic permutations (*). Therefore, once again, (P23) boils down to

$$\underset{U \in \mathbb{R}^{d \times p} : U^\top U = I_p}{\text{maximize}} \quad \text{tr} \left(U^\top \mathbb{E} \left(XX^\top \right) U \right). \quad (\text{P24})$$

Theorem 59. Let $C \in \mathbb{R}^{d \times d}$ be a PSD matrix and let us denote $C = \sum_{i=1}^d \lambda_i v_i v_i^\top$ its eigendecomposition, where for all $i \in [d]$, $v_i \in \mathbb{R}^d$ and $\lambda_i \in \mathbb{R}_+$, with sorted eigenvalues $\lambda_1 \leq \dots \leq \lambda_d$. For any $k \in [d]$, let us denote $V_- = [v_1 | \dots | v_k] \in \mathbb{R}^{d \times k}$ and $V_+ = [v_{d-k+1} | \dots | v_d] \in \mathbb{R}^{d \times k}$, respectively the matrices of the minor and major eigenvectors.

Then,

$$\inf_{\substack{U \in \mathbb{R}^{d \times k} \\ U^\top U = I_k}} \text{tr}(U^\top C U) = \text{tr}(V_-^\top C V_-)$$

and

$$\sup_{\substack{U \in \mathbb{R}^{d \times k} \\ U^\top U = I_k}} \text{tr}(U^\top C U) = \text{tr}(V_+^\top C V_+).$$

Proof. We only prove the first equality. The proof of the second is similar. In addition, we will proceed in three steps:

1. showing that $\inf_{\substack{U \in \mathbb{R}^{d \times k} \\ U^\top U = I_k}} \text{tr}(U^\top C U) \geq \inf_{\substack{\beta \in [0,1]^d \\ \mathbf{1}^\top \beta = k}} \sum_{i=1}^d \lambda_i \beta_i$;
2. showing that $\inf_{\substack{\beta \in [0,1]^d \\ \mathbf{1}^\top \beta = k}} \sum_{i=1}^n \lambda_i \beta_i = \sum_{i=1}^k \lambda_i$;
3. showing that $\sum_{i=1}^k \lambda_i = \text{tr}(V_-^\top C V_-)$.

First step Let $\Lambda \in \mathbb{R}^{d \times d}$ be the diagonal matrix of eigenvalues and $V = [v_1 | \dots | v_d] \in \mathbb{R}^{d \times d}$ be the matrix of eigenvectors. Thus, $C = V \Lambda V^\top$.

For any matrix $U \in \mathbb{R}^{d \times k}$ such that $U^\top U = I_k$, let $B = V^\top U \in \mathbb{R}^{d \times k}$. Then

$$\text{tr}(U^\top C U) = \text{tr}(U^\top V \Lambda V^\top U) = \text{tr}(B^\top \Lambda B) = \sum_{j=1}^k b_j^\top \Lambda b_j = \sum_{j=1}^k \sum_{i=1}^d \lambda_i B_{ij}^2 = \sum_{i=1}^d \lambda_i \sum_{j=1}^k B_{ij}^2,$$

where for all $j \in [k]$, $b_j \in \mathbb{R}^d$ is the j^{th} column of B . Now, we have $B^\top B = U^\top V V^\top U = U^\top U = I_k$ since V is an orthogonal matrix and by definition of U . This means that the columns of B , denoted $\{b_1, \dots, b_k\}$ are orthonormal vectors from \mathbb{R}^d . On the one hand, if $k = d$, then $\{b_1, \dots, b_k\}$ is an orthonormal basis of \mathbb{R}^d and consequently B is an orthogonal matrix. In particular, $BB^\top = I_d$. Thus, for all $i \in [d]$, $\sum_{j=1}^k B_{ij}^2 = 1$.

On the other hand, if $k < d$, by completion, there exists a set of vectors $\{c_1, \dots, c_{d-k}\}$ such that $\{b_1, \dots, b_k, c_1, \dots, c_{d-k}\}$ is an orthonormal basis of \mathbb{R}^d . Thus, the matrix $A = [b_1, \dots, b_k, c_1, \dots, c_{d-k}] \in \mathbb{R}^{d \times d}$ is orthogonal and in particular $AA^\top = I_d$. Consequently, for all $i \in [d]$,

$$\sum_{j=1}^k (b_j)_i^2 + \sum_{j=1}^{d-k} (c_j)_i^2 = \sum_{j=1}^k B_{ij}^2 + \sum_{j=1}^{d-k} (c_j)_i^2 = 1.$$

Since $\sum_{j=1}^{d-k} (c_j)_i^2 \geq 0$, it comes $\sum_{j=1}^k B_{ij}^2 \leq 1$.

To sum up, in both cases, we have, for all $i \in [d]$, $\sum_{j=1}^k B_{ij}^2 \leq 1$. In addition, $\sum_{j=1}^k B_{ij}^2 \geq 0$ and $\sum_{i=1}^d \sum_{j=1}^k B_{ij}^2 = \text{tr}(B^\top B) = \text{tr}(I_k) = k$. Thus, by setting $\tilde{\beta}_i = \sum_{j=1}^k B_{ij}^2$, for all $i \in [d]$, we have $\tilde{\beta} \in [0, 1]^d$ and $\mathbf{1}^\top \tilde{\beta} = k$. So

$$\text{tr}(U^\top C U) = \sum_{i=1}^d \lambda_i \sum_{j=1}^k B_{ij}^2 = \sum_{i=1}^d \lambda_i \tilde{\beta}_i \geq \inf_{\substack{\beta \in [0,1]^d \\ \mathbf{1}^\top \beta = k}} \sum_{i=1}^d \lambda_i \beta_i.$$

Second step Let $\tilde{\beta} \in [0, 1]^d$ such that $\tilde{\beta}_i = 1$ for all $i \in [k]$ and 0 otherwise. First, $\mathbb{1}^\top \tilde{\beta} = k$, so $\tilde{\beta}$ is admissible. Second, for any $\beta \in [0, 1]^d$ such that $\mathbb{1}^\top \beta = k$, we have

$$\begin{aligned}
\sum_{i=1}^d \lambda_i \beta_i - \sum_{i=1}^d \lambda_i \tilde{\beta}_i &= \sum_{i=1}^k \lambda_i (\beta_i - 1) + \sum_{i=k+1}^d \lambda_i \beta_i \\
&\geq \sum_{i=1}^k \lambda_i (\beta_i - 1) + \lambda_{k+1} \sum_{i=k+1}^d \beta_i && (\lambda_{k+1} \leq \dots \leq \lambda_d) \\
&\geq \lambda_k \sum_{i=1}^k (\beta_i - 1) + \lambda_{k+1} \sum_{i=k+1}^d \beta_i && (\lambda_1 \leq \dots \leq \lambda_k \text{ and } (\beta_i - 1) \leq 0) \\
&\geq \lambda_{k+1} \sum_{i=1}^k (\beta_i - 1) + \lambda_{k+1} \sum_{i=k+1}^d \beta_i && (\lambda_k \leq \lambda_{k+1} \text{ and } \sum_{i=1}^k (\beta_i - 1) \leq 0) \\
&= \lambda_{k+1} \left(\sum_{i=1}^d \beta_i - k \right) \\
&= 0 && \left(\sum_{i=1}^d \beta_i = k \right).
\end{aligned}$$

So $\tilde{\beta}$ is a minimizer.

Third step It is easy to see that $V^\top V_- = J$ with $J \in \mathbb{R}^{d \times k}$ diagonal and $J_{ii} = 1$ for all $i \in [k]$. Therefore,

$$\text{tr}(V_-^\top C V_-) = \text{tr}(V_-^\top V \wedge V^\top V_-) = \text{tr}(J^\top \wedge J) = \sum_{j=1}^k \sum_{i=1}^d \lambda_i J_{ij}^2 = \sum_{i=1}^k \lambda_i.$$

□

By PSDness of $\mathbb{E}(XX^\top)$ and Theorem 59, it results that a solution to (P24) is $U = V_+$, where $V_+ \in \mathbb{R}^{d \times p}$ is the matrix of the major eigenvectors of $\mathbb{E}(XX^\top)$. Then, dimensionality reduction mapping is $x \in \mathbb{R}^d \mapsto V_+^\top x \in \mathbb{R}^p$. This approach is called PCA, due to its relation to eigendecomposition.

Remark 3.1.1. The proof of Theorem 59 tells us that

$$\text{tr}(V_+^\top \mathbb{E}(XX^\top) V_+) = \sum_{i=d-p+1}^d \lambda_i \leq \sum_{i=1}^d \lambda_i,$$

where $0 \leq \lambda_1 \leq \dots \leq \lambda_d$ are the sorted eigenvalues of $\mathbb{E}(XX^\top)$. Since the bound is attained for $p = d$ (that is, there is no dimensionality reduction), the maximal value $\text{tr}(V_+^\top \mathbb{E}(XX^\top) V_+)$ serves as an indicator of the “quality” of the approximation made when reducing the dimensionality of the data.

Let us remark that we can go back to (P23):

$$\begin{aligned}
\mathbb{E} \left(\|X - V_+ V_+^\top X\|_{\ell_2}^2 \right) &= \mathbb{E} \left(\|X\|_{\ell_2}^2 \right) - \text{tr} \left(V_+^\top \mathbb{E} (X X^\top) V_+ \right) \\
&= \mathbb{E} \left(\text{tr}(X X^\top) \right) - \text{tr} \left(V_+^\top \mathbb{E} (X X^\top) V_+ \right) \\
&= \text{tr} \left(\mathbb{E}(X X^\top) \right) - \text{tr} \left(V_+^\top \mathbb{E} (X X^\top) V_+ \right) \\
&= \sum_{i=1}^{d-p} \lambda_i.
\end{aligned}$$

In addition, for all $x \in \mathbb{R}^d$ and $j \in [p]$, $(V_+^\top x)_j = \sum_{i=1}^d (V_+)_{ij} x_i$. Therefore, for all $i \in [d]$, $(V_+)_{ij}$ describes the “influence” of the explicative variable x_i to the j^{th} component. Therefore, if $|(V_+)_{ij}|$ is large, it likely explains disparities of points in the j^{th} direction of the reduced space.

3.1.2 Link with variance maximization

Up to now, PCA has been defined as building linear functions for compression and reconstruction (in other words, as projecting a random vector on a linear subspace with minimal error) but there is no reason not to consider an affine reconstruction (that is projecting on an affine subspace) :

$$\underset{\substack{W \in \mathbb{R}^{p \times d}, U \in \mathbb{R}^{d \times p} \\ \mu \in \mathbb{R}^d}}{\text{minimize}} \quad \mathbb{E} \left(\|X - UW X - \mu\|_{\ell_2}^2 \right).$$

It comes trivially that the optimal μ for (W, U) fixed is: $\mu = \mathbb{E} X - UW \mathbb{E} X$, leading to the affine PCA problem:

$$\underset{W \in \mathbb{R}^{p \times d}, U \in \mathbb{R}^{d \times p}}{\text{minimize}} \quad \mathbb{E} \left(\|(X - \mathbb{E} X) - UW(X - \mathbb{E} X)\|_{\ell_2}^2 \right).$$

That is why, it is a common practice to center the data before applying PCA, namely considering the random variable $Z = X - \mathbb{E} X$ instead of X . Then, PCA aims at solving

$$\underset{U \in \mathbb{R}^{d \times p}: U^\top U = I_p}{\text{maximize}} \quad \text{tr} \left(U^\top \mathbb{E} (Z Z^\top) U \right) = \text{tr} \left(U^\top \mathbb{V}(X) U \right),$$

where $\mathbb{V}(X) = \mathbb{E} \left((X - \mathbb{E} X)(X - \mathbb{E} X)^\top \right)$ is the covariance matrix of X . It follows that the principal components (the column vectors of V_+) are the orthonormal vectors that “maximize the variance of X ”.

