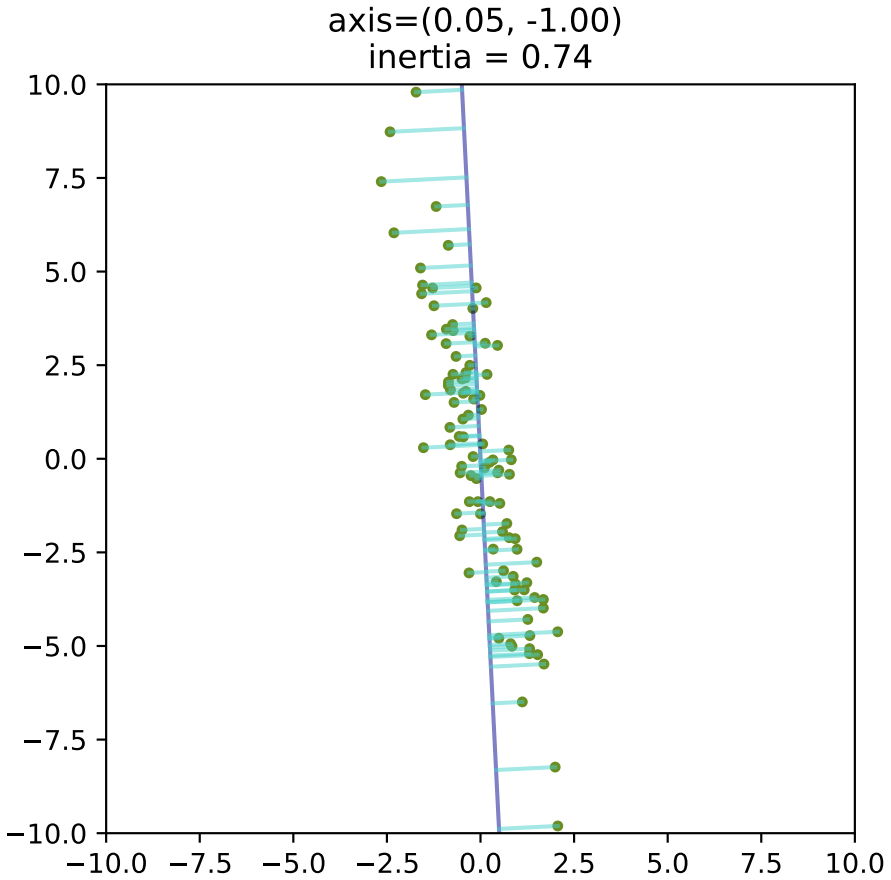
Fondamentaux théoriques du machine learning



Overview of lecture 5

Dimensionality reduction

- Principal component analysis

Probabilistic modelling

Classification

- Problem statement
- Convexification of the risk and calibration
- Logistic regression

Dimensionality reduction

We consider the space $\mathcal{X}$ that contains the data, for either supervised or unsupervised learning. In machine learning, we often have $\mathcal{X} \in \mathbb{R}^d$.

- If d is large (e.g. $\geq 10^4$), the algorithms that run on the data might become too slow to be used, as their algorithmic complexity depends on d (potentially in a quadratic or exponential way, curse of dimensionality)

Dimensionality reduction

We consider the space $\mathcal{X}$ that contains the data, for either supervised or unsupervised learning. In machine learning, we often have $\mathcal{X} \in \mathbb{R}^d$.

- If d is large (e.g. $\geq 10^4$), the algorithms that run on the data might become too slow to be used, as their algorithmic complexity depends on d (potentially in a quadratic or exponential way, curse of dimensionality)
- **However**, often the data might actually occupy a **subspace** of lower dimension q, or it may be possible to project the data on such a subspace without loosing too much information.
  - Working in a subspace of lower dimension might speed up the algorithms.
  - It may also allow visualization of the data.

Main methods of dimensionality reduction

- **feature selection** : selecting a subset of the original dimensions.
- **feature extraction** : computing new features from the original features.

Principal component analysis (PCA)

- PCA is a **linear feature extraction** technique.
- Points in $\mathbb{R}^d$ are linearly projected on a well chosen affine subspace of $\mathbb{R}^q$, with $q \leq d$.

Formalization as a (empirical) variance maximimisation problem

Without loss of generality, we assume the data are **centered**, which means that

$$\bar{x} = \sum_{i=1}^{n} x_n = 0 \in \mathbb{R}^d \tag{1}$$

X is the design matrix like in OLS. The **first principal componenent** is a vector $w \in \mathbb{R}^d$, with $||w|| = 1$, such that $\hat{Var}(w^T x)$ is maximal, where $\hat{Var}$ denotes the empirical variance.

Variance

$$\overline{w^T x} = w^T \overline{x} = 0 \tag{2}$$

Hence,

$$\begin{aligned} \hat{Var}(w^T x) &= \frac{1}{n-1} \sum_{i=1}^{n} \left((w^T x)_i - \overline{w^T x} \right)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^{n} \left(w^T x_i \right)^2 \end{aligned} \tag{3}$$

Variance maximisation problem

We can then formulate the problem as finding w, $||w|| = 1$ such that

$$\sum_{i=1}^{n} \left(w^T x_i\right)^2 \tag{4}$$

is maximal.

First principal component

We look for w, $||w|| = 1$ such that

$$\sum_{i=1}^{n} \left(w^T x_i\right)^2 \tag{5}$$

is maximal.

Proposition

w is the eigenvector of $X^T X$ with largest eigenvalue λ_{max}.

First principal component

We look for w, $||w|| = 1$ such that

$$\sum_{i=1}^{n} \left(w^T x_i\right)^2 \tag{6}$$

is maximal.

Proposition

w is the eigenvector of $X^T X$ with largest eigenvalue λ_{max}.

Exercice 1 : Show the proposition.

Reconstruction error

Alternately, we can formulate the problem as a **reconstruction error minimization.**

$$\mathbb{R}^d = \text{Vect}(w) \bigoplus \text{Vect}(w)^{\perp} \tag{7}$$

and if $||w|| = 1$,

$$\begin{aligned} \forall x \in \mathbb{R}^d, ||x||^2 &= ||(x^T w)w||^2 + ||x - (x^T w)w||^2 \\ &= (x^T w)^2 + ||x - (x^T w)w||^2 \end{aligned} \tag{8}$$

Reconstruction error

We can formulate the problem as a **reconstruction error minimization.** If $||w|| = 1$,

$$\begin{aligned} \forall x \in \mathbb{R}^d, ||x||^2 &= ||(x^T w)w||^2 + ||x - (x^T w)w||^2 \\ &= (x^T w)^2 + ||x - (x^T w)w||^2 \end{aligned} \tag{9}$$

Hence,

$$\begin{aligned} \sum_{i=1}^{n} ||x_i||^2 &= \sum_{i=1}^{n} (x_i{}^T w)^2 + \sum_{i=1}^{n} ||x_i - (x_i{}^T w)w||^2 \\ &= \hat{Var}(w^T x) + \sum_{i=1}^{n} ||x_i - (x_i{}^T w)w||^2 \end{aligned} \tag{10}$$

Reconstruction error

$$\sum_{i=1}^{n} ||x_i||^2 = \hat{Var}(w^T x) + \sum_{i=1}^{n} ||x_i - (x_i^T w)w||^2 \qquad (11)$$

We can see $\sum_{i=1}^{n} ||x_i - (x_i^T w)w||^2$ as a **reconstruction error**, when the data are projected on Vect(w).
Maximizing the variance of the projections is equivalent to minimizing the reconstruction errors obtained by projection.

Several principal components

Most of the time, we project the data on several principal components.

