Data Science for Geosciences

Classification

Filière SICOM, 3A

Classification problem

Variable terminology

- observed data referred to as *input* variables, *predictors* or *features* $\leftarrow$ usually denoted as X
- data to predict referred to as *output* variables, or *responses* $\leftarrow$ usually denoted as Y

Type of prediction problem : regression vs classification

Depending on the type of the *output* variables

- when Y are quantitative data (continuous variables, e.g. electrical load curve values) $\leftarrow$ regression
- when Y are categorical data (discrete qualitative variables, e.g. handwritten digits $Y \in \{0, \ldots, 9\}$) $\leftarrow$ classification

Classification outline

- Model based approaches for classification
  - Linear/Quadratic Discriminant Analysis
- Black box approaches for classification
  - K nearest neighbors ?
  - Support Vector Machine
- Clustering ?
  - K means ?
  - EM algorithm ?

Generative models

Two kinds of approaches based on a model :

1. Discriminative approaches : direct learning of $p(Y|X)$,
   e.g. Regression, logistic regression
2. Generative models : learning of the joint distribution $p(X,Y)$

$$p(X,Y) = \underbrace{p(X|Y)}_{\text{likelihood}}\underbrace{\Pr(Y)}_{\text{prior}},$$

e.g. linear/quadratic discriminant analysis, Naïve Bayes

Bayes classifier

- $Y \in \mathcal{Y} \quad \leftarrow$ discrete domain

Definition

The Bayes classification rule f^* is defined as

$$f^*(x) = \arg\max_{k \in \mathcal{Y}} \Pr(Y = k | X = x).$$

The associated misclassification error rate $\mathcal{E}[f^*] = \Pr\left(f^*(x) \neq Y\right)$ is refered to as the Bayesian error rate

Theorem

The Bayes classification rule f^* is optimal in the misclassification rate sense : for any rule f, $\mathcal{E}[f] \geq \mathcal{E}[f^*]$.

Remarks

- $f^*(X) \equiv$ *maximum a posteriori* (MAP) estimate
- In real-word applications, the distribution of (X, Y) is unknown $\Rightarrow$ no analytical expression of $f^*(X)$. But useful reference on academic examples.

Discriminant functions

For both model based approaches, Bayes classifier is defined as

$$f^*(x) = \arg\max_{k \in \mathcal{Y}} \Pr(Y = k | X = x)$$

- equivalent to consider a set of functions $\delta_k(x)$, for $k \in \mathcal{Y}$, derived from a monotone transformation of posterior probability $\Pr(Y = k | X = x)$
- decision boundary between classes k and l is then defined as the set $\{x \in \mathcal{X} \ : \ \delta_k(x) = \delta_l(x)\}$

Definition

$\delta_k(x)$ are called the discriminant functions of each class k

☞ x is predicted in the k_0 class such that $k_0 = \arg\max_{k \in \mathcal{Y}} \delta_k(x)$

Generative models : Estimation problem

Assumptions

- classification problem with K classes : $Y \in \mathcal{Y} = \{1, \dots, K\}$,
- input variables : $X \in \mathbb{R}^p$

Bayes rule :

$$\Pr(Y = k | X = x) = \frac{p(x|Y = k) \Pr(Y = k)}{p(x)} = \frac{p(x|Y = k) \Pr(Y = k)}{\sum_{j=1}^{K} p(x|Y = j) \Pr(Y = j)}.$$

In practice, the following quantities are unknown :

- densities of each class $p_k(x) \equiv p(x|Y = k)$
- weights, or prior probabilities, of each class $\pi_k \equiv \Pr(Y = k)$

Estimation problem

These quantities must be learned on a training set :

learning problem $\Leftrightarrow$ estimation problem in a parametric or not way

Quadratic Discriminant Analysis (QDA)

Supervised classification assumptions

- $X \in \mathbb{R}^p$, $Y \in \mathcal{Y} = \{1, \ldots, K\}$,
- sized n training set $(X_1, Y_1), \ldots (X_n, Y_n)$

QDA Assumptions

The input variables X, given a class $Y = k$, are distributed according to a parametric and Gaussian distribution :

$$X|Y = k \ \sim \ \mathcal{N}(\mu_k, \Sigma_k) \ \Leftrightarrow \ p_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)}$$

The Gaussian parameters are, for each class $k = 1, \ldots, K$

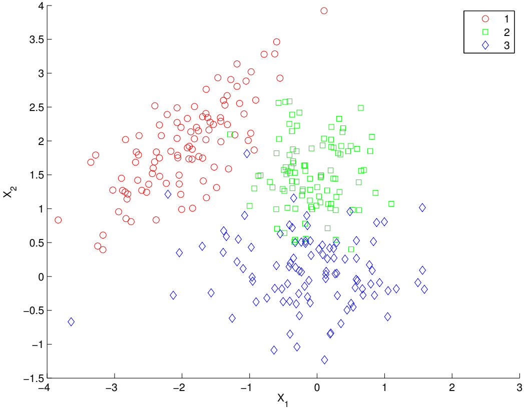
- mean vectors $\mu_k \in \mathbb{R}^p$,
- covariance matrices $\Sigma_k \in \mathbb{R}^{p \times p}$,

☞ set of parameters $\theta_k \equiv \{\mu_k, \Sigma_k\}$, plus the weights π_k, for $k = 1, \ldots, K$.