To be more formal, let us describe what we mean by “maximizing” the variance of X . For this purpose, we define recursively the sequence (u_1, \dots, u_p) . Let $u_1 \in \mathbb{R}^d$ be the normalized vector (direction) that maximizes the unidirectional variance of X , namely a solution to:

$$\begin{aligned}
&\underset{u \in \mathbb{R}^d}{\text{maximize}} \quad \mathbb{V}(u^\top X) \\
&\text{s. t.} \quad \|u\|_{\ell_2} = 1.
\end{aligned} \tag{P25}$$

Then, for all $k \in [p - 1]$, given (u_1, \dots, u_k) , let u_{k+1} be solution to

$$\begin{aligned}
&\underset{u \in \mathbb{R}^d}{\text{maximize}} \quad \mathbb{V}(u^\top X) \\
&\text{s. t.} \quad \begin{cases} \|u\|_{\ell_2} = 1 \\ \forall j \in [k]: u^\top u_j = 0. \end{cases}
\end{aligned} \tag{P26}$$

Now, we will show by induction that (u_p, \dots, u_1) corresponds to the principal components, namely the column vectors of V_+ , which are also the major eigenvectors of $\mathbb{V}(X)$.

First, since $\mathbb{V}(u^\top X) = u^\top \mathbb{V}(X)u = \text{tr}(u^\top \mathbb{V}(X)u)$, by Theorem 59, the major eigenvector of $\mathbb{V}(X)$ is solution to (P25). So we can set $u_1 = v_d$ (with the preceding notation).

Second, for any $k \in [p-1]$, let us assume that $(u_k, \dots, u_1) = (v_{d-k+1}, \dots, v_d)$. Since (u_1, \dots, u_k) are fixed, maximizing $\mathbb{V}(u^\top X) = u^\top \mathbb{V}(X)u$ is the same as maximizing $u^\top \mathbb{V}(X)u + \sum_{j=1}^k u_j^\top \mathbb{V}(X)u_j$. Therefore, (P26) is similar to

$$\begin{aligned} & \underset{u \in \mathbb{R}^d, U \in \mathbb{R}^{d \times (k+1)}}{\text{maximize}} && \sum_{j=1}^k u_j^\top \mathbb{V}(X)u_j + u^\top \mathbb{V}(X)u = \text{tr}(U^\top \mathbb{V}(X)U) \\ & \text{s. t.} && \begin{cases} U = [u|u_k| \dots |u_1] = [u|v_{d-k+1}| \dots |v_d] \\ U^\top U = I_{k+1}. \end{cases} \end{aligned} \quad (\text{P27})$$

Let us remark that (P27) has a solution since (v_{d-k+1}, \dots, v_d) are orthonormal. Now, let us consider $V_+ = [v_{d-k}|v_{d-k+1}| \dots |v_d]$. Then, by Theorem 59, $\text{tr}(V_+^\top \mathbb{V}(X)V_+) \geq \text{tr}(U^\top \mathbb{V}(X)U)$ for all $U \in \mathbb{R}^{d \times (k+1)}$ such that $U^\top U = I_{k+1}$. In particular, this is also true for all U that fulfill the constraint of (P27) (said *admissible*). Since V_+ is also admissible, this proves that V_+ , along with its first column $u = v_{d-k}$, is a maximizer of (P26).

Remark 3.1.2. As explained previously, $\sum_{i=n-p+1}^n \lambda_i$ measures the quality of the approximation made by PCA with p components. In view of the variance maximization paradigm, $r = \frac{\sum_{i=n-p+1}^n \lambda_i}{\sum_{i=1}^n \lambda_i}$ is often called the “ratio of explained variance” and is a normalized indicator of the quality of approximation.

3.1.3 Link with the Gram matrix

In practice, we are provided with a sample $\{X_1, \dots, X_n\} \subseteq \mathbb{R}^d$. Then, considering the empirical twin of $\mathbb{V}(X)$, which is $\frac{1}{n} \sum_{i=1}^n X_i X_i^\top$, PCA boils down to finding the p major eigenvectors of the empirical (and scaled by n) covariance matrix $C = \sum_{i=1}^n X_i X_i^\top = \mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{d \times d}$, where $\mathbf{X} \in \mathbb{R}^{n \times d}$ is the matrix whose rows are the observations X_i ($i \in [n]$).

As we are interested in dimensionality reduction, it is quite licit to assume the dimension d to be very big and that $d \geq n$. Therefore, PCA has to diagonalize a (big) $d \times d$ matrix, while saving only the p major eigenvectors. The forthcoming property tells us that, as long as the reduced dimension $p \leq n$, PCA can be implemented in a cheaper manner by diagonalizing the (small) Gram matrix $K = (X_i^\top X_j)_{1 \leq i, j \leq n} = \mathbf{X} \mathbf{X}^\top \in \mathbb{R}^{n \times n}$.

Property 60. Let us assume that $d \geq n$. Then if $\lambda_1 \leq \dots \leq \lambda_n$ are the eigenvalues of K with eigenvectors (v_1, \dots, v_n) , then there exists $(u_1, \dots, u_{d-n}) \in (\mathbb{R}^d)^{d-n}$ such that $0 \leq \dots \leq 0 \leq \lambda_1 \leq \dots \leq \lambda_n$ are eigenvalues of C with eigenvectors $(u_1, \dots, u_{d-n}, \mathbf{X}^\top v_1, \dots, \mathbf{X}^\top v_n)$.

Proof. For all $i \in [n]$, one has

$$CX^\top v_i = X^\top XX^\top v_i = X^\top K v_i = \lambda_i X^\top v_i,$$

which proves that λ_i is an eigenvalue of C with eigenvector $X^\top v_i$. In addition, let $k = \text{rank}(X) \leq n$. Then, $\text{rank}(C) = \text{rank}(X^\top X) = \text{rank}(XX^\top) = \text{rank}(K) = k$, so C and K have exactly k non-zero eigenvalues. As we have already found n eigenvalues of C , with k of which are non-zero, this means that all other eigenvalues of C are 0. This concludes the proof. \square

Then, in order to build the dimensionality reduction mapping $x \in \mathbb{R}^d \mapsto V_+^\top x \in \mathbb{R}^p$, we have to normalize the major eigenvectors (v_{n-p+1}, \dots, v_n) of K to unit vectors:

$$V_+ = \left[\frac{1}{\|X^\top v_{n-p+1}\|_{\ell_2}} X^\top v_{n-p+1} \mid \dots \mid \frac{1}{\|X^\top v_n\|_{\ell_2}} X^\top v_n \right] \in \mathbb{R}^{d \times p}.$$

Remark 3.1.3. Then for all $i \in [p]$, we have

$$\|X^\top v_{n-p+i}\|_{\ell_2}^2 = v_{n-p+i}^\top K v_{n-p+i} = \lambda_{n-p+i} \|v_{n-p+i}\|_{\ell_2}^2 = \lambda_{n-p+i}.$$

In addition, the matrix of reduced representations is

$$\begin{aligned} (V_+^\top X^\top)^\top &= X V_+ \\ &= \left[\frac{K v_{n-p+1}}{\|X^\top v_{n-p+1}\|_{\ell_2}} \mid \dots \mid \frac{K v_n}{\|X^\top v_n\|_{\ell_2}} \right] \\ &= \left[\frac{\lambda_{n-p+1} v_{n-p+1}}{\sqrt{\lambda_{n-p+1}}} \mid \dots \mid \frac{\lambda_n v_n}{\sqrt{\lambda_n}} \right] \\ &= \left[\sqrt{\lambda_{n-p+1}} v_{n-p+1} \mid \dots \mid \sqrt{\lambda_n} v_n \right]. \end{aligned}$$

Remark 3.1.4 (PCA and spectral clustering). Let us go back to spectral clustering: the Gram matrix K can legitimately be viewed as an adjacency matrix. Moreover, if all points have approximately the same degree, that is $D \approx \gamma I_n$ with $\gamma \approx \sum_{j=1}^n K_{1,j} > 0$, and γ is bigger than the leading eigenvalue of K , then the Laplacian $L = D - K$ is PSD and the minor eigenvectors of L correspond to the major eigenvectors of K . As a result, PCA and spectral clustering boils down to perform the same dimensionality reduction, while their purposes are completely different.

3.1.4 Link with singular values

The next theorem gives a more general solution to PCA.

Theorem 61 (singular value decomposition (SVD) [Shalev-Shwartz and Ben-David, 2014, Appendix C.4]). Let m and n be two positive integers and $A \in \mathbb{R}^{m \times n}$. Let $r = \text{rank}(A)$ be the rank of A , then there exist $U \in \mathbb{R}^{m \times r}$, $D \in \mathbb{R}^{r \times r}$ and $V \in \mathbb{R}^{n \times r}$ such that

$$A = UDV^\top,$$

and

- ◇ the columns of U are orthonormal: $U^\top U = I_r$;
- ◇ D is diagonal with positive and uniquely defined entries (called singular values);
- ◇ the columns of V are orthonormal: $V^\top V = I_r$.

Furthermore, denoting $U = [u_1 | \dots | u_r]$ and $V = [v_1 | \dots | v_r]$, one has:

- ◇ $A = \sum_{i=1}^r D_{ii} u_i v_i^\top$;
- ◇ for all $i \in [r]$, u_i and v_i are left and right singular vectors: $Av_i = D_{ii} u_i$ and $A^\top u_i = D_{ii} v_i$;
- ◇ for all $i \in [r]$, u_i is an eigenvector of AA^\top with eigenvalue D_{ii}^2 ;
- ◇ for all $i \in [r]$, v_i is an eigenvector of $A^\top A$ with eigenvalue D_{ii}^2 .

In practice, SVD is a tool from linear algebra, that is more robust than eigendecomposition for solving PCA. It computes neither $\mathbf{X}^\top \mathbf{X}$ nor $\mathbf{X}\mathbf{X}^\top$. In addition, PCA is performed even faster by computing the top p singular values, along with the left and right singular vectors of \mathbf{X} thanks to a truncated SVD. Let us remark that for computing the dimensionality reduction mapping, we need the eigenvectors of $C = \mathbf{X}^\top \mathbf{X}$, i.e. the right singular vectors, and for the matrix of reduced representations, the eigenvectors of $K = \mathbf{X}\mathbf{X}^\top$, i.e. the left singular vectors.

Algorithm 14 Reduced representation by PCA.

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$ (data matrix), p (reduced dimension).

Second order matrix

$C \leftarrow \mathbf{X}^\top \mathbf{X}$

$V \leftarrow p$ major eigenvectors of C

$U \leftarrow \mathbf{X}V$

Gram matrix

$K \leftarrow \mathbf{X}\mathbf{X}^\top$

$\lambda_1, \dots, \lambda_p \leftarrow p$ major eigenvalues

$V \leftarrow p$ major eigenvectors of K

$U \leftarrow [\sqrt{\lambda_1} v_1 | \dots | \sqrt{\lambda_p} v_p]$

SVD

$\lambda_1, \dots, \lambda_p \leftarrow p$ major singular values of \mathbf{X}

$V \leftarrow p$ major left singular vectors of \mathbf{X}

$U \leftarrow [\lambda_1 v_1 | \dots | \lambda_p v_p]$

Output: $U \in \mathbb{R}^{n \times p}$.

3.1.5 Random projection

Although PCA is very appealing by its simplicity and probabilistic interpretation, it requires computing singular vectors, which can be very expensive for very high-dimensional data, or very big samples. That is why, we describe in this section a very cheap way of reducing dimension. Here, the criterion of interest is not a reasonable recovery of the original data but saving the original placement of the data points $\{X_1, \dots, X_n\}$ with respect to each other. In other words, we would like to preserve pairwise distances.

Roughly speaking, Theorem 63 states that if the reduced dimension p is proportional to $\log(n)/\epsilon^2$, then a random matrix $W \in \mathbb{R}^{p \times d}$ produces a dimensionality reduction mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$, that preserves pairwise distances up to an error ϵ .

Lemma 62 (Concentration of a chi-squared variable). *Let Z be a chi-squared random variable with p degrees of freedom, denoted by $Z \sim \chi_p^2$, and $\epsilon \in (0, 1)$. Then, one has $\mathbb{E}(Z/p) = 1$ and*

$$\mathbb{P}(|Z/p - 1| > \epsilon) \leq 2e^{-p\epsilon^2/8}.$$

The proof is a good exercise.