- 1] compute the first principal component w_1
- 2] project the data on $\text{Vect}(w_1)^\perp$
- 3] start again on the projected data

Reconstruction error

The interpretation stays the same. If $p_F(x)$ is the projection of x on the subspace spanned by the principal components :

$$||x||^2 = ||p_F(x)||^2 + ||x - p_F(x)||^2 \tag{12}$$

The principal components are the largest eigenvectors of $X^T X \in \mathbb{R}^d, d$ with norm 1. They are **orthogonal** to each other.

Inertia

We can define an inertia :

$$I_F = \sum_{i=1}^{n} ||x - p_F(x)||^2 \tag{13}$$

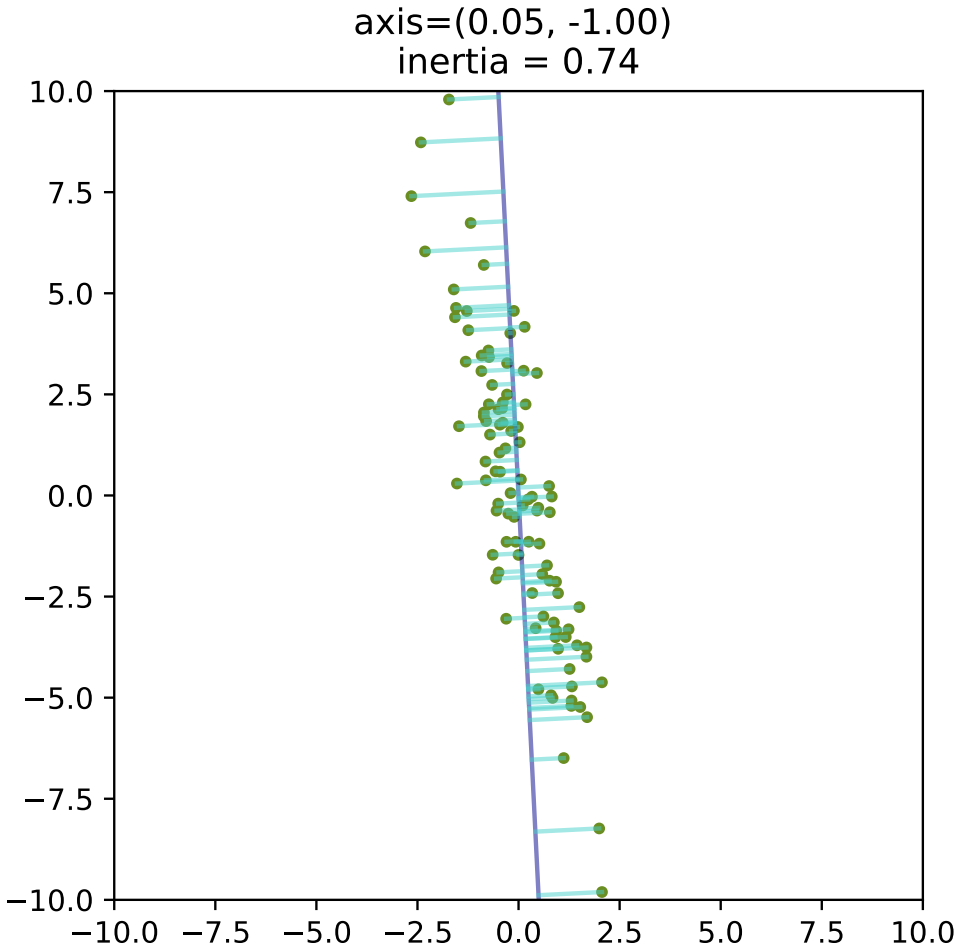
We lool for the subspace that minimizes the inertia I_F.

Inertia

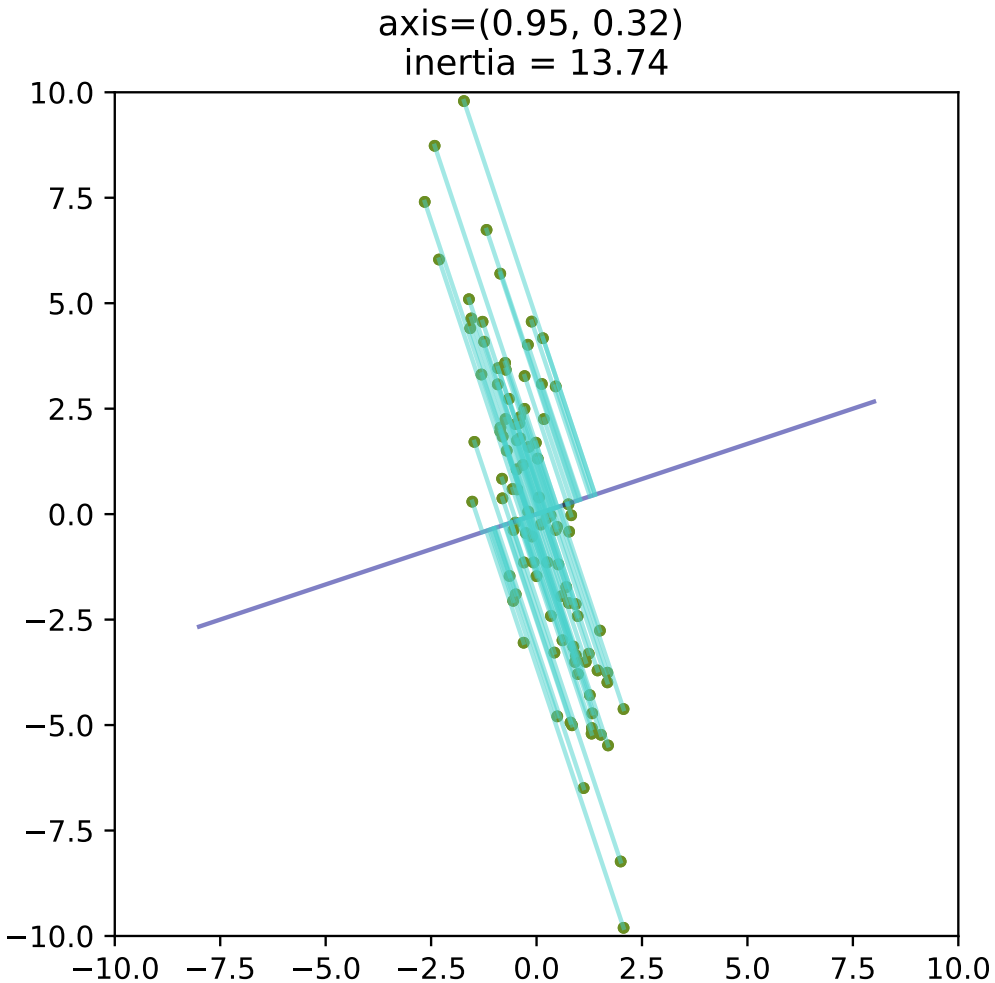
Exercice 2 : **No inertia**

In what situations could we have $I_F = 0$?