Example

Mixture of $K = 3$ Gaussians

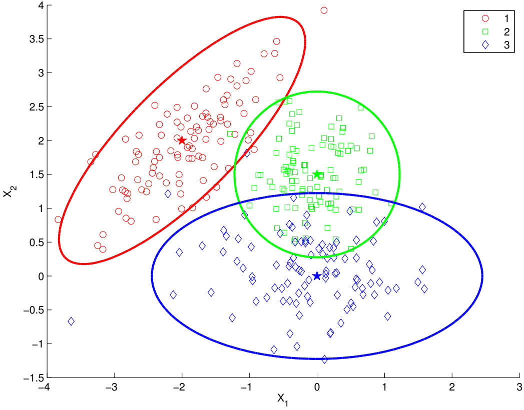
- $Y \in \{1, 2, 3\}$
- $X \in \mathbb{R}^2$



Example

Mixture of $K = 3$ Gaussians

- $Y \in \{1, 2, 3\}$
- $X \in \mathbb{R}^2$



95% theoretical confidence regions

QDA parameter estimation

Log-likelihood

For the training set,

$$
\begin{aligned}
\ell(\theta_1, \dots, \theta_K, \pi_1, \dots, \pi_{K-1}) &= \log p\left((x_1, y_1), \dots, (x_n, y_n)\right), \\
&= \sum_{i=1}^{n} \log p\left((x_i, y_i)\right), \quad \leftarrow \text{ i.i.d. training set}, \\
&= \sum_{i=1}^{n} \log \left[p\left(x_i | y_i\right) \Pr\left(y_i\right)\right], \\
&= \sum_{i=1}^{n} \log \left[\pi_{y_i}\ p_{y_i}\left(x_i; \theta_{y_i}\right)\right].
\end{aligned}
$$

Rk : $\pi_K = 1 - \sum_{j=1}^{K-1} \pi_j$ is not a parameter

QDA parameter estimation (Cont'd)

Notations

- $n_k = \#\{y_i = k\}$ is the number of training samples in class k,
- $\sum_{y_i=k}$ is the sum over all the indices i of the training samples in class k

(Unbiased) Maximum likelihood estimators (MLE)

- $\widehat{\pi}_k = \dfrac{n_k}{n}$, $\leftarrow$ sample proportion
- $\widehat{\mu}_k = \dfrac{\sum_{y_i=k} x_i}{n_k}$, $\leftarrow$ sample mean
- $\widehat{\Sigma}_k = \frac{1}{n_k-1} \sum_{y_i=k} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^T$, $\leftarrow$ sample covariance

Rk : $\frac{1}{n_k-1}$ is a bias correction factor for the covariance MLE (otherwise $\frac{1}{n_k}$)

QDA decision rule

The classification rule becomes

$$\begin{aligned} f(x) &= \arg\max_{k \in \mathcal{Y}} \Pr(Y = k | X = x, , \widehat{\theta}, \widehat{\pi}), \\ &= \arg\max_{k \in \mathcal{Y}} \underbrace{\log \Pr(Y = k | X = x, \widehat{\theta}, \widehat{\pi})}_{\delta_k(x)}, \end{aligned}$$

where

$$\delta_k(x) = -\frac{1}{2} \log \left| \widehat{\Sigma}_k \right| - \frac{1}{2} (x - \widehat{\mu}_k)^T \widehat{\Sigma}_k^{-1} (x - \widehat{\mu}_k) + \log \widehat{\pi}_k + \cancel{Cst},$$

is the discriminant function

Remarks

1. different rule than the Bayes classifier as θ replaced by $\widehat{\theta}$ (and π replaced by $\widehat{\pi}$)
2. when $n \gg p$, $\widehat{\theta} \to \theta$ (and $\widehat{\pi} \to \pi$) : convergence to the optimal classifier if the Gaussian model is correct...

QDA decision boundary

The boundary between two classes k and l is described by the equation

$$\delta_k(x) = \delta_l(x) \Leftrightarrow C_{k,l} + L_{k,l}^T x + x^T Q_{k,l}^T x = 0, \quad \leftarrow \text{quadratic equation}$$

where

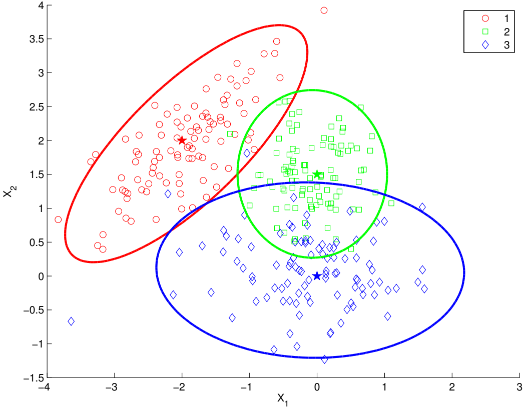
- $C_{k,l} = -\frac{1}{2} \log \frac{|\widehat{\Sigma}_k|}{|\widehat{\Sigma}_l|} + \log \frac{\widehat{\pi}_k}{\widehat{\pi}_l} - \frac{1}{2} \widehat{\mu}_k^T \widehat{\Sigma}_k^{-1} \widehat{\mu}_k + \frac{1}{2} \widehat{\mu}_l^T \widehat{\Sigma}_l^{-1} \widehat{\mu}_l, \quad \leftarrow$ scalar
- $L_{k,l} = \widehat{\Sigma}_k^{-1} \widehat{\mu}_k - \widehat{\Sigma}_l^{-1} \widehat{\mu}_l, \quad \leftarrow$ vector in $\mathbb{R}^p$
- $Q_{k,l} = \frac{1}{2} \left(-\widehat{\Sigma}_k^{-1} + \widehat{\Sigma}_l^{-1} \right), \quad \leftarrow$ matrix in $\mathbb{R}^{p \times p}$