Theorem 63 (Johnson–Lindenstrauss Lemma). *Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a finite set of vectors with cardinality $n \geq 2$ and $W \in \mathbb{R}^{p \times d}$ be a random matrix such that its entries $\{W_{ij}\}_{\substack{1 \leq i \leq p \\ 1 \leq j \leq d}}$ are iid and distributed according to $\mathcal{N}\left(0, \frac{1}{p}\right)$. For any $(\epsilon, \delta) \in (0, 1)^2$, if*

$$p \geq 16\epsilon^{-2} \log(n/\sqrt{\delta}),$$

then with probability at least $1 - \delta$ on the random matrix W ,

$$\forall (x, x') \in \mathcal{S}: \quad (1 - \epsilon) \|x - x'\|_{\ell_2}^2 \leq \|Wx - Wx'\|_{\ell_2}^2 \leq (1 + \epsilon) \|x - x'\|_{\ell_2}^2.$$

The mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ is called an ϵ -isometry on \mathcal{S} .

Proof. For any $y \in \mathbb{R}^d$ such that $y \neq 0$, let $Z = Wy \in \mathbb{R}^p$. Then, $\|Z\|_{\ell_2}^2 = \|Wy\|_{\ell_2}^2 = \sum_{i=1}^p \left(\sum_{j=1}^d W_{ij} y_j \right)^2$. But for all $i \in [p]$, the random variable $Z_i = \sum_{j=1}^d W_{ij} y_j$ is a weighted sum of independent normal random variables, so it is normally distributed with $\mathbb{E}(Z_i) = \sum_{j=1}^d y_j \mathbb{E}(W_{ij}) = 0$ and $\mathbb{V}(Z_i) = \sum_{j=1}^d y_j^2 \mathbb{V}(W_{ij}) = \frac{\|y\|_{\ell_2}^2}{p}$ (by independence). In other words, $\sqrt{p} \frac{Z_i}{\|y\|_{\ell_2}} \sim \mathcal{N}(0, 1)$. Therefore, since $\{Z_1, \dots, Z_p\}$ are independent, $\frac{p\|Z\|_{\ell_2}^2}{\|y\|_{\ell_2}^2} = \sum_{i=1}^p \left(\sqrt{p} \frac{Z_i}{\|y\|_{\ell_2}} \right)^2 \sim \chi_p^2$. By the preceding lemma, for all $\epsilon \in (0, 1)$:

$$\mathbb{P} \left(\left| \frac{\|Wy\|_{\ell_2}^2}{\|y\|_{\ell_2}^2} - 1 \right| > \epsilon \right) = \mathbb{P} \left(\|Wy\|_{\ell_2}^2 \notin \left[(1 - \epsilon) \|y\|_{\ell_2}^2, (1 + \epsilon) \|y\|_{\ell_2}^2 \right] \right) \leq 2e^{-p\epsilon^2/8}.$$

As a consequence, for all $x \in \mathcal{S}$ and $x' \in \mathcal{S}$, let $y = x - x'$. On the one hand, if $y = 0$, then $\|Wy\|_{\ell_2}^2 = 0 \notin \left[(1 - \epsilon)\|y\|_{\ell_2}^2, (1 + \epsilon)\|y\|_{\ell_2}^2\right] = \{0\}$ is true with probability 0. On the other hand, $\|Wy\|_{\ell_2}^2 \notin \left[(1 - \epsilon)\|y\|_{\ell_2}^2, (1 + \epsilon)\|y\|_{\ell_2}^2\right]$ holds with probability less than $2e^{-p\epsilon^2/8}$. Therefore, by a union bound on the at most $\binom{n}{2} = \frac{n(n-1)}{2}$ distinct pairs of vectors and n identical pairs,

$$\begin{aligned} \mathbb{P}\left(\exists (x, x') \in \mathcal{S}^2 : \|Wx - Wx'\|_{\ell_2}^2 \notin \left[(1 - \epsilon)\|x - x'\|_{\ell_2}^2, (1 + \epsilon)\|x - x'\|_{\ell_2}^2\right]\right) &\leq \binom{n}{2} 2e^{-p\epsilon^2/8} + n \times 0 \\ &= n(n-1)e^{-p\epsilon^2/8} \\ &\leq n^2 e^{-p\epsilon^2/8}. \end{aligned}$$

Thus, for any $\delta \in (0, 1)$, as long as $p \geq 8\epsilon^{-2} \log(n^2/\delta)$, we have $n^2 e^{-p\epsilon^2/8} \leq \delta$, which leads to

$$\mathbb{P}\left(\forall (x, x') \in \mathcal{S}^2 : \|Wx - Wx'\|_{\ell_2}^2 \in \left[(1 - \epsilon)\|x - x'\|_{\ell_2}^2, (1 + \epsilon)\|x - x'\|_{\ell_2}^2\right]\right) \geq 1 - \delta.$$

□

Remark 3.1.5. *The underlying idea of this approach is that the reduction mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ is an exact isometry “in expectation”:*

$$\forall x \in \mathbb{R}^d : \quad \mathbb{E}\left(\|Wx\|_{\ell_2}^2\right) = \|x\|_{\ell_2}^2.$$

Indeed, since for all $x \in \mathbb{R}^d$ such that $x \neq 0$, $\frac{p\|Wx\|_{\ell_2}^2}{\|x\|_{\ell_2}^2} \sim \chi_p^2$, one has

$$\mathbb{E}\left(\|Wx\|_{\ell_2}^2\right) = \frac{\|x\|_{\ell_2}^2}{p} \mathbb{E}\left(\frac{p\|Wx\|_{\ell_2}^2}{\|x\|_{\ell_2}^2}\right) = \frac{\|x\|_{\ell_2}^2}{p} p = \|x\|_{\ell_2}^2.$$

Let us remark that it is enough for $\{W_{ij}\}_{1 \leq i \leq p, 1 \leq j \leq d}$ to be independent with $\mathbb{E} W_{ij} = 0$ and $\mathbb{V}(W_{ij}) = \frac{1}{p}$ (for all $i \in [p]$ and $j \in [d]$) in order to get an exact isometry “in expectation”. Indeed, denoting $Z = Wx$, we have for all $i \in [p]$, $\mathbb{E} Z_i = 0$ and $\mathbb{V}(Z_i) = \sum_{j=1}^d x_j^2 \mathbb{V}(W_{ij}) = \frac{\|x\|_{\ell_2}^2}{p}$. Then, it comes:

$$\mathbb{E}\left[\|Wx\|_{\ell_2}^2\right] = \mathbb{E}\left[\|Z\|_{\ell_2}^2\right] = \sum_{i=1}^p \mathbb{E}[Z_i^2] = \sum_{i=1}^p \mathbb{E}[\mathbb{V}(Z_i) + (\mathbb{E} Z_i)^2] = \|x\|_{\ell_2}^2.$$

Remark 3.1.6. *It is remarkable that the requirement on the reduced dimension $p \geq 16\epsilon^{-2} \log(n/\delta)$ does not depend on the original dimension d . This means that we could consider data in infinite-dimensional Hilbert space.*

The forthcoming corollary goes a step further Theorem 63, telling us that we can find very quickly a linear dimensionality reduction mapping, that is an exact ϵ -isometry on a given dataset. This is done by sampling a matrix and checking that it provides the expected result.

Corollary 64. *Let $\mathcal{S} \subseteq \mathbb{R}^d$ be a finite set of vectors with cardinality $n \geq 2$. For any $\epsilon \in (0, 1)$, let p be an integer such that*

$$p \geq 16\epsilon^{-2} \log(n),$$

then there exists a matrix $W \in \mathbb{R}^{p \times d}$ such that

$$\forall (x, x') \in \mathcal{S}^2: \quad (1 - \epsilon) \|x - x'\|_{\ell_2}^2 \leq \|Wx - Wx'\|_{\ell_2}^2 \leq (1 + \epsilon) \|x - x'\|_{\ell_2}^2.$$

In addition, such a matrix can be found by a randomized algorithm, for which the expected time is linear in n .

Proof. The proof of Theorem 63 tells us that, for a random matrix $W \in \mathbb{R}^{p \times d}$ such that its entries $\{W_{ij}\}_{1 \leq i \leq p, 1 \leq j \leq d}$ are iid and distributed according to $\mathcal{N}\left(0, \frac{1}{p}\right)$,

$$\mathbb{P}\left(\exists (x, x') \in \mathcal{S}^2: \|Wx - Wx'\|_{\ell_2}^2 \notin \left[(1 - \epsilon) \|x - x'\|_{\ell_2}^2, (1 + \epsilon) \|x - x'\|_{\ell_2}^2\right]\right) \leq n(n - 1)e^{-p\epsilon^2/8}.$$

Yet, if $p \geq 8\epsilon^{-2} \log(n^2)$, then $e^{-p\epsilon^2/8} \leq 1/n^2$, so the preceding probability is bounded by $1 - 1/n$. Consequently,

$$\mathbb{P}\left(\forall (x, x') \in \mathcal{S}^2: \|Wx - Wx'\|_{\ell_2}^2 \in \left[(1 - \epsilon) \|x - x'\|_{\ell_2}^2, (1 + \epsilon) \|x - x'\|_{\ell_2}^2\right]\right) \geq \frac{1}{n} > 0.$$

This proves the existence of a matrix W satisfying the quasi-isometry condition.

Besides, the number of random matrices to draw in order to find a suitable one can be modeled by a random variable Z , distributed according to a geometric distribution with probability of success $p \geq 1/n$. Thus, the expected number of random matrices to draw in order to find a suitable one is $\mathbb{E}Z = 1/p \leq n$. This proves the linear time needed (in expectation), to find a suitable matrix W . \square

Figure 3.1 illustrates Corollary 64 regarding the minimal dimension required to get an ϵ -isometry. We observe that this is quite huge. Thus, random projection is a workable method only for very high-dimensional data. Otherwise, PCA should be preferred.

Besides, Figure 3.2 depicts the ratio $\frac{Wx_i - Wx_j}{x_i - x_j}^2$ for two trials of random matrices W and a dataset of 50 points uniformly sample on the hypercube $[0, 1]^{10000}$. It appears that, in practice, a single trial is sufficient to get a suitable matrix W (which is dramatically faster than linear in n) and even if the first trial does not work, the ϵ -isometry requirement is violated only for few points. A possible explanation is that the lower bound of the probability of success of finding a suitable matrix W ($1/n$) is very loose because of the union bound used in the proof and of the lack of hypothesis on the data distribution.

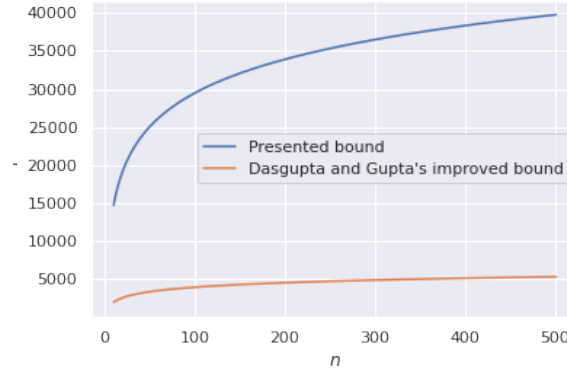


Figure 3.1: Curves $p = 16\epsilon^{-2} \log(n)$ (presented bound) and $p = \frac{4}{\epsilon^2/2 - \epsilon^2/3} \log(n)$ (Dasgupta and Gupta's improved bound) for $\epsilon = 0.05$.

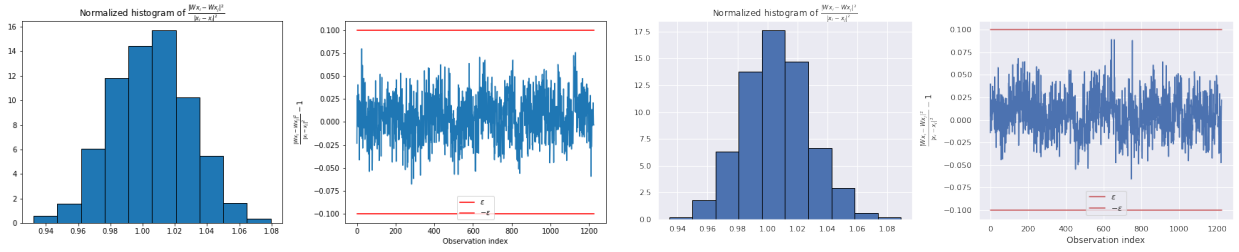


Figure 3.2: Analysis of the requirement $\left| \frac{Wx_i - Wx_j}{x_i - x_j} - 1 \right| \leq \epsilon$, which is fulfilled for the left trial and violated for the right one.