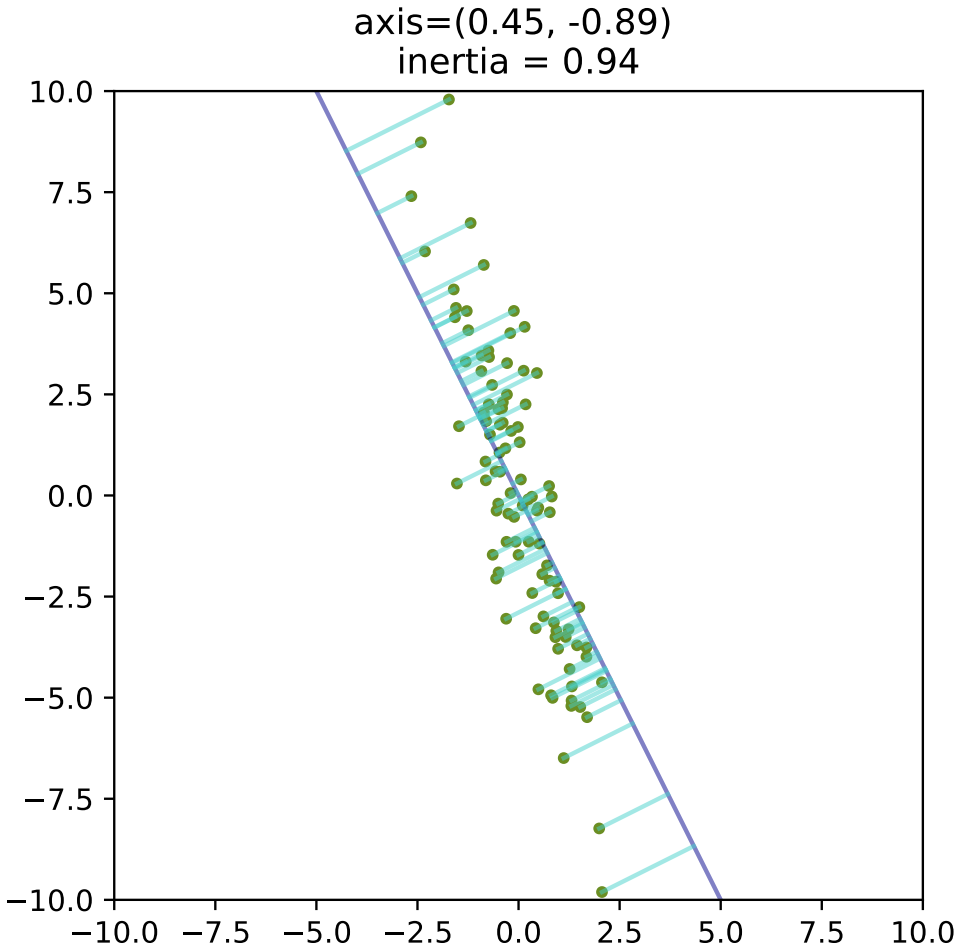
Inertia


axis=(0.05, -1.00)
inertia = 0.74
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Inertia


axis=(0.95, 0.32)
inertia = 13.74
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Inertia


axis=(0.45, -0.89)
inertia = 0.94
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

Iris dataset

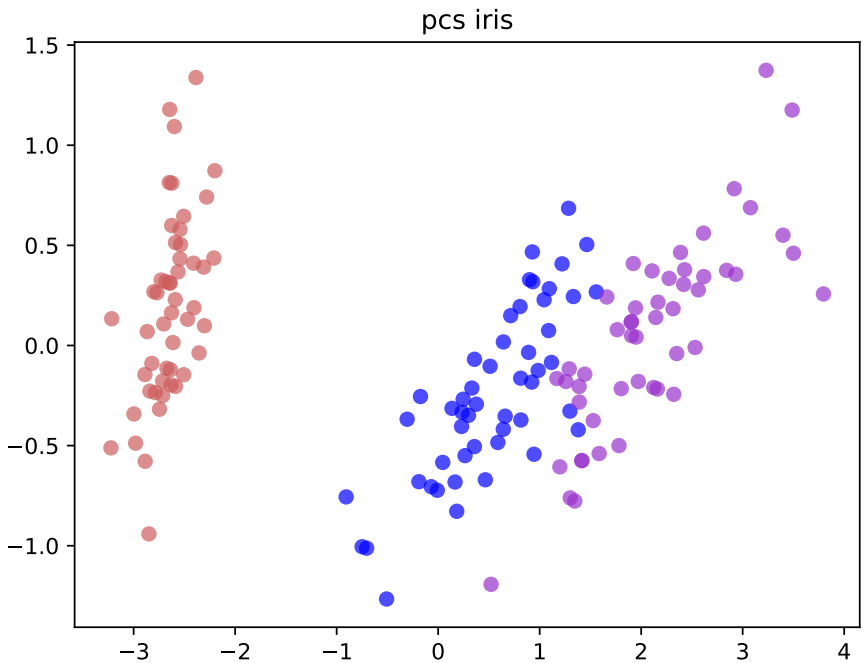


Figure – PCA performed on the iris dataset, keeping 2 dimensions. We see that the principal components are able to separate the data.

In this paper, astrophysicists use PCA in order to test a new star temperature prediction method [Bermejo et al., 2013]

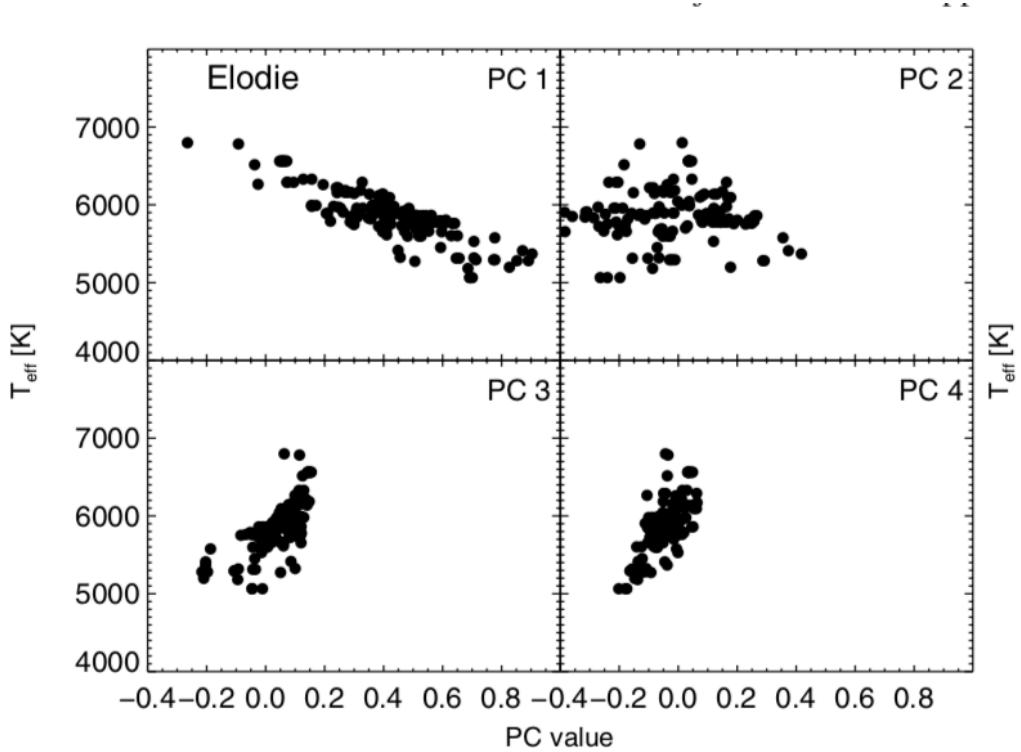
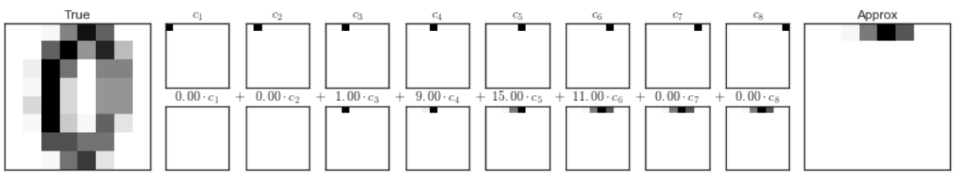