☞ Quadratic discriminant analysis

QDA example

Mixture of $K = 3$ Gaussians

- Estimation of the parameters $\hat{\mu}_k$, $\hat{\Sigma}_k$ and $\hat{\pi}_k$, for $k = 1, 2, 3$

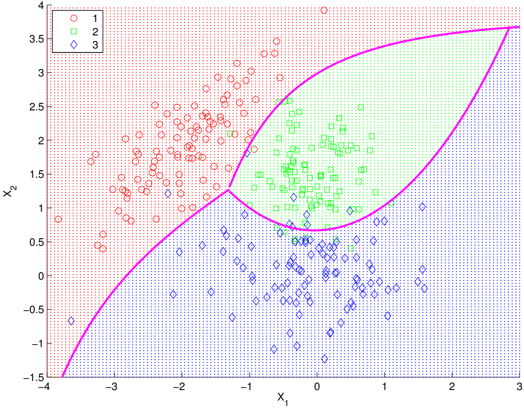


95% estimated confidence regions

QDA example (Cont'd)

Mixture of $K = 3$ Gaussians

- Classification rule : $\arg\max_{k=1,2,3} \delta_k(x)$
- Quadratic boundaries $\{x; \delta_k(x) = \delta_l(x)\}$



LDA principle

LDA Assumptions

Additional simplifying assumption w.r.t. QDA : all the class covariance matrices are identical ("homoscedasticity"), i.e. $\Sigma_k = \Sigma$, for $k = 1, \ldots, K$

(Unbiased) Maximum likelihood estimators (MLE)

- $\widehat{\pi}_k$ and $\widehat{\mu}_k$ are unchanged,
- $\widehat{\Sigma} = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{y_i=k} (x_i - \widehat{\mu}_k)(x_i - \widehat{\mu}_k)^T$, $\leftarrow$ pooled covariance

Rk : $\frac{1}{n-K}$ is a bias correction factor for the covariance MLE (otherwise $\frac{1}{n}$)

LDA discriminant function

$$\delta_k(x) = -\frac{1}{2} \log \left| \widehat{\Sigma} \right| - \frac{1}{2} (x - \widehat{\mu}_k)^T \widehat{\Sigma}^{-1} (x - \widehat{\mu}_k) + \log \widehat{\pi}_k + \cancel{Cst},$$

LDA decision boundary

The boundary between two classes k and l reduces to the equation

$$\delta_k(x) = \delta_l(x) \Leftrightarrow C_{k,l} + L_{k,l}^T x = 0, \quad \leftarrow \text{linear equation}$$

where

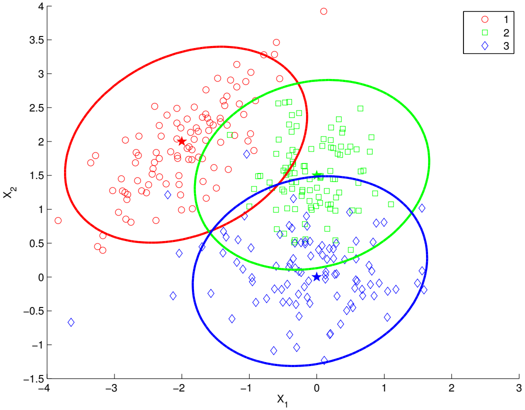
- $C_{k,l} = \log \frac{\widehat{\pi}_k}{\widehat{\pi}_l} - \frac{1}{2}\widehat{\mu}_k^T \widehat{\Sigma}^{-1} \widehat{\mu}_k + \frac{1}{2}\widehat{\mu}_l^T \widehat{\Sigma}^{-1} \widehat{\mu}_l, \quad \leftarrow$ scalar
- $L_{k,l} = \widehat{\Sigma}^{-1}\left(\widehat{\mu}_k - \widehat{\mu}_l\right), \quad \leftarrow$ vector in $\mathbb{R}^p$
- $Q_{k,l} = 0$,

☞ Linear discriminant analysis

Linear Discriminant Analysis (LDA)

Mixture of $K = 3$ Gaussians

- Estimation of the parameters $\hat{\mu}_k$, $\hat{\pi}_k$, for $k = 1, 2, 3$, and $\hat{\Sigma}$

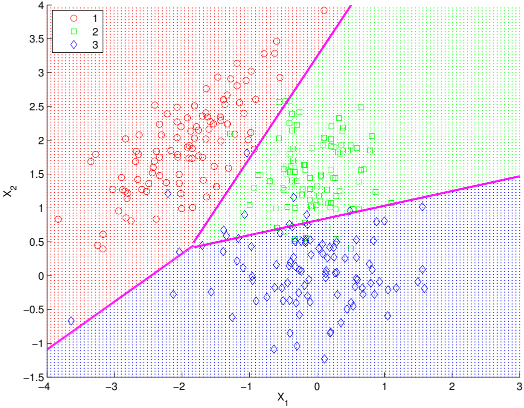


95% estimated confidence regions

Linear Discriminant Analysis (LDA)

Mixture of $K = 3$ Gaussians

- Classification rule : $\arg\max_{k=1,2,3} \delta_k(x)$
- linear boundaries $\{x; \delta_k(x) = \delta_l(x)\}$



Complexity of discriminant analysis methods

Effective number of parameters

- LDA : $(K-1) \times (p+1) = O(Kp)$
- QDA : $(K-1) \times \left(\frac{p(p+3)}{2} + 1\right) = O(Kp^2)$

Remarks

- in high dimension, i.e. $p \approx n$ or $p > n$, LDA is more stable than QDA which is more prone to overfitting,
- both methods appear however to be robust on a large number of real-word datasets
- LDA can be viewed in some cases as a least squares regression method
- LDA performs a dimension reduction to a subspace of dimension $\leq K-1$ generated by the vectors $z_k = \Sigma^{-1}\widehat{\mu}_k \leftarrow$ dimension reduction from p to $K-1$!