3.1.6 Reconstruction of random projections

When we moved from PCA to random projections, we changed the paradigm of dimensionality reduction from a *reasonable recovery* to *preserving pairwise distances*. It is entirely licit to wonder if one has a reasonable recovery of the original data when pairwise distances are preserved up to an error ϵ .

An answer comes from the mathematical domain of *compressed sensing* (see Claire Boyer's class), which requires nevertheless to modify our assumptions: from now on, we do not consider being provided with a finite set of points \mathcal{S} any longer, but, given an integer s , we focus on all s -sparse vectors. A vector $x \in \mathbb{R}^d$ is said s -sparse if its pseudo ℓ_0 -norm is bounded by s :

$$\|x\|_{\ell_0} = \sum_{i=1}^d \mathbf{1}_{x_i \neq 0} \leq s.$$

Theorem 65 ([Shalev-Shwartz and Ben-David, 2014, Theorem 23.9]). Let $W \in \mathbb{R}^{p \times d}$ be a random matrix such that its entries $\{W_{ij}\}_{\substack{1 \leq i \leq p \\ 1 \leq j \leq d}}$ are iid and distributed according to $\mathcal{N}\left(0, \frac{1}{p}\right)$, and $s \in [d]$

a sparsity level. For any $(\epsilon, \delta) \in (0, 1)^2$, if

$$p \geq 100s\epsilon^{-2} \log(40d/(\delta\epsilon)),$$

then with probability at least $1 - \delta$ on the random matrix W ,

$$\forall x \in \mathbb{R}^d : \|x\|_{\ell_0} \leq s, \quad (1 - \epsilon) \|x\|_{\ell_2}^2 \leq \|Wx\|_{\ell_2}^2 \leq (1 + \epsilon) \|x\|_{\ell_2}^2.$$

The matrix W is said to have the (ϵ, s) -restricted isometry property (RIP).

Theorem 65 gives a condition on the reduced dimension p for the dimensionality reduction mapping $x \in \mathbb{R}^d \mapsto Wx \in \mathbb{R}^p$ to be ϵ -isometry on the set of all s -sparse vectors “in expectation”. It should be noticed that, contrarily to the Johnson-Lindenstrauss Lemma (Theorem 63), this condition on p depends on the dimension d of the original data. Figure 3.3 compares both requirements.

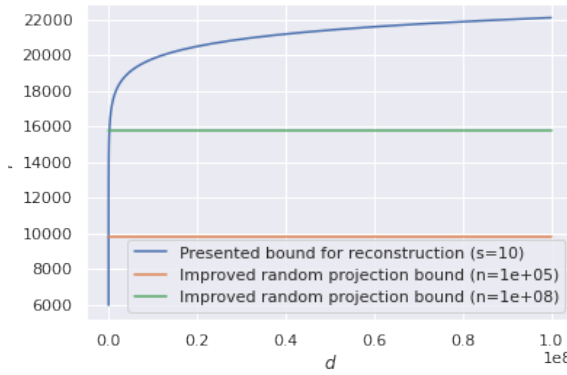


Figure 3.3: Curves $p = 100s\epsilon^{-2} \log(40d/(\delta\epsilon))$ (presented bound for reconstruction) and $p = \frac{4}{\epsilon^2(2-\epsilon^2/3)} \log(n)$ (Improved random projection bound) for $\epsilon = 0.1$.

Now, Theorem 66 states that we can recover the original data $x \in \mathbb{R}^d$ from its Gaussian random projection $Wx \in \mathbb{R}^p$ by solving a convex optimization problem. This topic is studied in the minutest detail in Claire Boyer’s class.

Theorem 66 ([Shalev-Shwartz and Ben-David, 2014, Theorem 23.7]). Let $\epsilon \in (0, 2/5)$, $s \in [d]$ be a sparsity level and $W \in \mathbb{R}^{p \times d}$ be an $(\epsilon, 2s)$ -RIP matrix. Then,

$$\forall x \in \mathbb{R}^d : \|x\|_{\ell_0} \leq s, \quad x \in \arg \min_{\substack{u \in \mathbb{R}^d: \\ Wu=Wx}} \|u\|_{\ell_1},$$

where $\|u\|_{\ell_1} = \sum_{i=1}^d |u_i|$.

Remark 3.1.7. It is quite natural to wonder which of PCA or random projection is preferable. To answer this question, one can focus on the recovery property of each method.

On the one hand, PCA guarantees perfect recovery whenever the variable X lies in a linear subspace of \mathbb{R}^d , with dimension k less than the number p of selected components. Indeed, if \mathcal{R} is the subspace in which lies X and \mathcal{R}_\perp is its orthogonal subspace, the k directions $u \in \mathbb{R}^d$ ($\|u\|_{\ell_2} = 1$) that maximize the variance of $u^\top X$ are necessarily in \mathcal{R} ($u^\top X = 0$ as soon as $u \in \mathcal{R}_\perp$). Thus, if we are interested in $p \geq k$ components, then the p directions that maximize the variance contain an orthonormal basis of \mathcal{R} , which guarantees perfect recovery of X from its projection.

On the other hand, Gaussian random projection guarantees perfect recovery whenever the original data is sparse (in a well chosen basis).

3.2 Nonlinear methods

3.2.1 Kernel principal component analysis

With the kernel trick

Let $\{X_i\}_{1 \leq i \leq n} \subset \mathbb{R}^d$ be iid copies of X and $k: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ a kernel with feature map $\phi: \mathbb{R}^d \rightarrow \mathcal{G}$, where \mathcal{G} is an appropriate Hilbert space (of dimension D , potentially infinite). As a reminder, we have $\forall (x, x') \in \mathbb{R}^d \times \mathbb{R}^d: k(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$.

Similarly to kernel Fisher discriminant analysis (Section 1.1.4), we consider the general method of applying PCA to the random variable $Z = \phi(X) - \mathbb{E}(\phi(X)) \in \mathcal{G}$. As shown in Section 3.1.1, this boils down to diagonalizing $\mathbb{E}(ZZ^\top)$, which may be an infinitely dimensional matrix (as soon as \mathcal{G} is of dimension ∞), that is a linear operator. From now on, we may use as a notation for all $x \in \mathbb{R}^d$ and $x' \in \mathbb{R}^d$, $\phi(x)^\top \phi(x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{G}}$.

Even though it seems quite difficult, we rely on Section 3.1.3, in which we have shown that, in its empirical version, PCA can be performed by diagonalizing the Gram matrix $K_Z = \left(\langle Z_i, Z_j \rangle_{\mathcal{G}} \right)_{1 \leq i, j \leq n}$ of the sample $\{Z_i\}_{1 \leq i \leq n}$, where for each $i \in [n]$, $Z_i = \phi(X_i) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell)$.

Then, we can write the sample matrices

$$\mathbf{X} = [\phi(X_1) | \dots | \phi(X_n)]^\top \in \mathbb{R}^{n \times D}, \quad \mathbf{Z} = [Z_1 | \dots | Z_n]^\top \in \mathbb{R}^{n \times D},$$

whose rows are the sample vectors. Since $\frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) = \mathbf{X}^\top \mathbf{1}/n$, it is easy to show that

$$\mathbf{Z} = \mathbf{X} - \left[\frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) | \dots | \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right]^\top = \mathbf{X} - \mathbf{1} \left(\frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right)^\top = (\mathbf{I}_n - M)\mathbf{X},$$

where $M = \mathbf{1}\mathbf{1}^\top/n \in \mathbb{R}^{n \times n}$. Therefore

$$K_Z = \mathbf{Z}\mathbf{Z}^\top = (\mathbf{I}_n - M)K_X(\mathbf{I}_n - M),$$

where $K_X = \left(\langle \phi(X_i), \phi(X_j) \rangle_{\mathcal{G}} \right)_{1 \leq i, j \leq n}$.

Remark 3.2.1. More formally, one has, for all $(i, j) \in [n]^2$:

$$\begin{aligned}
(K_Z)_{ij} &= \langle Z_i, Z_j \rangle_{\mathcal{G}} \\
&= \langle \phi(X_i), \phi(X_j) \rangle_{\mathcal{G}} + \left\| \frac{1}{n} \sum_{\ell=1}^n \phi(X_{\ell}) \right\|_{\mathcal{G}}^2 - \left\langle \phi(X_i), \frac{1}{n} \sum_{\ell=1}^n \phi(X_{\ell}) \right\rangle_{\mathcal{G}} - \left\langle \phi(X_j), \frac{1}{n} \sum_{\ell=1}^n \phi(X_{\ell}) \right\rangle_{\mathcal{G}} \\
&= \langle \phi(X_i), \phi(X_j) \rangle_{\mathcal{G}} + \frac{1}{n^2} \sum_{1 \leq \ell, h \leq n} \langle \phi(X_{\ell}), \phi(X_h) \rangle_{\mathcal{G}} - \frac{1}{n} \sum_{\ell=1}^n \langle \phi(X_i), \phi(X_{\ell}) \rangle_{\mathcal{G}} - \frac{1}{n} \sum_{\ell=1}^n \langle \phi(X_j), \phi(X_{\ell}) \rangle_{\mathcal{G}} \\
&= (K_X)_{ij} + (MK_X M)_{ij} - (KM)_{ij} - (MK)_{ij} \\
&= ((I_n - M)K_X(I_n - M))_{ij}.
\end{aligned}$$

Remark 3.2.2. If the data is not centered, that is PCA is applied on the sample $\{\phi(X_i)\}_{1 \leq i \leq n}$, then the matrix to diagonalize is K_X instead of K_Z .

Let now $(v_1, \dots, v_p) \subset \mathbb{R}^n$ be the major unit eigenvectors of K_Z . The dimensionality reduction mapping is $x \in \mathbb{R}^D \mapsto V_+^T x \in \mathbb{R}^p$, where

$$V_+ = \left[\frac{1}{\|Z^T v_1\|_{\ell_2}} Z^T v_1 \mid \dots \mid \frac{1}{\|Z^T v_p\|_{\ell_2}} Z^T v_p \right] \in \mathbb{R}^{D \times p}.$$

We have, for all $i \in [p]$,

$$\|Z^T v_i\|_{\mathcal{G}}^2 = v_i^T K_Z v_i = \lambda_i \|v_i\|_{\ell_2}^2 = \lambda_i.$$

In addition, let $\mathbf{U} \in \mathbb{R}^{n \times p}$ be the matrix of reduced representations (that is $(V_+^T [\phi(X_i) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_{\ell})])_{1 \leq i \leq n}$ are the rows of \mathbf{U}) is

$$\mathbf{U} = (V_+^T Z^T)^T = Z V_+ = \left[\frac{K_Z v_1}{\sqrt{\lambda_1}} \mid \dots \mid \frac{K_Z v_p}{\sqrt{\lambda_p}} \right] = \left[\sqrt{\lambda_1} v_1 \mid \dots \mid \sqrt{\lambda_p} v_p \right]. \quad (3.1)$$

Moreover, given a new point $x \in \mathbb{R}^d$, its reduced representation $u \in \mathbb{R}^p$ is $u = V_+^T (\phi(x) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_{\ell}))$,

with, for all $j \in [p]$:

$$\begin{aligned}
u_j &= \lambda_j^{-1/2} v_j^\top \mathbf{Z} \left(\phi(x) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right) \\
&= \lambda_j^{-1/2} \sum_{i=1}^n (v_j)_i \left(\phi(X_i) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right)^\top \left(\phi(x) - \frac{1}{n} \sum_{\ell=1}^n \phi(X_\ell) \right) \\
&= \lambda_j^{-1/2} \sum_{i=1}^n (v_j)_i \left(k(x, X_i) + \frac{1}{n^2} \sum_{1 \leq \ell, \ell' \leq n} k(X_\ell, X_{\ell'}) - \frac{1}{n} \sum_{\ell=1}^n (k(x, X_\ell) + k(X_i, X_\ell)) \right) \\
&= \lambda_j^{-1/2} \left[\sum_{i=1}^n (v_j)_i \left(k(x, X_i) - \frac{1}{n} \sum_{\ell=1}^n k(X_\ell, X_i) \right) + \frac{\mathbf{1}^\top v_j}{n} \sum_{i=1}^n \left(\frac{1}{n} \sum_{\ell=1}^n k(X_i, X_\ell) - k(x, X_i) \right) \right] \\
&= \lambda_j^{-1/2} \left[\sum_{i=1}^n \left((v_j)_i - \frac{\mathbf{1}^\top v_j}{n} \right) k(x, X_i) + \frac{1}{n} \sum_{1 \leq i, \ell \leq n} \left(\frac{\mathbf{1}^\top v_j}{n} - (v_j)_i \right) k(X_i, X_\ell) \right] \\
&= \sum_{i=1}^n (\alpha_j)_i k(x, X_i) - \frac{1}{n} \sum_{1 \leq i, \ell \leq n} (\alpha_j)_i k(X_i, X_\ell),
\end{aligned}$$

where $\alpha_j = \lambda_j^{-1/2} \left(v_j - \left(\frac{1}{n} \mathbf{1}^\top v_j \right) \mathbf{1} \right) = \lambda_j^{-1/2} (I_n - M) v_j \in \mathbb{R}^n$ for all $j \in [p]$.