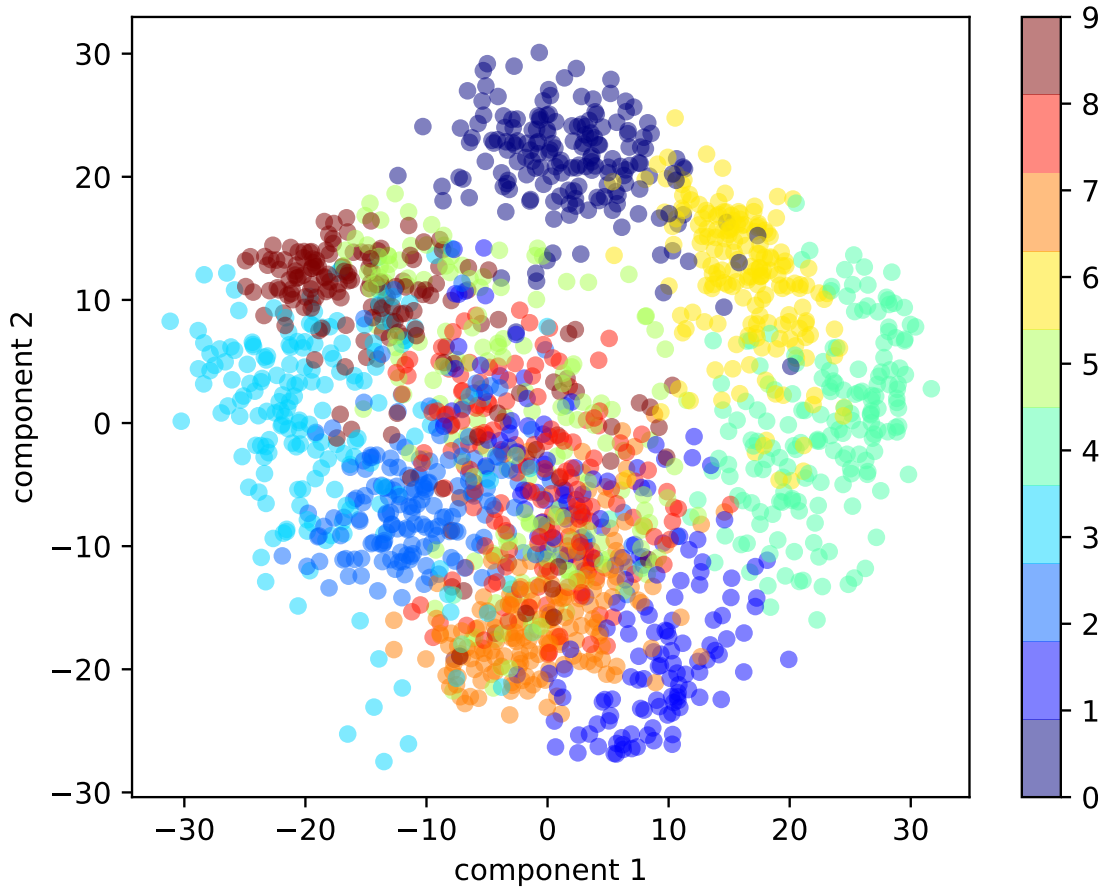
Figure – PCA used in order to predict temperature.

PCA on digits

- We can perform the PCA on a dataset consisting in 8×8 pixels images of digits, in order to see if the PCA allows a visualization of some structure in the data.



PCA on digits



PCA on digits : reconstruction

With 8 principal components, we can monitor the reconstruction of the images (originally in 64 dimensions)

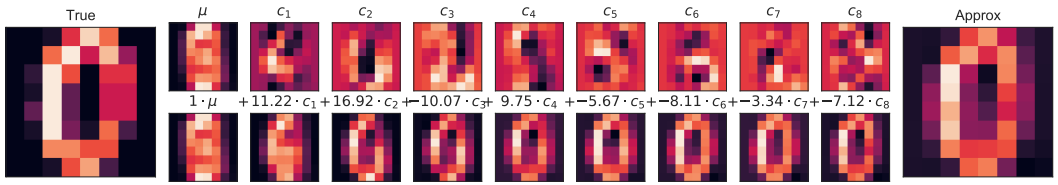


Figure – Reconstruction of 0

`https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html`

PCA on digits : reconstruction

With 8 principal components, we can monitor the reconstruction of the images (originally in 64 dimensions)

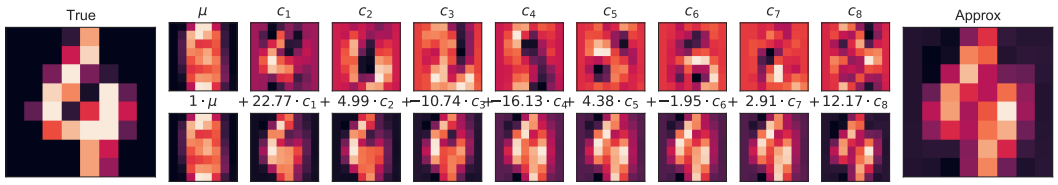


Figure – Reconstruction of 4

`https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html`

Explained variance

A natural question is : what is a relevant number of principal components ?
A common quantity that is used is **explained variance**. Each component w_k carries a percentage of the total variance of the data.

$$\frac{\hat{Var}(w_k^T x)}{\sum_{j=1}^{d} \hat{Var}(x^j)} \tag{14}$$

where $\hat{Var}(x^j)$ is the variance of the component j.