Conclusions

Generative models

- learning/estimation of $p(X,Y) = p(X|Y)\,\Pr(Y)$,
- derivation of $\Pr(Y|X)$ from Bayes rule,

Different assumptions on the class densities $p_k(x) = p(X = x|Y = k)$

- QDA/LDA : Gaussian parametric model

☞ performs well on many real-word datasets

☞ LDA is especially useful when n is small

Perspectives

Black box approaches : direct learning of the prediction rule f

Support Vector Machine (SVM)

Theory elaborated in the early 1990's (Vapnik *et al*) based on the idea of 'maximum margin'

- deterministic criterion learned on the training set ← supervised classification
- ☞ general, i.e. model free, linear classification rule
- ☞ classification rule is linear in a transformed space of higher (possible infinite) dimension than the original input feature/predictor space

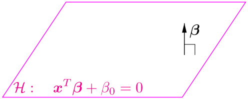
Linear discrimination and Separating hyperplane

Binary classification problem

- $X \in \mathbb{R}^p$
- $Y \in \{-1, 1\} \leftarrow 2$ classes
- Training set (x_i, y_i), for $i = 1, \ldots, n$

Defining a linear discriminant function $h(x) \Leftrightarrow$ defining a separating hyperplane $\mathcal{H}$ with equation

$$\boldsymbol{x}^T \boldsymbol{\beta} + \beta_0 = 0,$$



- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the normal vector (vector normal to the hyperplane $\mathcal{H}$),
- $\beta_0 \in \mathbb{R}$ is the intercept/offset (regression or geometrical interpretation)

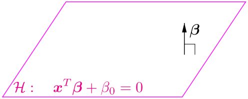
☞ $\mathcal{H}$ is an *affine subspace* of codimension 1

☞ $h(x) \equiv \boldsymbol{x}^T \boldsymbol{\beta} + \beta_0$ is the associated (linear) discriminant function

Separating hyperplane and prediction rule

For a given separating hyperplane $\mathcal{H}$ with equation

$$\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0,$$



the prediction rule can be expressed as

- $\widehat{y} = +1$, if $h(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 \geq 0$,
- $\widehat{y} = -1$, otherwise,

or in an equivalent way :

$$\widehat{y} \equiv G(\boldsymbol{x}) = \text{sign}\left[\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0\right]$$

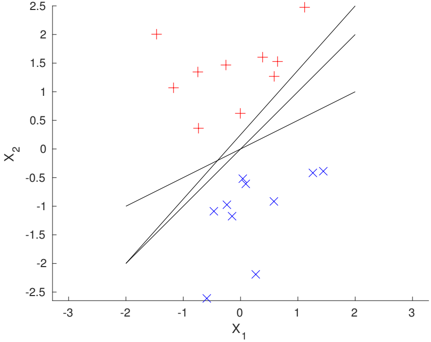
Rk : $\boldsymbol{x}$ is in class $y \in \{-1, 1\}$: prediction $G(\boldsymbol{x})$ is correct iff $y\left(\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0\right) \geq 0$

Separating Hyperplane : separable case

Linear separability assumption : $\exists \boldsymbol{\beta} \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$ s.t. the hyperplane $\boldsymbol{x}^T \boldsymbol{\beta} + \beta_0 = 0$ perfectly separates the two classes on the training set :

$$y_k \left(x_k^T \boldsymbol{\beta} + \beta_0 \right) \geq 0, \quad \text{for } k = 1, \ldots, n,$$

Separable case ($p = 2$ example)



Pb : infinitely many possible perfect separating hyperplanes $\boldsymbol{x}^T \boldsymbol{\beta} + \beta_0 = 0$

☞ Find the 'optimal' separating hyperplane

Maximum margin separating hyperplane (separable case)

Distance of a point $\boldsymbol{x_k}$ to an hyperplane $\mathcal{H}$ s.t. $\boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0$,

$$d(x_k, \mathcal{H}) \equiv \min_{\boldsymbol{x}} \left\{ \|\boldsymbol{x} - \boldsymbol{x}_k\| \ : \ \boldsymbol{x}^T\boldsymbol{\beta} + \beta_0 = 0 \right\}$$

Maximum margin principle

We are interested in the 'optimal' perfect separating hyperplane maximizing the distance $M > 0$, called the margin, between the samples of each class and the separating hyperplane

$\Rightarrow$ Find $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\beta_0 \in \mathbb{R}$ s.t. the margin

$$M = \min_{1 \leq k \leq n} \{d(x_k, \mathcal{H})\}$$

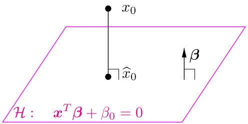
is maximized

Signed distance

From the orthogonality principle,

$$d(x_0, \mathcal{H}) = \|\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0\|\,,$$

where $\widehat{\boldsymbol{x}}_0$ is the orthogonal projection of $\boldsymbol{x}_0$ on $\mathcal{H}$