This derivation shows that, as expected, we only need the kernel k (and not the — possibly infinite dimensional — feature mapping ϕ) to apply PCA in the feature space \mathcal{G} .

Remark 3.2.3 (PCA and spectral clustering). See Remark 3.1.4.

RKHS point of view

For the sake of simplicity, let us assume that the dataset is centered in \mathcal{G} : $\frac{1}{n} \sum_{i=1}^n \phi(X_i) = 0$. The previous derivation boils down to $u_j = \sum_{i=1}^n (\alpha_j)_i k(x, X_i)$, where $\alpha_j = \lambda_j^{-1/2} v_j$ for all $j \in [p]$.

Let now \mathcal{H} be the RKHS associated to k and $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^p$ a reduction mapping of the form $\psi(x) = (h_1(x), \dots, h_p(x))$, where $h_j \in \mathcal{H}$ for all $j \in [p]$. Consider the problem of determining ψ such that the components are orthonormal with maximal variance:

$$\begin{aligned}
&\underset{h_1, \dots, h_p \in \mathcal{H}}{\text{maximize}} && \sum_{j=1}^p \mathbb{V}(h_j(X)) \\
&\text{s. t.} && \begin{cases} \forall j \in [p], \|h_j\|_{\mathcal{H}} = 1 \\ \forall i, j \in [p], i \neq j \implies \langle h_i, h_j \rangle_{\mathcal{H}} = 0. \end{cases}
\end{aligned}$$

Remarking that $\frac{1}{n} \sum_{i=1}^n \phi(X_i) = 0 \implies \frac{1}{n} \sum_{i=1}^n h(X_i) = 0, \forall h \in \mathcal{H}$, the empirical point of view of the

latter problem is

$$\begin{aligned} & \underset{h_1, \dots, h_p \in \mathcal{H}}{\text{maximize}} \quad \frac{1}{n} \sum_{\substack{1 \leq j \leq p \\ 1 \leq i \leq n}} h_j(X_i)^2 \\ & \text{s. t.} \quad \begin{cases} \forall j \in [p], \|h_j\|_{\mathcal{H}} = 1 \\ \forall i, j \in [p], i \neq j \implies \langle h_i, h_j \rangle_{\mathcal{H}} = 0, \end{cases} \end{aligned}$$

the maximizers of which being necessarily in $\text{span} \{k(X_i, \cdot), i \in [n]\}$. Thus, considering $h_j = \sum_{i=1}^n (\alpha'_j)_i k(\cdot, X_i)$ for some $\alpha'_j \in \mathbb{R}^n$, the problem of maximal variance becomes

$$\begin{aligned} & \underset{\alpha'_1, \dots, \alpha'_p \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{n} \sum_{j=1}^p \alpha_j'^{\top} K_X^2 \alpha'_j \\ & \text{s. t.} \quad \begin{cases} \forall j \in [p], \alpha_j'^{\top} K_X \alpha'_j = 1 \\ \forall i, j \in [p], i \neq j \implies \alpha_i'^{\top} K_X \alpha'_j = 0. \end{cases} \end{aligned}$$

With the change of variable $v'_j = K_X^{-\frac{1}{2}} \alpha'_j$ (assuming K_X invertible), this boils down to solve

$$\begin{aligned} & \underset{v'_1, \dots, v'_p \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{n} \sum_{j=1}^p v_j'^{\top} K_X v'_j \\ & \text{s. t.} \quad \begin{cases} \forall j \in [p], \|v'_j\|_{\ell_2} = 1 \\ \forall i, j \in [p], i \neq j \implies v_i'^{\top} v'_j = 0. \end{cases} \end{aligned}$$

Remembering that $K_Z = K_X$, it is clear that the p leading eigenvectors v_1, \dots, v_p of K_Z are solution to the latter problem, meaning that $\alpha'_j = K_Z^{-\frac{1}{2}} v_j = \lambda_j^{-1/2} v_j = \alpha_j$ are solutions to the former problem.

To sum up, applying kernel PCA with centered data in \mathcal{G} is equivalent to building a nonlinear reduction mapping $\psi(x) = (h_1(x), \dots, h_p(x))$, where $h_1, \dots, h_p \in \mathcal{H}$ are such that the empirical variance of $h_j(X)$ is maximal and h_1, \dots, h_p are orthonormal.

Remark 3.2.4. When the data is not centered, it should be considered $\psi_j = h_j - \frac{1}{n} \sum_{\ell=1}^p h_{\ell}$, with $h_j = \sum_{i=1}^n (\alpha'_j)_i k(\cdot, X_i)$ for some $\alpha'_j \in \mathbb{R}^n$, which leads for $\alpha'_1, \dots, \alpha'_p$ to be solution to

$$\begin{aligned} & \underset{\alpha'_1, \dots, \alpha'_p \in \mathbb{R}^n}{\text{maximize}} \quad \frac{1}{n} \sum_{j=1}^p \alpha_j'^{\top} K_X (I_n - M) K_X \alpha'_j \\ & \text{s. t.} \quad \begin{cases} \forall j \in [p], \alpha_j'^{\top} K_X \alpha'_j = 1 \\ \forall i, j \in [p], i \neq j \implies \alpha_i'^{\top} K_X \alpha'_j = 0. \end{cases} \end{aligned}$$

It can be shown that $\alpha'_j = (I_n - M) K_Z^{-\frac{1}{2}} v_j = \alpha_j$ is solution to the latter problem, which is consistent with the initial derivation.

3.2.2 Classical multidimensional scaling

In Section 3.1.5, we introduced the paradigm of preserving pairwise distances and showed that it was conceivable with random projections (based on Gaussian matrices). More formally, for any $\epsilon \in (0, 1)$, we exhibited a matrix $W \in \mathbb{R}^{p \times d}$ such that for all pairs of points of interest $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$(1 - \epsilon) \|x - x'\|_{\ell_2}^2 \leq \|Wx - Wx'\|_{\ell_2}^2 \leq (1 + \epsilon) \|x - x'\|_{\ell_2}^2,$$

namely

$$\left| \|x - x'\|_{\ell_2}^2 - \|Wx - Wx'\|_{\ell_2}^2 \right| \leq \epsilon \|x - x'\|_{\ell_2}^2.$$

The approach, called multidimensional scaling (MDS), goes a step further by building representations, not necessarily linear, that tend to preserve pairwise distances. Given a training sample $\{X_i\}_{1 \leq i \leq n} \subseteq \mathbb{R}^d$, MDS proceeds by defining a *stress function* S and minimizing it over the reduced representations $\{Z_i\}_{1 \leq i \leq n}$.

Classical scaling considers that the distance of each pair of points should be preserved, regardless of how far points are, namely

$$\left| \|x - x'\|_{\ell_2}^2 - \|Wx - Wx'\|_{\ell_2}^2 \right| \leq \epsilon.$$

Let $\mathbf{Z} \in \mathbb{R}^{n \times p}$ be the matrix of which the rows are the reduced representations $\{Z_i\}_{1 \leq i \leq n}$. A natural variational formulation of the preceding criterion is to minimize the stress function

$$S_C(\mathbf{Z}) = \sum_{1 \leq i \neq j \leq n} \left(\|X_i - X_j\|_{\ell_2}^2 - \|Z_i - Z_j\|_{\ell_2}^2 \right)^2.$$

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be the matrix of which the rows are the training points $\{X_i\}_{1 \leq i \leq n}$, $D_X = \left(\|X_i - X_j\|_{\ell_2}^2 \right)_{1 \leq i, j \leq n}$ and $D_Z = \left(\|Z_i - Z_j\|_{\ell_2}^2 \right)_{1 \leq i, j \leq n}$ be respectively the squared pairwise distances. Then, we have

$$S_C(\mathbf{Z}) = \|D_X - D_Z\|_F^2.$$

Since minimizing such a function with respect to \mathbf{Z} may be difficult, classical scaling introduces the Gram matrices $K_X = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n}$ and $K_Z = \mathbf{Z}\mathbf{Z}^\top \in \mathbb{R}^{n \times n}$. This makes the problem simple since as soon as we know K_Z , \mathbf{Z} can be obtained by factorization (see below).

It should be noticed that D_Z and K_Z are linked together: let $\delta_X = \text{diag}(K_X) \in \mathbb{R}^n$ be the vector of diagonal items of K_X . Then, one has

$$D_X = \delta_X \mathbf{1}^\top + \mathbf{1} \delta_X^\top - 2K_X. \quad (3.2)$$

However, obtaining K_X from D_X (what we need in practice) is not so easy. That is why we make use of matrix of centered data:

$$\mathbf{X}' = \mathbf{X} - \mathbf{1} \bar{X}^\top,$$

where $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Let now $K_{X'} = \mathbf{X}'\mathbf{X}'^\top$ be the Gram matrix of the centered data.

Property 67. One has

$$K_{X'} = -\frac{1}{2}HD_XH,$$

where $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top \in \mathbb{R}^{n \times n}$.

Proof. The centered data matrix is

$$X' = X - \mathbf{1}\bar{X}^\top = X - \mathbf{1} \left(\frac{1}{n}X^\top \mathbf{1} \right)^\top = HX.$$

So

$$K_{X'} = X'X'^\top = HK_XH.$$

In addition, from Equation (3.2), we have

$$K_X = \frac{1}{2} \left(\delta_X \mathbf{1}^\top + \mathbf{1} \delta_X^\top - D_X \right).$$

Yet, since $H\mathbf{1} = 0$ and $\mathbf{1}^\top H = 0$, we obtain

$$K_{X'} = HK_XH = \frac{1}{2}H \left(\delta_X \mathbf{1}^\top + \mathbf{1} \delta_X^\top - D_X \right) H = -\frac{1}{2}HD_XH.$$

□

Remark 3.2.5. H is the PSD matrix of the orthogonal projection onto the vector space orthogonal to $\text{span}(\mathbf{1})$.

Property 68. One has

$$\|D_X - D_Z\|_F \leq 2(1 + \sqrt{n}) \|K_{X'} - K_Z\|_F.$$

Proof. Using the same notation as in the preceding paragraph and since pairwise distances are the same when the data is centered, $D_X = \delta_{X'} \mathbf{1}^\top + \mathbf{1} \delta_{X'}^\top - 2K_{X'}$. So,

$$\begin{aligned} \|D_X - D_Z\|_F &= \|(\delta_{X'} - \delta_Z) \mathbf{1}^\top + \mathbf{1}(\delta_{X'} - \delta_Z)^\top - 2(K_{X'} - K_Z)\|_F \\ &\leq \|(\delta_{X'} - \delta_Z) \mathbf{1}^\top\|_F + \|\mathbf{1}(\delta_{X'} - \delta_Z)^\top\|_F + 2\|K_{X'} - K_Z\|_F \\ &= 2\sqrt{n} \|\delta_{X'} - \delta_Z\|_{\ell_2} + 2\|K_{X'} - K_Z\|_F \\ &\leq 2(1 + \sqrt{n}) \|K_{X'} - K_Z\|_F. \end{aligned}$$

□

Property 68 tells us that minimizing the distance between $K_{X'}$ and K_Z makes the squared pairwise distances closer. Therefore, we now aim at minimizing the stress function

$$S'_C(Z) = \|K_{X'} - ZZ^T\|_F^2 = \sum_{1 \leq i, j \leq n} \left(\langle X_i - \bar{X}, X_j - \bar{X} \rangle_{\ell_2} - \langle z_i, z_j \rangle_{\ell_2} \right)^2.$$

Such a problem is a low rank approximation problem. As explained in the forthcoming theorem, it is solved by the truncation of the smallest singular values of $K_{X'}$.