$$\hat{Var}(x^j) = \sum_{i=1}^{n} (x_i^j)^2 \tag{15}$$

Number of components

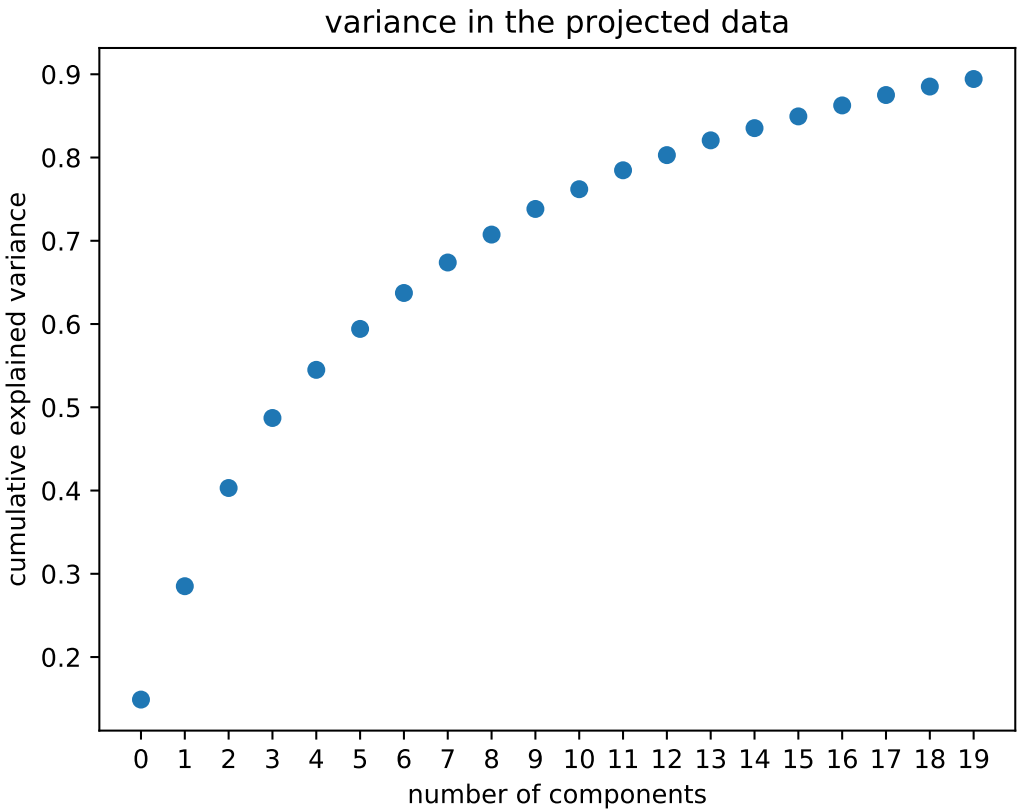
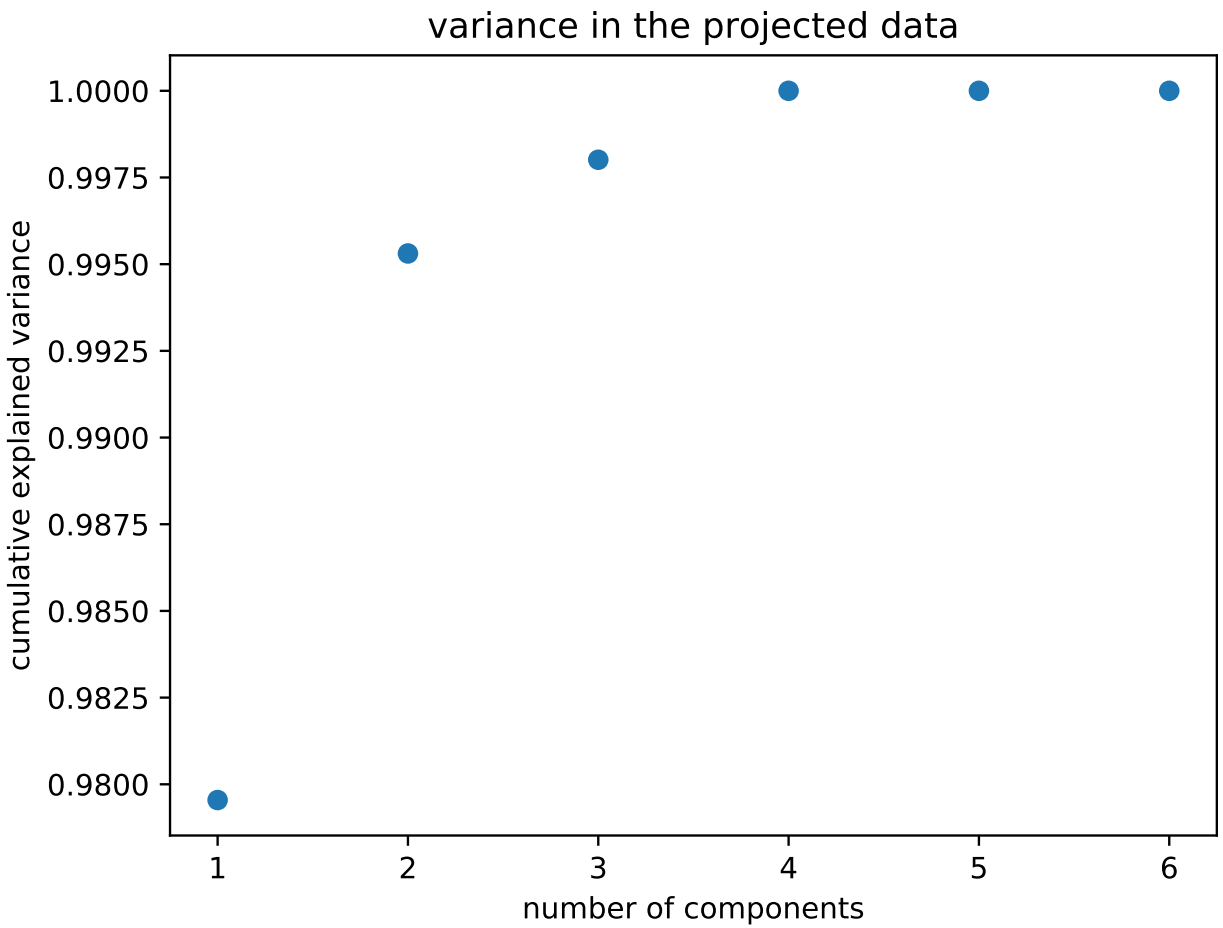


Figure – Variance of the projected data as a function of the number of components (digits dataset)

Number of components

Exercice 3 : What happens with this dataset ?



Number of components

Conclusion : PCA can help detemine whether some components carry no information in the data.

Shortcomings of PCA

PCA is sensitive to :

- outliers
- initial data scaling

Context

We are given a set of observations $\{y_1, \ldots, y_n\} \in \mathcal{Y}$ that we assume are generated i.i.d from an unknown distribution. We look for a **probabilistic model** that explains well the data. We could for instance use this model to generate new data, that would be statistically similar to the observed ones.

Density estimation

We will consider **parametric models** for density estimation.

Definition

Parametric model

Let $\Theta \subset \mathbb{R}^p$. A parametric model $\mathcal{P}$ is a set of probability distributions on $\mathcal{Y}$, indexed by elements of Θ.

$$\mathcal{P} = \{p_\theta | \theta \in \Theta\}$$

Examples :

- Bernoulli model (parameter p)
- Gaussian model (parameter (μ, σ))
- Binomial model (parameter (n, p))

Objective

If we assume that the data were generated from some $p_{\theta^*} \in \mathcal{P}$, with a unknown parameter θ^*, our goal if to find a good estimation of θ. If the data are indeed generated by a distribution in $\mathcal{P}$, the problem is said to be **well specified**. Otherwise, the problem is said to be **misspecified**.

Definition

Likelihood

Let $\mathcal{P} = \{p_\theta, \theta \in \Theta\}$ be a parametric model.
Given $y \in \mathcal{Y}$, the **likelihood** is the function :

$$\theta \mapsto L(\theta|y) = p_\theta(y) \tag{16}$$

Given $D_n = (y_1, \ldots, y_n)$, the likelihood $L(.|D_n)$ is the function :

$$\theta \mapsto L(\theta|D_n) = \prod_{i=1}^{n} p_\theta(y_i) \tag{17}$$

The **maximum likelihood estimator** (MLE) is the parameter θ that maximises the likelihood :

$$\hat{\theta}_n \in \arg\max_{\theta \in \Theta}(L(\theta|D_n)) \tag{18}$$

Remarks

- Since the samples y_i are assumed to be independent, the likelihood corresponds to the probability (or probability density) of observing the dataset according to p_θ.
- We often maximise the log of the likelihood, as it is easier to differentiate a sum. Since log is an increasing function, the MLE is also the maximiser of the log of L.

Example 1

Exercice 4 : We observe the data $(1, 0)$. We model these data with a Bernoulli distribution of parameter p.

- What is the likelihood of these observations as a function of p ?
- What is the value $\hat{p}$ that maximizes this likelihood ?

Example 2

Exercice 5 : We observe the data $(1, 0, 1)$ (same hypotheses)

- What is the likelihood of these observations as a function of p ?
- What is the value $\hat{p}$ that maximizes this likelihood ?