$\widehat{\boldsymbol{x}}_0 \in \mathcal{H} \Rightarrow -\widehat{\boldsymbol{x}}_0^T\boldsymbol{\beta} = \beta_0$

$\Rightarrow$ $\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0$ and $\boldsymbol{\beta}$ are collinear,

$\Rightarrow$ $\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0 = \underbrace{\langle \boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0, \boldsymbol{\beta}^* \rangle}_{\text{signed distance}} \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^* = \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|}$,

$$\Rightarrow \text{signed distance } = (\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0)^T \frac{\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} = \frac{\boldsymbol{x}_0^T\boldsymbol{\beta} - \widehat{\boldsymbol{x}}_0^T\boldsymbol{\beta}}{\|\boldsymbol{\beta}\|} = \frac{\boldsymbol{x}_0^T\boldsymbol{\beta} + \beta_0}{\|\boldsymbol{\beta}\|},$$

Remarks

- $|\langle \boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0, \boldsymbol{\beta}^* \rangle| = \|\boldsymbol{x}_0 - \widehat{\boldsymbol{x}}_0\| = d(\boldsymbol{x}_0, \mathcal{H}) \leftarrow$ “signed distance”
- for any perfect separating hyperplane $y_k\langle \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_k, \boldsymbol{\beta}^* \rangle = \frac{1}{\|\boldsymbol{\beta}\|} y_k(\boldsymbol{x}_k^T\boldsymbol{\beta} + \beta_0) \geq 0$, for $k = 1, \ldots, n$,

Canonical separating hyperplane

For any perfect separating hyperplane, for $k = 1, \ldots, n$

$$y_k \langle \boldsymbol{x}_k - \widehat{\boldsymbol{x}}_k, \boldsymbol{\beta}^* \rangle = d(x_k, \mathcal{H})$$

Hence, the margin reads

$$M \equiv \min_{1 \leq k \leq n} \{d(x_k, \mathcal{H})\} = \frac{1}{\|\boldsymbol{\beta}\|} \min_{1 \leq k \leq n} \left\{ y_k(\boldsymbol{x}_k^T \boldsymbol{\beta} + \beta_0) \right\}$$

Remarks

- The bound M is reached (min of a countable set),
- ☞ the samples at the margin are denoted as $\boldsymbol{x}_{\text{margin}}$

Canonical expression of the separating hyperplane

$\boldsymbol{\beta}$ and β_0 are normalized s.t.

$$y_{\text{margin}}(\boldsymbol{x}_{\text{margin}}^T \boldsymbol{\beta} + \beta_0) = 1, \quad \text{thus } M = \frac{1}{\|\boldsymbol{\beta}\|}$$

Primal problem (separable case)

Canonical hyperplane expression :

$$\text{maximizing the margin } M = \frac{1}{\|\boldsymbol{\beta}\|} \quad \Leftrightarrow \quad \text{minimizing} \quad \|\boldsymbol{\beta}\|$$
$$\Leftrightarrow \quad \text{minimizing} \quad \frac{1}{2}\|\boldsymbol{\beta}\|^2$$

Primal optimization problem

$$\begin{cases} \min_{\boldsymbol{\beta},\beta_0} & \frac{1}{2}\|\boldsymbol{\beta}\|^2, \\ \text{subject to} & y_k\left(\boldsymbol{x}_k^T\boldsymbol{\beta} + \beta_0\right) \geq 1, \ \text{for } 1 \leq k \leq n. \end{cases}$$

- quadratic criterion + linear inequality constraints

☞ convex optimization problem

Lagrangian (separable case)

Convex constraints of positivity $\Rightarrow$ introduction of the Lagrange multipliers

Lagrangian

$$L(\boldsymbol{\beta}, \beta_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\beta}\|^2 - \sum_{i=1}^{n} \alpha_i \underbrace{\left[y_i(\boldsymbol{x}_i^T \boldsymbol{\beta} + \beta_0) - 1\right]}_{\geq 0},$$

where α_i are the Lagrange multipliers

First order Kuhn-Tucker necessary conditions

Setting the partial derivatives w.r.t. $\boldsymbol{\beta}$ and β_0 to zero yields

$$\begin{cases} \widehat{\boldsymbol{\beta}} & = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i, \\ 0 & = \sum_{i=1}^{n} \alpha_i y_i, \end{cases}$$

- plugging these expression in the Lagrangian yields the dual expression

Dual problem (separable case)

Dual optimization problem

$$\begin{cases} \max_{\boldsymbol{\alpha}} & \widetilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j, \\ \text{subject to} & \alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0. \end{cases}$$

- ☞ simple convex optimization problem for which standard numerical procedure are available
- ☞ calculation of the optimum multipliers $\widehat{\alpha}_i$

Support vectors and maximum margin hyperplane (separable case)

Complementary slackness Kuhn-Tucker necessary conditions

$$\widehat{\alpha}_i[y_i h(\boldsymbol{x}_i) - 1] = 0 \quad \Rightarrow \quad \widehat{\alpha}_i = 0 \text{ as } y_i h(\boldsymbol{x}_i) > 1$$