Lemma 69. *Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank r , with $A = UDV^T$ being its SVD. Then*

$$\|A\|_F^2 = \sum_{i=1}^r D_{ii}^2.$$

Proof.

$$\|A\|_F^2 = \text{tr}(A^T A) = \text{tr}(VDU^T UDV^T) = \text{tr}(D^2),$$

since U and V are orthogonal matrices. □

Lemma 70 (Weyl's inequality). *Let $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{m \times n}$ be two matrices and let us denote $q = \min(m, n)$. Let $(\sigma_1(\cdot), \dots, \sigma_q(\cdot))$ be the set of singular values (of a given matrix) sorted in decreasing order and completed with 0 after the rank.*

Then, for all $(i, j) \in [q]^2$ such that $i + j - 1 \in [q]$,

$$\sigma_{i+j-1}(A + B) \leq \sigma_i(A) + \sigma_j(B).$$

Theorem 71 (Eckart-Young-Mirsky theorem). *Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank r and let us denote $q = \min(m, n)$. Then, for any $p \in [q]$, a solution to*

$$\underset{B \in \mathbb{R}^{m \times n} : \text{rank}(B) \leq p}{\text{minimize}} \|A - B\|_F$$

is $B^ = A$ if $p \geq r$ and if $p < r$, $B^* = UD'V^T$, where*

- ◇ $A = UDV^T$ is the SVD of A with singular values $D_{11} \geq \dots \geq D_r$;
- ◇ $D' \in \mathbb{R}^{r \times r}$ is such that $D'_{ii} = D_{ii}$ for all $i \in [p]$ and 0 otherwise.

Proof. The result is straightforward if $p \geq r$. Thus, let us consider $p < r$. First, it is easy to see that $\text{rank}(B^*) = p$, since it has p nonzero singular values. Thus, B^* is admissible for the minimization problem.

Second, one has

$$\begin{aligned}
\|A - B^*\|_F^2 &= \|U(D - D')V^\top\|_F^2 \\
&= \text{tr}(V(D - D')U^\top U(D - D')V^\top) \\
&= \text{tr}((D - D')(D - D')) \\
&= \|D - D'\|_F^2 \\
&= \sum_{i=1}^{q-p} \sigma_{i+p}(A)^2, \\
&= \sum_{i=1}^{r-p} \sigma_{i+p}(A)^2,
\end{aligned}$$

since U and V are orthogonal matrices and $D - D'$ is exactly the diagonal matrix of the smallest singular values.

Third, for any matrix $B \in \mathbb{R}^{m \times n}$ such that $\text{rank}(B) \leq p$, for all $i \in [q - p]$, by Weyl's inequality,

$$\sigma_{i+p}(A) = \sigma_{i+(p+1)-1}(A - B + B) \leq \sigma_i(A - B) + \sigma_{p+1}(B) = \sigma_i(A - B),$$

since the rank of B is at most p . Also, $\sigma_{i+p}(A)^2 \leq \sigma_i(A - B)^2$ since singular values are nonnegative.

Therefore,

$$\begin{aligned}
\|A - B^*\|_F^2 &= \sum_{i=1}^{r-p} \sigma_{i+p}(A)^2 \\
&\leq \sum_{i=1}^{r-p} \sigma_i(A - B)^2 \\
&\leq \sum_{i=1}^q \sigma_i(A - B)^2 \\
&= \|A - B\|_F^2.
\end{aligned}$$

□

From Theorem 71, we obtain that S'_C can be minimized by computing a low rank approximation \tilde{K} of K_X and by factorizing \tilde{K} in $\mathbf{Z}\mathbf{Z}^\top$. This is described in Algorithm 15. We can remark that the result obtained for kernel PCA (or centered linear PCA solved with the Gram matrix) is similar to classical MDS (see Equation (3.1) and Remark 3.1.3). In fact, kernel PCA (with centered data) can be seen as classical MDS applied in the feature space \mathcal{G} . There is however a big difference: kernel PCA is a predictive (or inductive) model, while MDS is not (we have to know all points beforehand to transform them).

Remark 3.2.6. *All this derivation is true for Euclidean distance matrices. However, classical MDS may be performed with simple dissimilarity matrices. One should only take care of negative eigen-*

Algorithm 15 Classical multidimensional scaling.

Input: $D \in \mathbb{R}^{n \times n}$ (matrix of squared pairwise distances), $p \in [n]$ (reduced dimension).

$$K_{X'} \leftarrow -\frac{1}{2}HD_XH$$

Compute the eigendecomposition $\sum_{i=1}^n \lambda_i v_i v_i^\top$ of $K_{X'}$, with $\lambda_1 \geq \dots \geq \lambda_n$

$$\mathbf{Z} \leftarrow [\sqrt{\lambda_1} v_1 | \dots | \sqrt{\lambda_p} v_p] \in \mathbb{R}^{n \times p}$$

$\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of \mathbf{Z}

Output: $\{z_i\}_{1 \leq i \leq n}$.

values.

3.2.3 Metric and nonmetric multidimensional scaling

Classical scaling is part of the family of metric scaling, because it tends to preserve pairwise distances. Two other approaches are included in this family.

Kruskal-Shepard

Kruskal-Shepard scaling is a variant of classical scaling, where squares have been dropped. The stress function to minimize is

$$S_{KS}(\mathbf{Z}) = \sum_{1 \leq i \neq j \leq n} \left(\|X_i - X_j\|_{\ell_2} - \|z_i - z_j\|_{\ell_2} \right)^2.$$

In practice, Kruskal-Shepard scaling is solved by the scaling by majorizing a complicated function (SMACOF) algorithm. It consists in majorizing S_{KS} by a convex quadratic function.

Property 72. Let $\alpha \in \mathbb{R}^{n \times n}$ be a symmetric matrix. For any $\mathbf{Y} \in \mathbb{R}^{n \times p}$, with rows denoted $\{y_i\}_{1 \leq i \leq n} \subseteq \mathbb{R}^p$,

$$\sum_{1 \leq i \neq j \leq n} \alpha_{ij} (z_i - z_j)^\top (y_i - y_j) = \text{tr}(\mathbf{Z}^\top \mathbf{V} \mathbf{Y}),$$

where $\mathbf{V} \in \mathbb{R}^{n \times n}$ and for all $i \in [n]$, $V_{ii} = 2 \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \alpha_{ij}$ and for all $j \in [n]$, such that $i \neq j$, $V_{ij} = -2\alpha_{ij}$.

Proof. First,

$$\begin{aligned} \sum_{1 \leq i \neq j \leq n} \alpha_{ij} (z_i - z_j)^\top (y_i - y_j) &= \sum_{1 \leq i \neq j \leq n} \alpha_{ij} (z_i^\top y_i + z_j^\top y_j - z_i^\top y_j - z_j^\top y_i) \\ &= 2 \sum_{i=1}^n \left(\sum_{\substack{1 \leq j \leq n \\ j \neq i}} \alpha_{ij} \right) z_i^\top y_i - 2 \sum_{1 \leq i \neq j \leq n} \alpha_{ij} z_i^\top y_j, \end{aligned}$$

by symmetry of α . Second, for any symmetric matrix $V \in \mathbb{R}^{n \times n}$,

$$\begin{aligned}
\text{tr}(\mathbf{Z}^\top V \mathbf{Y}) &= \sum_{\ell=1}^p \sum_{i=1}^n \sum_{j=1}^n \mathbf{z}_{i,\ell} V_{ij} \mathbf{y}_{j,\ell} \\
&= \sum_{\ell=1}^p \sum_{i=1}^n \sum_{j=1}^n (z_i)_\ell V_{ij} (y_j)_\ell \\
&= \sum_{i=1}^n \sum_{j=1}^n V_{ij} \mathbf{z}_i^\top \mathbf{y}_j \\
&= \sum_{i=1}^n V_{ii} \mathbf{z}_i^\top \mathbf{y}_i + \sum_{1 \leq i \neq j \leq n} V_{ij} \mathbf{z}_i^\top \mathbf{y}_j.
\end{aligned}$$

Finally, identification concludes the proof. \square

Let, for all $i \in [n]$ and $j \in [n]$, $d_{ij} = \|X_i - X_j\|_{\ell_2}$ and $\delta_{ij} = \|z_i - z_j\|_{\ell_2}$. Then,

$$\begin{aligned}
S_{KS}(\mathbf{Z}) &= \sum_{1 \leq i \neq j \leq n} (d_{ij} - \delta_{ij})^2 \\
&= \sum_{1 \leq i \neq j \leq n} d_{ij}^2 + \sum_{1 \leq i \neq j \leq n} \delta_{ij}^2 - 2 \sum_{1 \leq i \neq j \leq n} d_{ij} \delta_{ij}.
\end{aligned}$$

In addition,

$$\sum_{1 \leq i \neq j \leq n} \delta_{ij}^2 = \sum_{1 \leq i \neq j \leq n} (z_i - z_j)^\top (z_i - z_j) = \text{tr}(\mathbf{Z}^\top V \mathbf{Z}),$$

where $V \in \mathbb{R}^{n \times n}$ has $2(n-1)$ on the diagonal and -2 elsewhere, and

$$\sum_{1 \leq i \neq j \leq n} d_{ij} \delta_{ij} = \sum_{1 \leq i \neq j \leq n} \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0} \delta_{ij}^2 = \sum_{1 \leq i \neq j \leq n} \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0} (z_i - z_j)^\top (z_i - z_j) = \text{tr}(\mathbf{Z}^\top V'(\mathbf{Z}) \mathbf{Z}),$$

where $V'(\mathbf{Z}) \in \mathbb{R}^{n \times n}$ has $\left(2 \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \leq n}$ on the diagonal and $\left(-2 \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \neq j \leq n}$ elsewhere.

Thus,

$$S_{KS}(\mathbf{Z}) = \sum_{1 \leq i \neq j \leq n} d_{ij}^2 + \text{tr}(\mathbf{Z}^\top V \mathbf{Z}) - 2 \text{tr}(\mathbf{Z}^\top V'(\mathbf{Z}) \mathbf{Z}).$$

But, for any $\mathbf{Y} \in \mathbb{R}^{n \times p}$,

$$\begin{aligned}
\text{tr}(\mathbf{Z}^\top V'(\mathbf{Y})\mathbf{Y}) &= \sum_{1 \leq i \neq j \leq n} \frac{d_{ij}}{\|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2}} \mathbf{1}_{\|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2} \neq 0} (\mathbf{z}_i - \mathbf{z}_j)^\top (\mathbf{y}_i - \mathbf{y}_j) \\
&\leq \sum_{1 \leq i \neq j \leq n} \frac{d_{ij}}{\|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2}} \mathbf{1}_{\|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2} \neq 0} \|\mathbf{z}_i - \mathbf{z}_j\|_{\ell_2} \|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2} \\
&= \sum_{1 \leq i \neq j \leq n} d_{ij} \delta_{ij} \mathbf{1}_{\|\mathbf{y}_i - \mathbf{y}_j\|_{\ell_2} \neq 0} \\
&\leq \sum_{1 \leq i \neq j \leq n} d_{ij} \delta_{ij},
\end{aligned}$$

by Cauchy-Schwarz and non-negativity of the weights inside the sum. As a consequence, by denoting

$$M_{KS}(\mathbf{Z}, \mathbf{Y}) = \sum_{1 \leq i \neq j \leq n} d_{ij}^2 + \text{tr}(\mathbf{Z}^\top V \mathbf{Z}) - 2 \text{tr}(\mathbf{Z}^\top V'(\mathbf{Y})\mathbf{Y}),$$

which is a convex quadratic function in \mathbf{Z} , we obtain, for all $\mathbf{Y} \in \mathbb{R}^{n \times p}$,

$$S_{KS}(\mathbf{Z}) \leq M_{KS}(\mathbf{Z}, \mathbf{Y}) \quad \text{and} \quad S_{KS}(\mathbf{Z}) = M_{KS}(\mathbf{Z}, \mathbf{Z}).$$

Given $\mathbf{Y} \in \mathbb{R}^{n \times p}$, the minimum of $M_{KS}(\cdot, \mathbf{Y})$ can be obtained by Fermat's rule:

$$0 = \nabla_{\mathbf{Z}} M_{KS}(\mathbf{Z}, \mathbf{Y}) = 2V\mathbf{Z} - 2V'(\mathbf{Y})\mathbf{Y},$$

by symmetry of V . Since V is not necessarily full rank, it cannot be inverted. That is why we resort to the Moore-Penrose inverse of V , denoted V^+ , in order to obtain a minimizer of $M_{KS}(\cdot, \mathbf{Y})$:

$$\mathbf{Z} = V^+ V'(\mathbf{Y})\mathbf{Y}.$$

The resulting procedure is described in Algorithm 16. It provides naturally a series of non-increasing stress values.