Example 3

We observe the data $(2.5, 3.5)$. We assume that these data come from a normal law of parameters μ and σ.

$$\begin{aligned} L &= p(2.5|\mu,\sigma)p(3.5|\mu,\sigma) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{2.5-\mu}{\sigma})^2} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}(\frac{3.5-\mu}{\sigma})^2} \end{aligned} \tag{19}$$

We wan show that the likelihood is maximum for :

- $\hat{\mu} = \frac{2.5+3.5}{2}$
- $\hat{\sigma^2} = \frac{(2.5-\hat{\mu})^2+(3.5-\hat{\mu})^2}{2}$

ERM

In the context of density estimation, we can define a loss function as the **negative log-likelihood.**

$$\Theta \times \mathcal{Y} \mapsto -\log(p_\theta(y))$$

Given this loss, the risk writes :

$$R(\theta) = E_Y[-\log\left(p_\theta(y)\right)]$$

and the empirical risk :

$$R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(p_\theta(y_i)\right)]$$

The MLE is then also the empirical risk minimizer.

KL divergence

The Kullback-Leibler divergence is a quantity used to compare two probability distributions.

Definition

Kullback-Leibler divergence
Given two distributions p and q, the KL divergence from p to q is defined as :

$$KL(p||q) = E_{Y\sim p}\left[\log \frac{p(Y)}{q(Y)}\right]$$

Lemma

If the data are generated by p_{θ^*}, then $KL(p_{\theta^*}||p_\theta)$ is the excess risk of p_θ, with the legative log-likelihood loss.

Link with supervised learning methods

Probabilistic modelling can provide an interesting interpretation of several supervised learning methods, such as :

- logistic regression
- ordinary least squares

In a supervised learning context, we replace the likelihood $p_\theta(y)$ by a **conditional** likelihood $p_\theta(y|x)$ (conditional modelling).

$$R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(p_\theta(y_i|x_i)\right) \tag{20}$$

Link with logistic regression

We consider a binary classification problem, with $\mathcal{Y} = \{0, 1\}$.
Let us now consider the probabilistic model such that

$$p_\theta(1|x) = \sigma(\theta^T x)$$

Equivalently, this model can be written (remember that $y = 0$ or $y = 1$)

$$p_\theta(y|x) = \big(\sigma(\theta^T x)\big)^y \big(1 - \sigma(\theta^T x)\big)^{1-y} \tag{21}$$

Exercice 6 : Show that the parameter θ with maximum likelihood is the logistic regression estimator θ_{logit} (cross entropy version).

General classification problem

- $\mathcal{X} = \mathbb{R}^d$
- $\mathcal{Y} = \{-1, 1\}$ or $\mathcal{Y} = \{0, 1\}$.
- $l(y, z) = 1_{y \neq z}$ ("0-1" loss)
- $F = \mathcal{Y}^{\mathcal{X}}$

Problem

Optimizing on $F = \mathcal{Y}^{\mathcal{X}}$ is equivalent to optimizing in the set of subsets of $\mathcal{X}$.
We cannot differentiate on this hypothesis space and it is not clear how to regularize.

Subsets

Exercice 7 : **Combinatorial problem**

If we wanted to try all applications in $\mathcal{Y}^{\mathcal{X}}$, if $|\mathcal{X}| = n$, how many applications would there be ?

Real-valued function

Instead of an application in $\mathcal{Y}^{\mathcal{X}}$, we will learn $g : \mathcal{X} \to \mathbb{R}$ and define $f(x) = \text{sign}(g(x))$ with

$$\text{sign}(x) = \begin{cases} 1 \text{ if } x \geq 0 \\ -1 \text{ if } x < 0 \end{cases}$$

Risk

The risk (generalization error) of $f = \text{sign} \circ g$ is.

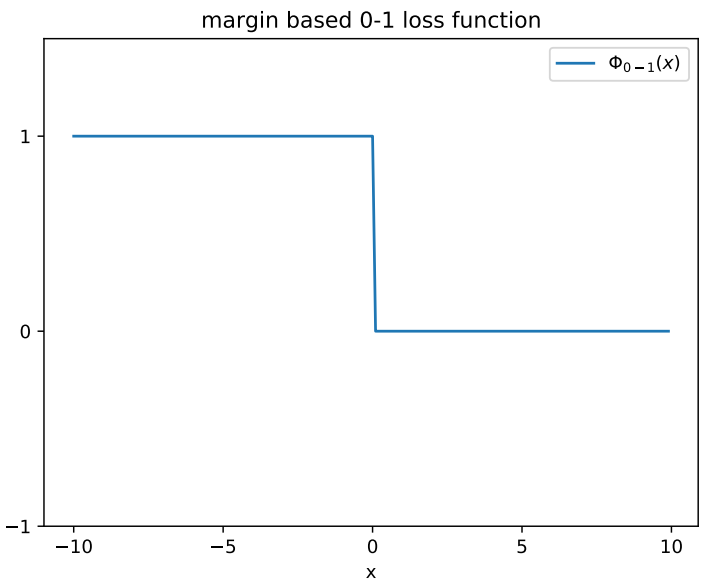
$$\begin{aligned} R(g) &= P(\text{sign}(g(x)) \neq y) \\ &= E\left[1_{\text{sign}(g(x)) \neq y}\right] \\ &= E\left[1_{yg(x)<0}\right] \end{aligned} \tag{22}$$

Remark : several solutions

If f^* is the Bayes predictor, there might be many optimal functions g, i.e : such that $\text{sign}(g(x)) = f^*(x)$.

Margin based 0-1 loss function Φ_{0-1}

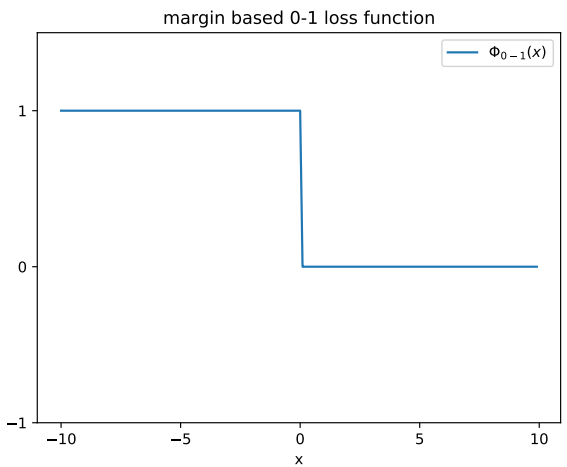
$$\begin{aligned} R(g) &= E\left[1_{\text{sign}(g(x))\neq y}\right] \\ &= E\left[1_{yg(x)<0}\right] \\ &= E\left[\Phi_{0-1}(yg(x))\right] \end{aligned} \tag{23}$$



Empirical risk minimization

The corresponding empirical risk writes :

$$\frac{1}{n}\sum_{i=1}^{n}\Phi_{0-1}(y_i g(x_i)) \tag{24}$$

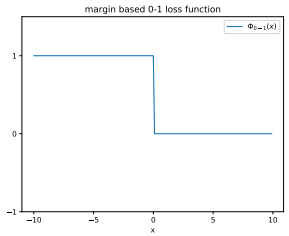


What is the issue with this objective function ?

Empirical risk minimization

The corresponding empirical risk writes :

$$\frac{1}{n}\sum_{i=1}^{n}\Phi_{0-1}(y_i g(x_i)) \tag{25}$$



What is the issue with this objective function ?

- non-convex
- the differential is not defined or non informative.

Convex surrogate

Key idea : replace Φ_{0-1} by another function Φ that is easier to optimize (convexity) but still represents the correctness of the classification.

Definition

The Φ-risk is defined as

$$R_\Phi(g) = E\Big[\Phi(yg(x))\Big] \tag{26}$$

The empirical Φ-risk is defined as

$$R_{\Phi,n}(g) = \frac{1}{n}\sum_{i=1}^{n}\Phi(y_i g(x_i)) \tag{27}$$

Most common convex surrogates

Definition

Logistic loss

$$\Phi(u) = \log(1 + e^{-u}) \tag{28}$$

With linear predictors, this loss will lead to **logistic regression** (which is classification despite its name).