- since $\widehat{\boldsymbol{\beta}} = \sum_{i=1}^{n} \widehat{\alpha}_i y_i \boldsymbol{x}_i$, $\widehat{\boldsymbol{\beta}}$ depends only on the points at the margin $\leftarrow$ support vectors
- $\widehat{\beta}_0$ can be derived from the complementary slackness expression for any of support vectors $\boldsymbol{x}_{\text{margin}}$

$$\begin{aligned} y_{\text{margin}} h(\boldsymbol{x}_{\text{margin}}) - 1 = 0 \quad & \Rightarrow \quad \widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_{\text{margin}} + \widehat{\beta}_0 = y_{\text{margin}}, \\ & \Rightarrow \quad \widehat{\beta}_0 = -\widehat{\boldsymbol{\beta}}^T \boldsymbol{x}_{\text{margin}} + y_{\text{margin}} \end{aligned}$$

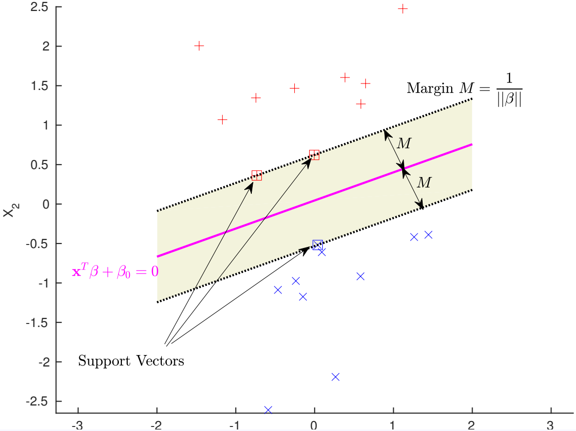
☞ the only inputs used to construct the maximum margin hyperplane are the support vectors and the discriminant function reads

$$h(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i y_i (\boldsymbol{x} - \boldsymbol{x}_{\text{margin}})^T \boldsymbol{x}_i + y_{\text{margin}}$$

Maximum margin separating hyperplane (separable case)

Separable case

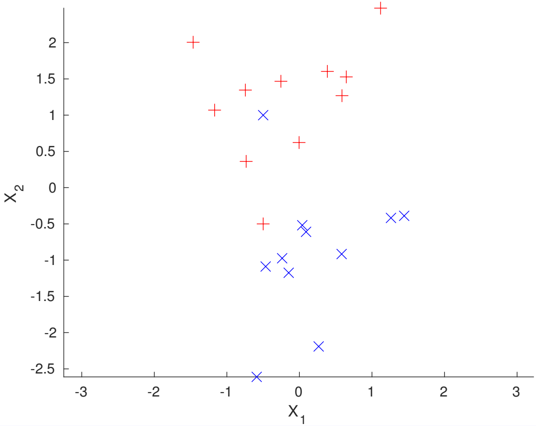
☞ Maximizing the *margin* M between the separating hyperplane and the training data :



Nonseparable case

- in general, overlap of the 2 classes

☞ No hyperplane that perfectly separates the training data



Maximum margin separating hyperplane (nonseparable case)

Solution for the nonseparable case

Considering a *soft-margin* that allows wrong classifications

- introduction of *slack variables* $\xi_i \geq 0$ s.t.

$$y_i(\boldsymbol{x_i}^T\boldsymbol{\beta} + \beta_0) \geq (1 - \xi_i)$$

  Support vectors include now the wrong classified points, and the points inside the margins ($\xi_i > 0$)

- Primal problem : adding a penalty in the criterion

$$\begin{cases} \min_{\boldsymbol{\beta},\beta_0,\xi} & \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i=1}^{n}\xi_i, \\ \text{subject to} & y_i(\boldsymbol{x_i}^T\boldsymbol{\beta} + \beta_0) \geq 1 - \xi_i, \end{cases}$$

  where $C > 0$ is the "cost" parameter

Cost parameter (nonseparable case)

$$\text{Criterion to be minimized :} \quad \frac{1}{2}||\boldsymbol{\beta}||^2 + C\sum_{i=1}^{n}\xi_i,$$

Influence of the cost parameter $C > 0$

C drives the margin size, thus the number of support vectors

- $C \gg 0$: $\sim$ underfitting (small margin, less support vectors)
- $C \to 0^+$: $\sim$ overfitting (large margin, more support vectors)
- $C \to +\infty$: converges to the separable case

Choosing the cost parameter $C > 0$

- the optimal C can be estimated by cross validation

☞ performance might not be very sensitive to choices of C (because of the rigidity of a linear boundary)

☞ usually $C \approx 1$ yields a good trade-off

Dual problem (nonseparable case)

Introducing the Lagrangian and substituting the first order KT conditions w.r.t. $\boldsymbol{\beta}$, β_0, $\boldsymbol{\xi}$ yields the dual expression

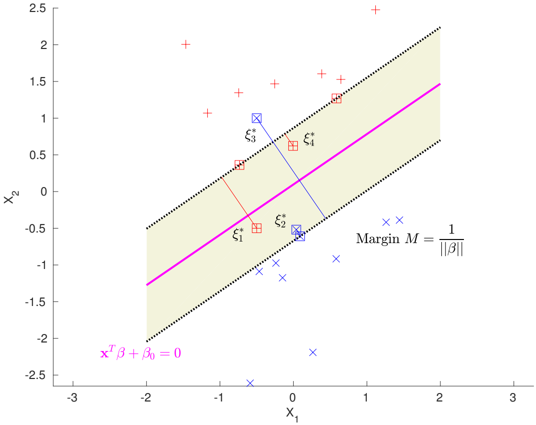
Dual optimization problem

$$\begin{cases} \max_{\boldsymbol{\alpha}} & \widetilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^n \alpha_i - \frac{1}{2}\sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j, \\ \text{subject to} & 0 \leq \alpha_i \leq C \text{ and } \sum_{i=1}^n \alpha_i y_i = 0. \end{cases}$$