Sammon scaling

Sammon scaling goes back to the roots by considering the original criterion for preserving pairwise distances: for all pairs of points of interest $(x, x') \in \mathbb{R}^d \times \mathbb{R}^d$,

$$\left| \|\mathbf{x} - \mathbf{x}'\|_{\ell_2}^2 - \|\mathbf{W}\mathbf{x} - \mathbf{W}\mathbf{x}'\|_{\ell_2}^2 \right| \leq \epsilon \|\mathbf{x} - \mathbf{x}'\|_{\ell_2}^2.$$

A natural variational formulation is to minimize the Sammon mapping with respect to \mathbf{Z} :

$$S_S(\mathbf{Z}) = \sum_{1 \leq i \neq j \leq n} \frac{\left(\|\mathbf{x}_i - \mathbf{x}_j\|_{\ell_2} - \|\mathbf{z}_i - \mathbf{z}_j\|_{\ell_2} \right)^2}{\|\mathbf{x}_i - \mathbf{x}_j\|_{\ell_2}}.$$

Algorithm 16 SMACOF.

Input: $d \in \mathbb{R}^{n \times n}$ (matrix of pairwise distances), $p \in [n]$ (reduced dimension).

$V \leftarrow$ matrix from $\mathbb{R}^{n \times n}$ with $2(n-1)$ on the diagonal and -2 elsewhere

$V^+ \leftarrow$ Moore-Penrose inverse of V

$Z \leftarrow$ random matrix from $\mathbb{R}^{n \times p}$ (*initialization*)

while not converged **do**

$\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of Z

$\delta_{ij} \leftarrow \|z_i - z_j\|_{\ell_2}$ for all $(i, j) \in [n]$

$V' \leftarrow$ matrix from $\mathbb{R}^{n \times n}$ with $\left(2 \sum_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \leq n}$ on the diagonal and $\left(-2 \frac{d_{ij}}{\delta_{ij}} \mathbf{1}_{\delta_{ij} \neq 0}\right)_{1 \leq i \neq j \leq n}$ elsewhere.

$Z \leftarrow V^+ V' Z$

end while

$\{z_i\}_{1 \leq i \leq n} \leftarrow$ rows of Z

Output: $\{z_i\}_{1 \leq i \leq n}$.

Nonmetric scaling

In some applications, like wine tasting for instance, pairwise distances are not as important as ranking of them: if for some $(i, j, i', j') \in [n]^4$, $\|X_i - X_j\|_{\ell_2} \geq \|X_{i'} - X_{j'}\|_{\ell_2}$, we would like a representation Z that fulfills $\|z_i - z_j\|_{\ell_2} \geq \|z_{i'} - z_{j'}\|_{\ell_2}$. The major interest here is preserving the ordinal properties of the data. For this reason, nonmetric scaling aims at minimizing the stress function

$$S_{NM}(Z, \varphi) = \frac{\sum_{1 \leq i \neq j \leq n} \left(\varphi \left(\|X_i - X_j\|_{\ell_2} \right) - \|z_i - z_j\|_{\ell_2} \right)^2}{\sum_{1 \leq i \neq j \leq n} \|z_i - z_j\|_{\ell_2}^2},$$

over representations Z and monotonically increasing functions φ .

A naive algorithm in order to approximate a minimizer of S_{NM} is to alternate minimization over Z (for instance thanks to a subgradient descent) for a fixed φ , and isotonic regression to approximate φ given Z .

3.3 Other methods

3.3.1 Spectral embedding

As explained in Section 2.2.5, spectral clustering boils down to finding a novel representation of the training data and then performing a k-means. This new representation is in fact a dimensionality reduction technique, called *spectral embedding*.

3.3.2 Linear discriminant analysis

Dimensionality reduction can be performed naturally in a supervised manner, taking into consideration the derivation of multiclass discriminant analysis (Section 1.1.5). It has been shown that the (let us say, p)

leading eigenvectors of $\Sigma^{-1}M$ (see notation in Section 1.1.5), denoted $(v_1, \dots, v_p) \subseteq \mathbb{R}^n$, concentrate the variability between features. Thus, $x \in \mathbb{R}^d \mapsto [v_1 | \dots | v_p]^\top x \in \mathbb{R}^p$ defines a dimensionality reduction mapping. In fact, when $p = C - 1$ (C being the number of classes), this mapping projects the data onto the subspace spanned by the class centers, which is enough to discriminate points.

3.4 Exercises

3.4.1 Random projection

Exercise 3.1 (Concentration of a chi-squared variable). Let k be a positive integer and $Z \sim \chi_k^2$. The moment-generating function of Z is defined for all $\lambda < \frac{1}{2}$ by:

$$\mathbb{E}(e^{\lambda Z}) = (1 - 2\lambda)^{-\frac{k}{2}}.$$

1. Show that for all $x \leq \frac{1}{4}$:

$$\frac{1}{\sqrt{1-2x}} \leq e^{x+2x^2} \mathbf{1}_{x \geq 0} + e^{x+x^2} \mathbf{1}_{x < 0}.$$

Prove that for all $\epsilon \in [0, 1]$,

$$\mathbb{P}\left(\frac{Z}{k} - 1 \geq \epsilon\right) \leq e^{-k \frac{\epsilon^2}{8}}.$$

Prove that for all $\epsilon \geq 0$,

$$\mathbb{P}\left(-\left(\frac{Z}{k} - 1\right) \geq \epsilon\right) \leq e^{-k \frac{\epsilon^2}{4}}.$$

Deduce that $\forall \epsilon \in [0, 1]$,

$$\mathbb{P}\left(\left|\frac{Z}{k} - 1\right| \geq \epsilon\right) \leq 2e^{-k \frac{\epsilon^2}{8}}.$$

Chapter 4

Previous exams

Examen : Introduction à l'apprentissage automatique

18 décembre 2019

Tous les documents et les ordinateurs connectés sont autorisés.

Les questions peuvent être traitées de manière indépendante en admettant le résultat des questions précédentes.

Le barème (sur 19 points, auxquels s'ajoutent 4 points bonus) n'est donné qu'à titre indicatif.

Exercice 1 (Une variante des SVM, 5½ points)

Soient $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un échantillon de n couples aléatoires et $C \geq 0$. Pour tout $i \in [n]$, on appelle

$$\ell_i : x \in \mathbb{R} \mapsto \frac{Y_i + 3}{2} \max(0, 1 - Y_i x)$$

et on considère le problème d'optimisation :

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{1}{2} b^2 + C \sum_{i=1}^n \ell_i(w^\top X_i + b). \quad (\text{P1})$$

1. a) (1 point) Tracer les graphes de ℓ_i pour $i \in [n]$ tel que $Y_i = 1$ et $Y_i = -1$. La résolution de (P1) permet-il de construire un régresseur ou un classifieur ? Expliquer votre réponse.
Quelle est la différence entre le problème d'intérêt et une machine à vecteurs supports telle que vue en cours ?
- b) (1 point) Donner la forme de la fonction de *prédiction* en fonction d'un couple (\hat{w}, \hat{b}) solution de (P1). Est-ce un prédicteur linéaire ou non-linéaire ? Réaliser une figure illustrant le problème en jeu et la prédiction.
2. (1½ points) En notant $p = (\mathbf{1}_{Y_1=1}, \dots, \mathbf{1}_{Y_n=1}) \in \mathbb{R}^n$, $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, $A = \begin{pmatrix} Y_1 X_1^\top \\ \vdots \\ Y_n X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ et $\mathbf{1}$ le vecteur rempli de 1 (de taille adéquate), montrer que (P1) peut se réécrire

$$\begin{aligned} & \underset{w \in \mathbb{R}^n, b \in \mathbb{R}, \xi \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{1}{2} b^2 + C \mathbf{1}^\top \xi + C p^\top \xi \\ & \text{s. t.} \quad \begin{cases} Aw + by \succcurlyeq \mathbf{1} - \xi \\ \xi \succcurlyeq 0. \end{cases} \end{aligned} \quad (\text{P2})$$

3. (2 points) Établir le problème d'optimisation dual à (P2).

Exercice 2 (Régression à noyau, 3 points)

Soient \mathcal{H} un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \mathbb{R}$ un échantillon de n couples aléatoires. On appelle $K = (k(X_i, X_j))_{1 \leq i, j \leq n}$ la matrice noyau, supposée inversible.

Pour $\alpha > 0$, on considère le régresseur \hat{f} défini par :

$$\{\hat{f}\} = \arg \min_{f \in \mathcal{H}} \left\{ \alpha \|f\|_{\mathcal{H}}^2 + \sum_{i=1}^n (Y_i - f(X_i))^2 \right\}.$$

1. (2 points) En faisant appel aux résultats du cours, justifier que $\hat{f} = \sum_{i=1}^n \hat{\beta}_i k(\cdot, X_i)$, où $\hat{\beta} \in \mathbb{R}^n$ est solution de

$$\arg \min_{\beta \in \mathbb{R}^n} \left\{ \alpha \beta^\top K \beta + \|y - K \beta\|_{\ell_2}^2 \right\},$$

où $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$.

2. (1 point) Montrer que $\hat{\beta} = (K + \alpha I_n)^{-1} y$, où $I_n \in \mathbb{R}^{n \times n}$ est la matrice identité.

Exercice 3 (Approximation de la matrice noyau, 10 $\frac{1}{2}$ points)

Soient \mathcal{H} un espace de Hilbert à noyau reproduisant de noyau $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, $\{x_1, \dots, x_n\} \subset \mathbb{R}^d$ un ensemble de n points fixés et $m \leq n$ un entier non-nul.

On souhaite analyser quelques propriétés d'une méthode d'approximation de la matrice noyau $K = (k(x_i, x_j))_{1 \leq i, j \leq n}$ par les m premiers points x_i . Pour ce faire, notons $K^m = (k(x_i, x_j))_{1 \leq i, j \leq m}$ la matrice noyau de $\{x_1, \dots, x_m\}$ (supposée inversible) et $P_m : \mathbb{R}^d \rightarrow \mathbb{R}^d$ le projecteur orthogonal sur $\mathcal{V} = \text{span}\{k(\cdot, x_1), \dots, k(\cdot, x_m)\}$.

Par ailleurs, pour tout $f \in \mathcal{H}$, on appelle *coordonnées de $P_m f$ dans \mathcal{V}* un vecteur $\alpha \in \mathbb{R}^m$ vérifiant $P_m f = \sum_{\ell=1}^m \alpha_\ell k(\cdot, x_\ell)$.

1. a) (1 point) Soit $f \in \mathcal{H}$. Montrer que les coordonnées de $P_m f$ dans \mathcal{V} s'expriment par $\alpha = (K^m)^{-1} F$, où $F \in \mathbb{R}^m$ est un vecteur à déterminer.
- b) (1 point) On appelle, pour tout $i, j \in [n]$, $f_i = k(\cdot, x_i)$ et $\tilde{K}_{i,j} = \langle P_m f_i, P_m f_j \rangle_{\mathcal{H}}$. Expliquer en quoi $\tilde{K}_{i,j}$ peut être vu comme une approximation de $K_{i,j}$ et justifier que la matrice $\tilde{K} \in \mathbb{R}^{n \times n}$ est symétrique semi-définie positive.
- c) (1 point) Montrer que $\sum_{i=1}^n \|f_i - P_m f_i\|_{\mathcal{H}}^2 = \text{tr}(K) - \text{tr}(\tilde{K})$.
- d) (1 point) Montrer que pour tout $i, j \in [n]$, $K_{i,j} - \tilde{K}_{i,j} = \langle f_i - P_m f_i, f_j - P_m f_j \rangle_{\mathcal{H}}$. En déduire que $K - \tilde{K}$ est symétrique semi-définie positive.
2. a) (1 point) Pour tout $i \in [n]$, on appelle $\alpha_i \in \mathbb{R}^m$ le vecteur des coordonnées de $P_m f_i$ dans \mathcal{V} . Exprimer de manière matricielle $\tilde{K}_{i,j}$ ($i, j \in [n]$) en fonction de α_i et α_j .