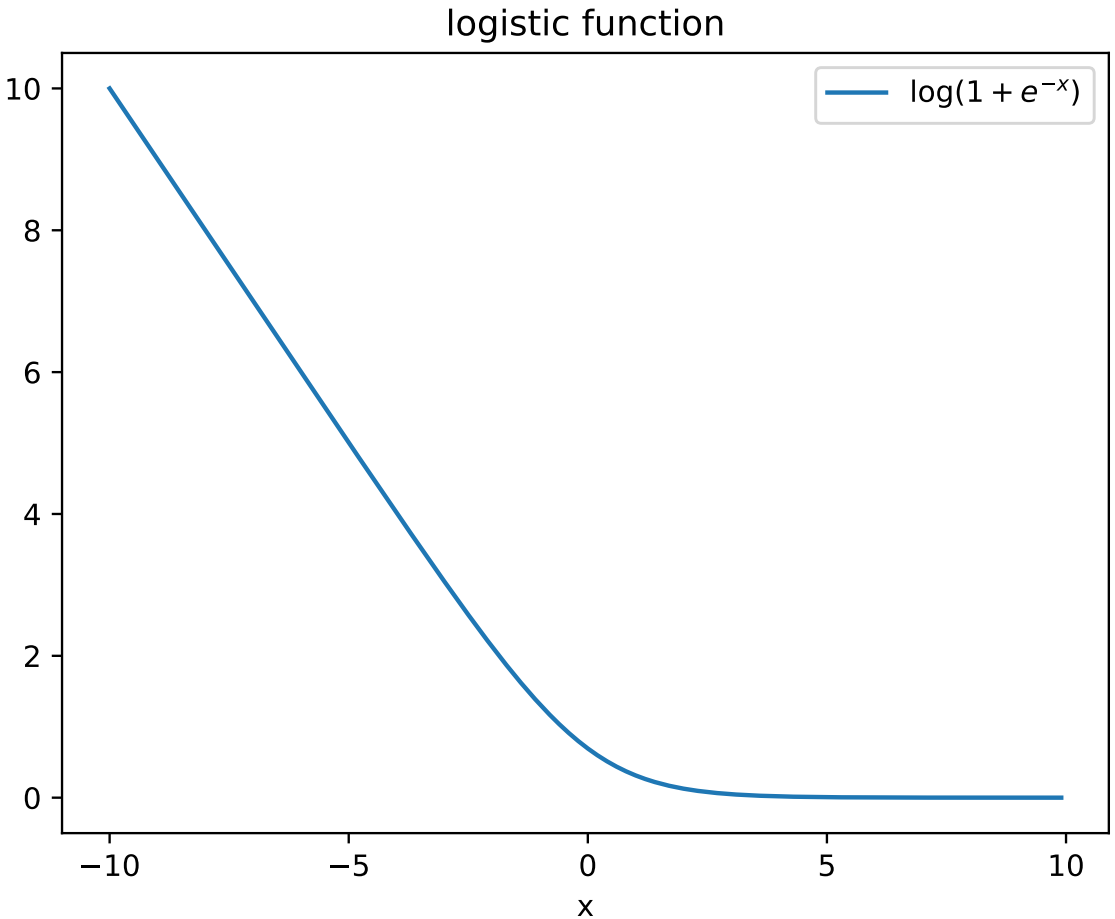
Most common convex surrogates

If $\mathcal{Y} = \{0, 1\}$, $\hat{y}$ is the prediction and y is the correct label, then we write :

$$l(\hat{y}, y) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}}) \qquad (29)$$

(cross entropy loss)

Logistic function



Most common convex surrogates

Definition

Hinge loss

$$\Phi(u) = max(1 - u, 0) \tag{30}$$

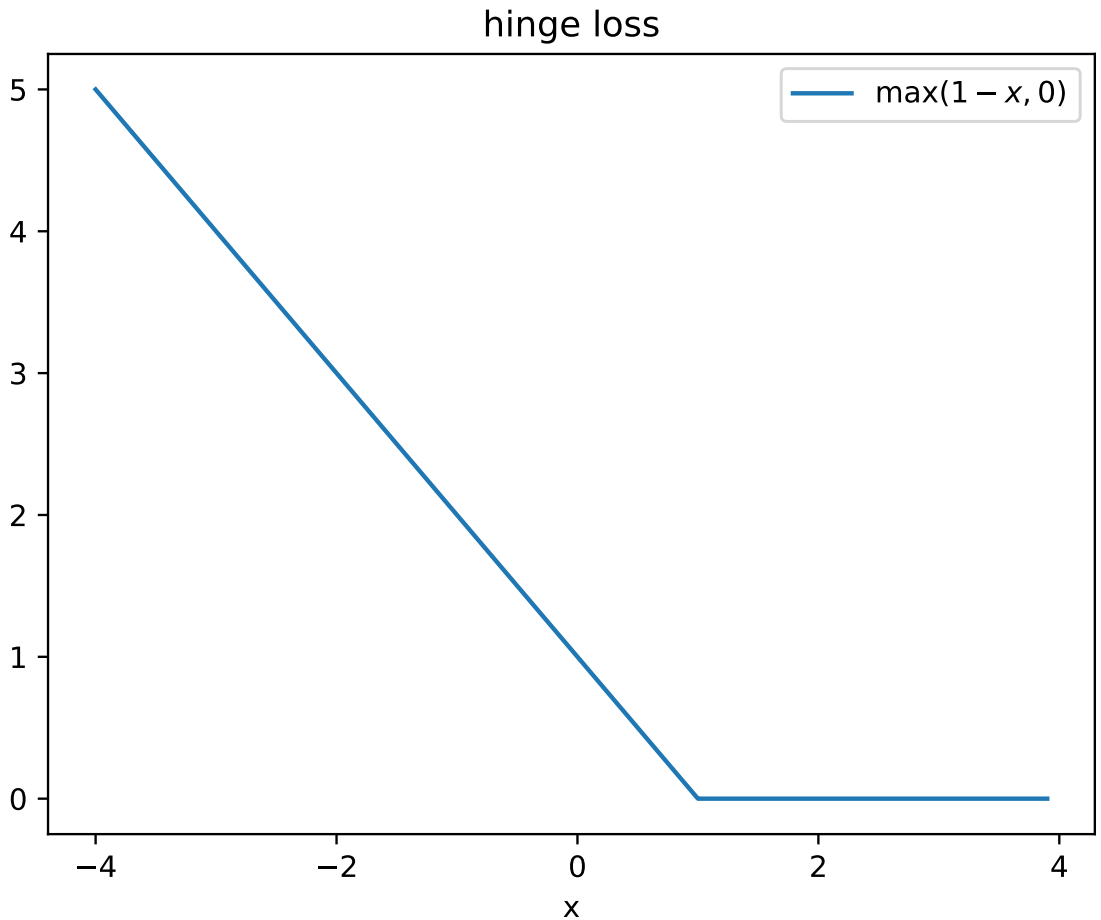
With linear predictors, this loss will lead to **Support vector machines**.