- ☞ only difference w.r.t the separable case : $\alpha_i \leq C$ constraint !
- ☞ simple convex optimization problem for which standard numerical procedure are available

Optimal separating hyperplane

Example (nonseparable case)

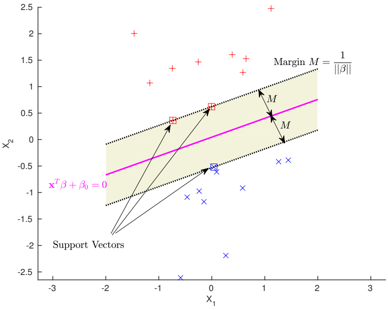


$\xi_i^* \equiv M\xi_i \leftarrow$ distance between a support vector and the margin

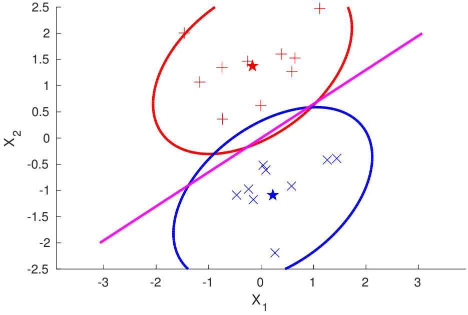
Linear discrimination : SVM vs LDA

Linear discrimination

- Linear Discriminant Analysis (LDA) : Gaussian generative model
- SVM : criterion optimization (maximizing the margin)



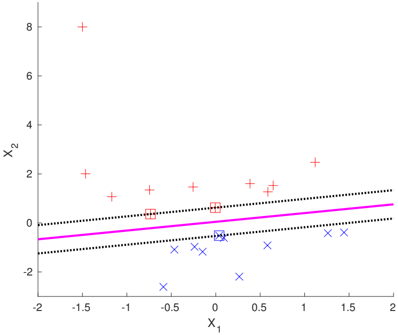
SVM



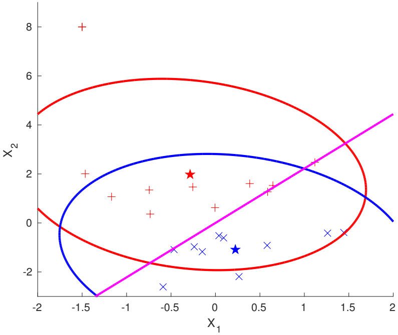
LDA

Linear discrimination : SVM vs LDA (Cont'd)

Adding one atypical data



SVM



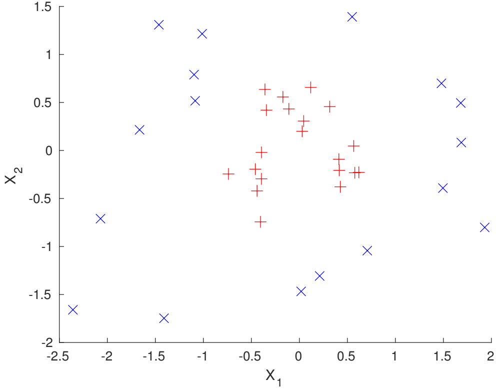
LDA

SVM property

- Nonsensitive to atypical points (outliers) far from the margin

☞ sparse method (information ≡ support vectors)

Nonlinear discrimination in the input space

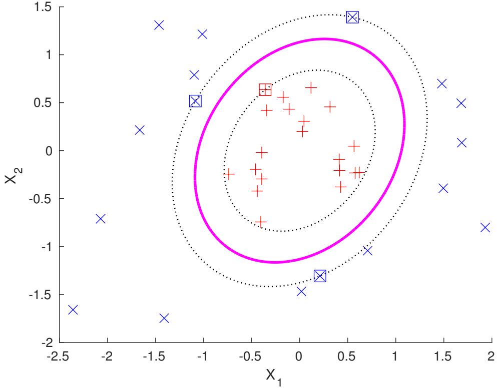


Transformed space $\mathcal{F}$

- Choice of a transformed space $\mathcal{F}$ (expansion space) where the linear separation assumption is more relevant
- Nonlinear expansion map $\phi\ : \mathbb{R}^p \to \mathcal{F}$, $\boldsymbol{x} \mapsto \phi(\boldsymbol{x}) \leftarrow$ enlarged features

Nonlinear discrimination in the input space

- $X \in \mathbb{R}^2$, $\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)^T$



Linear separation in the feature space $\mathcal{F} \Rightarrow$ Nonlinear separation in the input space

Kernel trick

The SVM solution depends only on the inner product between the input features $\phi(\boldsymbol{x})$ and the support vectors $\phi(\boldsymbol{x}_{\text{margin}})$

Kernel trick

Use of a kernel function k associated with an expansion/feature map ϕ :

$$\begin{array}{rccl} k: & \mathbb{R}^p \times \mathbb{R}^p & \to & \mathbb{R} \\ & (\boldsymbol{x}, \boldsymbol{x}') & \mapsto & k(\boldsymbol{x}, \boldsymbol{x}') \equiv \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle \end{array}$$

and the separating hyperplane reads $h(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\alpha}_i y_i k(\boldsymbol{x}_i, \boldsymbol{x}) + \widehat{\beta}_0$

Advantages

- computations are performed in the original input space : less expansive than in a high dimensional transformed space $\mathcal{F}$
- explicit representations of the feature map ϕ and enlarged feature space $\mathcal{F}$ are not necessary, the only expression of k is required !
- ☞ possibility of complex transformations in possible infinite space $\mathcal{F}$
- ☞ standard trick in machine learning not limited to SVM (kernel-PCA, gaussian process, kernel ridge regression, spectral clustering ...)