- b) (1 point) Soit $\Lambda = (\alpha_1 | \dots | \alpha_n) \in \mathbb{R}^{m \times n}$. En identifiant la matrice K par blocs : $K = \begin{pmatrix} K^m & Q \\ Q^\top & K' \end{pmatrix}$, exprimer Λ et en déduire que $\tilde{K} = (K^m | Q)^\top (K^m)^{-1} (K^m | Q)$.
3. a) ($\frac{1}{2}$ point) Comparer l'espace mémoire nécessaire pour stocker K et \tilde{K} (tel qu'exprimé à la question 2. b).
- b) (1 point) Montrer que $\|K - \tilde{K}\|_F = \|K' - Q^\top (K^m)^{-1} Q\|_F$, où $\|\cdot\|_F$ est la norme de Frobenius.
- c) (1 point) Soit $X \in \mathbb{R}^{n \times d}$ la matrice des données (rangées en ligne). En identifiant X par blocs : $X = \begin{pmatrix} X^m \\ X' \end{pmatrix}$, avec $X^m \in \mathbb{R}^{m \times d}$ la matrice des m premiers points, supposée de rang plein, on appelle $P_\top : \mathbb{R}^d \rightarrow \mathbb{R}^d$ le projecteur orthogonal sur l'espace engendré par les colonnes de $(X^m)^\top$ (c'est-à-dire les lignes de X^m). Montrer que $P_\top = (X^m)^\top (X^m (X^m)^\top)^{-1} X^m$.
- d) (2 points) On considère (dans cette question uniquement) que k est le noyau linéaire, autrement dit $K = X X^\top$. Déduire de la question précédente que $\tilde{K} = X P_\top X^\top$.
4. a) (1 point (bonus)) En nommant $K = R D R^\top$ la décomposition en éléments propres de K ($D \in \mathbb{R}^{n \times n}$ est une matrice diagonale, $R \in \mathbb{R}^{n \times n}$ une matrice orthogonale) et $R = \begin{pmatrix} R^m \\ R' \end{pmatrix}$ la décomposition par blocs de R avec $R^m \in \mathbb{R}^{m \times n}$, exprimer K en fonction de D , R^m et R' .
- b) (1 point (bonus)) On suppose que $\text{rank}(K) = \text{rank}(K^m) = m$. Que peut-on en déduire sur D ? Exprimer $(K^m)^{-1}$ en fonction de blocs de D et de R^m .
- c) (2 points (bonus)) En déduire que lorsque $\text{rank}(K) = \text{rank}(K^m) = m$, $\tilde{K} = K$.

Examen : Introduction à l'apprentissage automatique

3 juillet 2020

Tous les documents et les ordinateurs connectés sont autorisés.

Les exercices sont indépendants les uns des autres.

Les questions peuvent être traitées de manière indépendante en admettant le résultat des questions précédentes.

Le barème (sur 20 points, auxquels s'ajoutent 3 points bonus) n'est donné qu'à titre indicatif.

Notation

Dans tout le sujet, on notera

$$\text{sign} : x \in \mathbb{R} \mapsto \begin{cases} 1 & \text{si } x > 0 \\ -1 & \text{sinon.} \end{cases}$$

Exercice 1 (Classifieur de Bayes, 9 points)

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$. On appelle $\pi = \mathbb{P}(Y = 1)$.

1. (1 point) Rappeler la définition du classifieur de Bayes.
2. Supposons que $d = 2$, c'est-à-dire $X = (X_1, X_2)$ avec X_1 et X_2 les composantes de X , qui sont des variables aléatoires réelles. Supposons de plus que pour tout $y \in \{\pm 1\}$, $\exists m_y \in \mathbb{R} : [X_1|Y = y] \sim \mathcal{N}(m_y, 1)$, $[X_2|Y = y] \sim \mathcal{U}([0, 1])$ et $X_1 \perp\!\!\!\perp X_2$.
 - a) (1 point) Dessiner un schéma illustrant le problème de classification (on pourra représenter les densités).
 - b) (1 point) Quelle est la densité de $[X|Y = 1]$?
 - c) (1 point) Calculer le classifieur de Bayes associé au problème.
 - d) (1 point) Sans détailler les calculs, quel aurait été le classifieur de Bayes si nous avions supposé $d = 3$ et pour tout $y \in \{\pm 1\}$, $\exists m_y \in \mathbb{R}^2 : [(X_1, X_2)|Y = y] \sim \mathcal{N}(m_y, C)$ (où C une matrice 2×2 symétrique définie positive), $[X_3|Y = y] \sim \mathcal{U}([0, 1])$ et $(X_1, X_2) \perp\!\!\!\perp X_3$?
3. On appelle $\eta : x \in \mathbb{R}^d \mapsto \mathbb{E}[Y|X = x]$ la fonction de régression du problème et $h^* : x \in \mathbb{R}^d \mapsto \text{sign}(\eta(x))$.

- a) (1 point) En développant l'espérance, montrer que h^* est le classifieur de Bayes du problème.
- b) (1 point) Soit $h : \mathbb{R}^d \rightarrow \{\pm 1\}$. Montrer que $\mathbb{P}(Y \neq h(X)) = \frac{1 - \mathbb{E}[Yh(X)]}{2}$.
- c) (1 point) Soit $x \in \mathbb{R}^d$. En remarquant que pour tout $u \in \mathbb{R}^*$: $\text{sign}(u) = \frac{|u|}{u}$, déterminer le signe de $\mathbb{E}[Y(h^*(X) - h(X)) | X = x]$.
- d) (1 point) En déduire que h^* est un minimiseur de $h \mapsto \mathbb{P}(Y \neq h(X))$ sur l'ensemble $\{\pm 1\}^{\mathbb{R}^d}$ des fonctions de \mathbb{R}^d dans $\{\pm 1\}$.

Exercice 2 (Analyse linéaire discriminante, 6 points)

Soit (X, Y) un couple de variables aléatoires à valeurs dans $\mathbb{R}^d \times \{\pm 1\}$. On appelle $\pi = \mathbb{P}(Y = 1)$ et on suppose qu'il existe une matrice $\Sigma \in \mathbb{R}^{d \times d}$ symétrique définie positive telle que pour tout $y \in \{\pm 1\}$, $\exists \mu_y \in \mathbb{R}^d : [X | Y = y] \sim \mathcal{N}(\mu_y, \Sigma)$ et $\mu_1 \neq \mu_{-1}$.

- 1. (1 point) Expliciter la densité de probabilité du vecteur aléatoire X en fonction des densités respectives $p_1 : \mathbb{R}^d \rightarrow \mathbb{R}_+$ et $p_{-1} : \mathbb{R}^d \rightarrow \mathbb{R}_+$ de $[X | Y = 1]$ et $[X | Y = -1]$.
- 2. (1 point) Soit $h^* : x \in \mathbb{R}^d \mapsto \text{sign}(\pi p_1(x) - (1 - \pi)p_{-1}(x))$. Montrer l'équivalence, pour tout $x \in \mathbb{R}^d$:

$$h^*(x) = 1 \iff (\mu_1 - \mu_{-1})^\top \Sigma^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2} \right) > \log \left(\frac{1 - \pi}{\pi} \right).$$

- 3. (1 point) Donner une interprétation géométrique de la condition précédente lorsque $\Sigma = I_d$ (la matrice identité de taille d) et $\pi = \frac{1}{2}$.
- 4. (1 point) Soit $d_\Sigma : (x, y) \in \mathbb{R}^d \times \mathbb{R}^d \mapsto \sqrt{(x - y)^\top \Sigma^{-1} (x - y)}$ la distance de Mahalanobis induite par Σ . Donner une interprétation géométrique de cette distance.
- 5. (1 point) Soit $Z \sim \mathcal{N}(\mu_{-1}, \Sigma)$. Montrer que

$$(\mu_1 - \mu_{-1})^\top \Sigma^{-1} (Z - \mu_{-1}) \sim \mathcal{N}(0, d_\Sigma(\mu_1, \mu_{-1})^2).$$

- 6. (1 point) Soit $\Phi : \mathbb{R} \rightarrow]0, 1[$ la fonction de répartition de $\mathcal{N}(0, 1)$. Supposons que $\pi = \frac{1}{2}$. Montrer que

$$\mathbb{P}(h^*(X) = 1 | Y = -1) = 1 - \Phi \left(\frac{d_\Sigma(\mu_1, \mu_{-1})}{2} \right).$$

- 7. (1 point (bonus)) En déduire une expression de $\mathbb{P}(h^*(X) \neq Y)$ lorsque $\pi = \frac{1}{2}$.

Exercice 3 (Une variante des SVM, 5 points)

Soient $\{(X_1, Y_1), \dots, (X_n, Y_n)\} \subset \mathbb{R}^d \times \{\pm 1\}$ un échantillon de n couples aléatoires et

$C > 0$. On considère le problème d'optimisation :

$$\begin{aligned} & \underset{w \in \mathbb{R}^n, b \in \mathbb{R}, s \in \mathbb{R}^n}{\text{minimize}} && \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{C}{2} \sum_{i=1}^n s_i^2 \\ & \text{s. t.} && \begin{cases} \forall i \in \llbracket 1, n \rrbracket, s_i \geq 0 \\ Y_i(w^\top X_i + b) \geq 1 - s_i. \end{cases} \end{aligned} \quad (\text{P1})$$

- (1 point) La résolution de (P1) permet-il de construire un régresseur ou un classifieur ? Celui-ci est-il linéaire ou non-linéaire ? Expliquer votre réponse en réalisant une figure illustrant le problème en jeu.
- (1 point) Montrer que tout couple (\hat{w}, \hat{b}) solution de (P1) est aussi solution du problème

$$\underset{w \in \mathbb{R}^n, b \in \mathbb{R}}{\text{minimize}} \quad \frac{1}{2} \|w\|_{\ell_2}^2 + C \sum_{i=1}^n \ell_i(w^\top X_i + b),$$

où $\{\ell_i : \mathbb{R} \rightarrow \mathbb{R}_+\}_{1 \leq i \leq n}$ sont des fonctions à préciser.

Quelle est la différence entre le problème d'intérêt et une machine à vecteurs supports telle que vue en cours ?

- (1 point) En notant $y = (Y_1, \dots, Y_n) \in \mathbb{R}^n$, $A = \begin{pmatrix} Y_1 X_1^\top \\ \vdots \\ Y_n X_n^\top \end{pmatrix} \in \mathbb{R}^{n \times d}$ et $\mathbf{1}$ le vecteur rempli de 1 (de taille adéquate), montrer que (P1) peut se réécrire

$$\begin{aligned} & \underset{w \in \mathbb{R}^n, b \in \mathbb{R}, s \in \mathbb{R}^n}{\text{minimize}} && \frac{1}{2} \|w\|_{\ell_2}^2 + \frac{C}{2} \|s\|_{\ell_2}^2 \\ & \text{s. t.} && \begin{cases} Aw + by \succcurlyeq \mathbf{1} - s \\ s \succcurlyeq 0. \end{cases} \end{aligned} \quad (\text{P2})$$

- (2 points) Montrer que le problème d'optimisation dual à (P2) est

$$\begin{aligned} & \underset{\alpha \in \mathbb{R}^n}{\text{maximize}} && \mathbf{1}^\top \alpha - \frac{1}{2} \|A^\top \alpha\|_{\ell_2}^2 - \frac{1}{2C} \|\alpha\|_{\ell_2}^2 \\ & \text{s. t.} && \alpha \succcurlyeq 0, \quad y^\top \alpha = 0. \end{aligned}$$

- (2 points (bonus)) Soit $f : w \in \mathbb{R}^d \mapsto \ell_1(w^\top X_1)$. Déterminer ∇f et montrer que ∇f est une application L -lipschitzienne pour une constante L à préciser.

Bibliography

- T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2003.
- C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media, 2013.
- W.K. Härdle and L. Simar. *Applied Multivariate Statistical Analysis*. Springer, 2015.
- M. Mažeika. The singular value decomposition and low rank approximation. Technical report, University of Chicago, 2016.
- M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.
- S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- I. Steinwart and A. Christmann. *Support Vector Machines*. Springer, New York, NY, 2008.
- U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.