Definition

Squared hinge loss

$$\Phi(u) = (max(1 - u, 0))^2 \tag{31}$$

Hinge loss

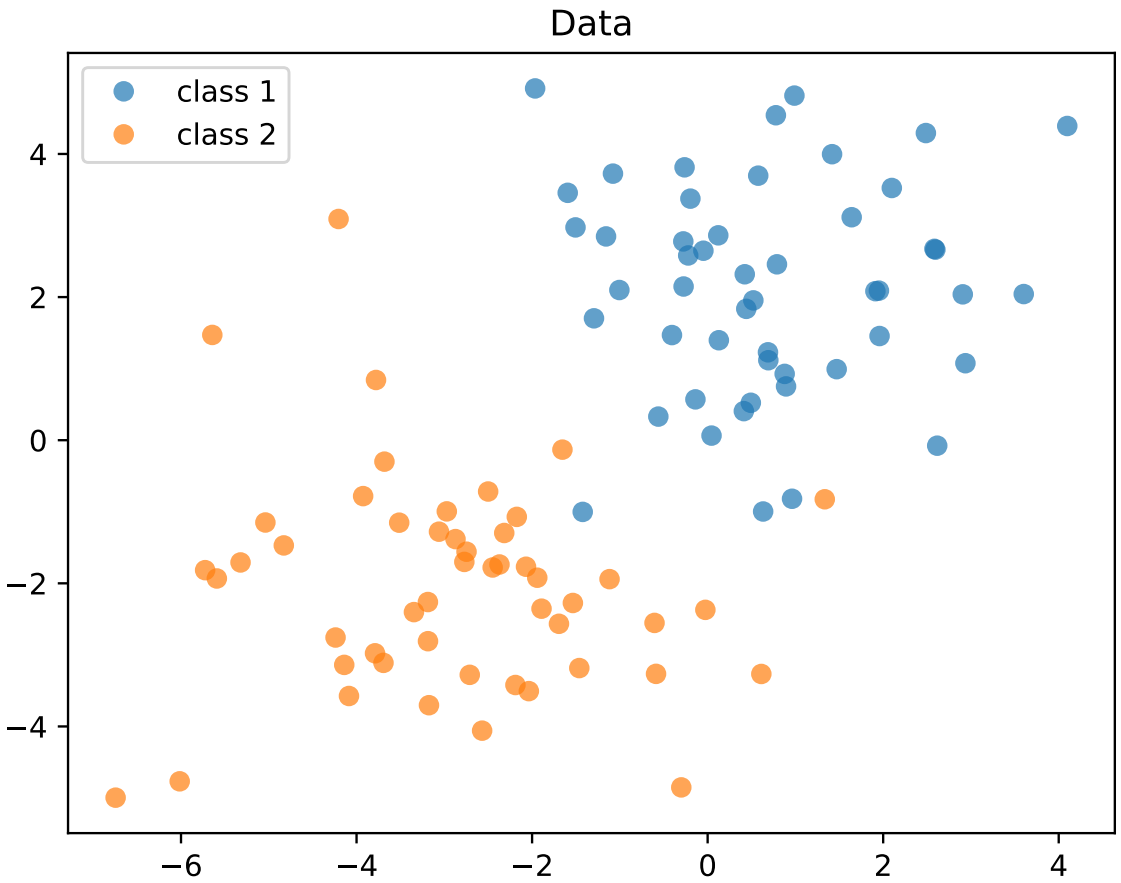


Under some technical hypotheses, minimizing the empirical Φ risk leads to a good generalization errror for the "0-1" loss (notion of calibration function).

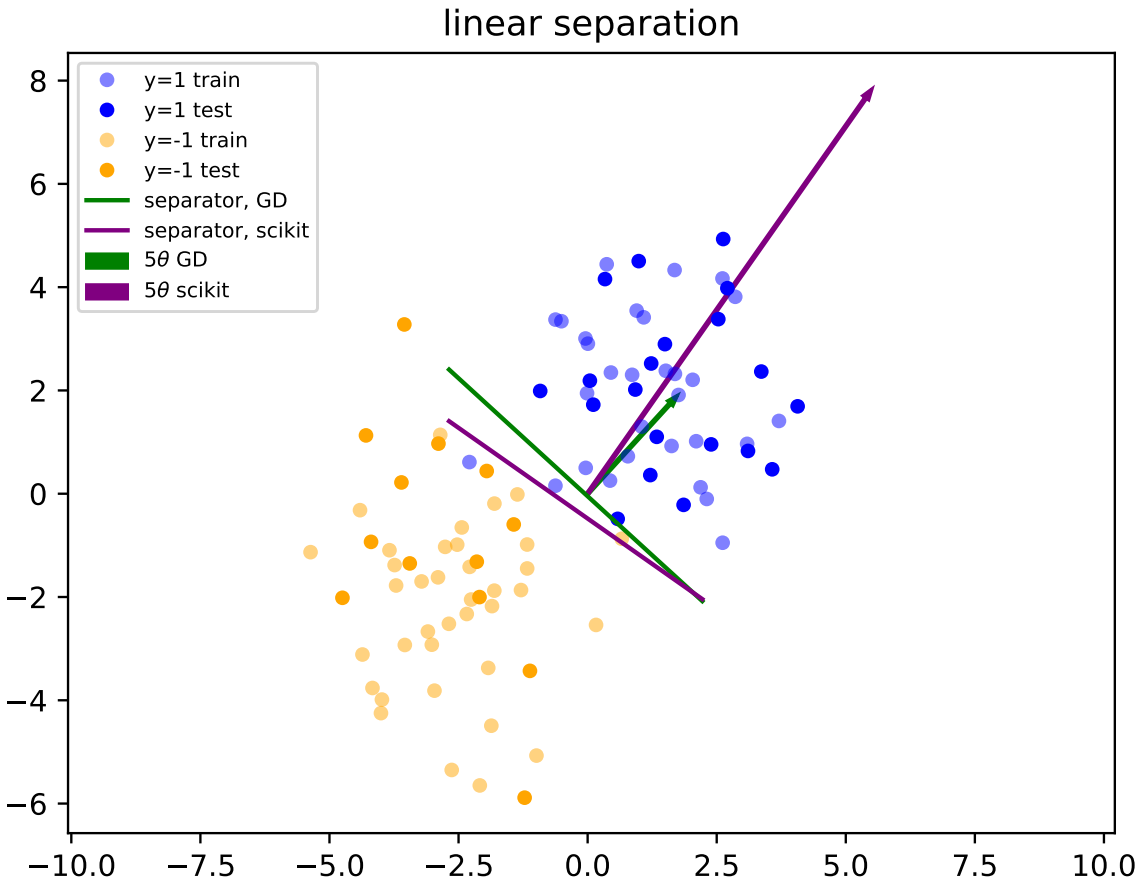
Logistic regression

- $g(x) = \langle x, \theta \rangle = x^T \theta$.
- $f(x) = \text{sign}(\langle x^T \theta \rangle)$
- It can be seen as "linear regression applied to classification".

Data to separate



Data to separate



Logistic regression

In this section we use the setting $\mathcal{Y} = \{0, 1\}$.

- prediction : $\hat{y} = x^T \theta$
- surrogate loss : cross-entropy loss.

$$l(\hat{y}, y) = y \log(1 + e^{-\hat{y}}) + (1 - y) \log(1 + e^{\hat{y}}) \qquad (32)$$

Logistic regression estimator

If l is the logistic loss, it is defined as

$$\hat{\theta}_{logit} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} l(x_i^T \theta, y_i) \tag{33}$$

We will show that this empirical risk is convex as a function of θ (exercises).

No closed-form solution

Since the loss is convex, to minimize it is sufficient to look for the cancellation of the gradient. However, the corresponding equation has no closed-form solution.
We thus need to use iterative algorithms in order to find a minimizer (e.g. : gradient descent, Newton's method, etc)

Practical usage of logistic regression

In practice, it is common practice to :

- regularize the logistic loss to avoid overfitting, for instance with a $L2$ penalty (as in ridge regression)
- use feature maps and classify with $\phi(x)$ instead of x.

References I


Bermejo, J. M., Ramos, A. A., and Prieto, C. A. (2013).
Astrophysics A PCA approach to stellar effective temperatures.
95 :1–9.