Choosing the Kernel function

Mercer theorem

$k(\cdot, \cdot)$ should be a symmetric positive (semi-) definite function

Usual kernel functions

- Linear kernel ($\mathcal{F} \equiv \mathbb{R}^p$) : $k(x, x') = x^T x'$
- Polynomial kernel (dimension of $\mathcal{F}$ increases with the order d)

$$k(x, x') = (x^T x')^d \quad \text{or} \quad (x^T x' + 1)^d$$

- Gaussian radial function ($\mathcal{F}$ with infinite dimension)

$$k(x, x') = \exp\left(-\gamma ||x - x'||^2\right)$$

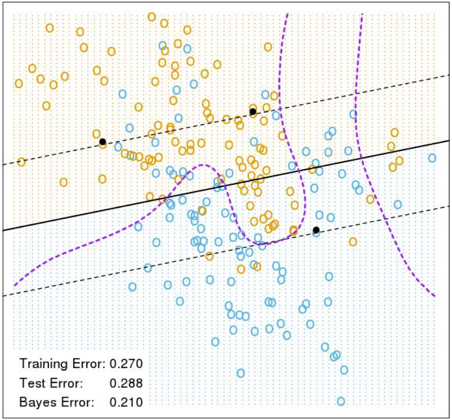
- Neural net kernel ($\mathcal{F}$ with infinite dimension)

$$k(x, x') = \tanh\left(\kappa_1 x^T x' + \kappa_2\right)$$

☞ optimal kernel parameters can be estimated by cross validation

Application : binary data (cf course 01)

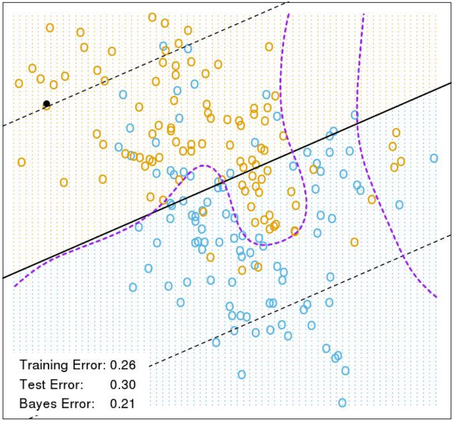
Linear kernel



$C = 10000$

Application : binary data (cf course 01)

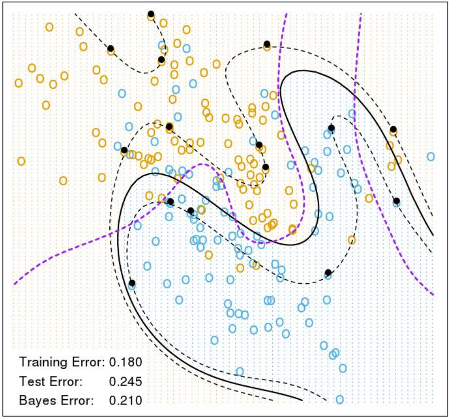
Linear kernel



$C = 0.01$

Application : binary data (cf course 01)

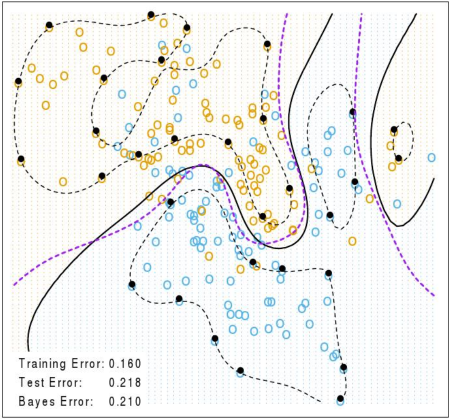
Polynomial kernel ($d = 4$)



$C \approx 1$

Application : binary data (cf course 01)

Gaussian radial kernel ($\gamma = 1$)



$C \approx 1$

Multiclass SVM

- $Y \in \{1, \ldots, K\} \leftarrow K$ classes

Standard approach : direct generalization by using multiple binary SVMs

OVA : one-versus-all strategy

- K classifiers between one class ($+1$ label) versus all the other classes (-1 label)

☞ classifier with the highest confidence value (e.g. the maximum distance to the separator hyperplane) assigns the class

OVO : one-versus-one strategy

- $\binom{K}{2} = K(K-1)/2$ classifiers between every pair of classes

☞ majority vote rule : the class with the most votes determines the instance classification

Which to choose ? if K is not too large, choose OVO

SVM vs Logistic regression (LR)

- When classes are nearly separable, SVM does better than LR. So does LDA.
- When not, LR (with ridge penalty) and SVM are very similar
- If one wants to estimate probabilities for each class, LR is the natural choice
- For non linear boudaries, kernel SVMs are popular. Can use kernels with LR and LDA as well, but computations are more expansive.

Conclusions

SVM

- maximum margin learning criterion $\leftarrow$ model free
- classification algorithm nonlinear in the original input space by performing an implicit linear classification in a higher dimensional space
- sparse solutions characterized by the support vectors
- popular algorithms, with a large literature