

Méthodes multivariées

Introduction à l'apprentissage automatique – GIF-4101 / GIF-7005

Professeur: Christian Gagné

Semaine 3



3.1 Données multivariées

Données multivariées

- Méthodes paramétriques telles que vues la semaine passée \Rightarrow estimation d'une variable X
 - En général, on mesure plusieurs variables $\{X_1, X_2, \dots, X_D\}$ pour une donnée

$$\mathcal{X} = \{\mathbf{x}^t, r^t\}_{t=1}^N, \quad \mathbf{x}^t = [x_1^t \ x_2^t \ \cdots \ x_D^t]^\top$$

- Appellations de variables (X_i)

- Entrées (*inputs*)
- Caractéristiques (*features*)
- Attributs

- Appellations de données (\mathbf{x}^t)

- Observations
- Exemples
- Instances

Représentation matricielle :

$$\mathbf{X} = \begin{bmatrix} x_1^1 & x_2^1 & \cdots & x_D^1 \\ x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \ddots & \vdots \\ x_1^N & x_2^N & \cdots & x_D^N \end{bmatrix}$$

Moyennes et variances, cas multivarié

- Vecteur moyen μ défini comme la moyenne selon chaque colonne (chaque variable) d'un ensemble \mathbf{X}

$$\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu} = [\mu_1 \ \mu_2 \ \cdots \ \mu_D]^\top$$

- Variance d'une variable X_i est σ_i^2
- Covariance de deux variables X_i et X_j est notée $\sigma_{i,j}$

$$\sigma_{i,j} \equiv \text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathbb{E}[X_i X_j] - \mu_i \mu_j$$

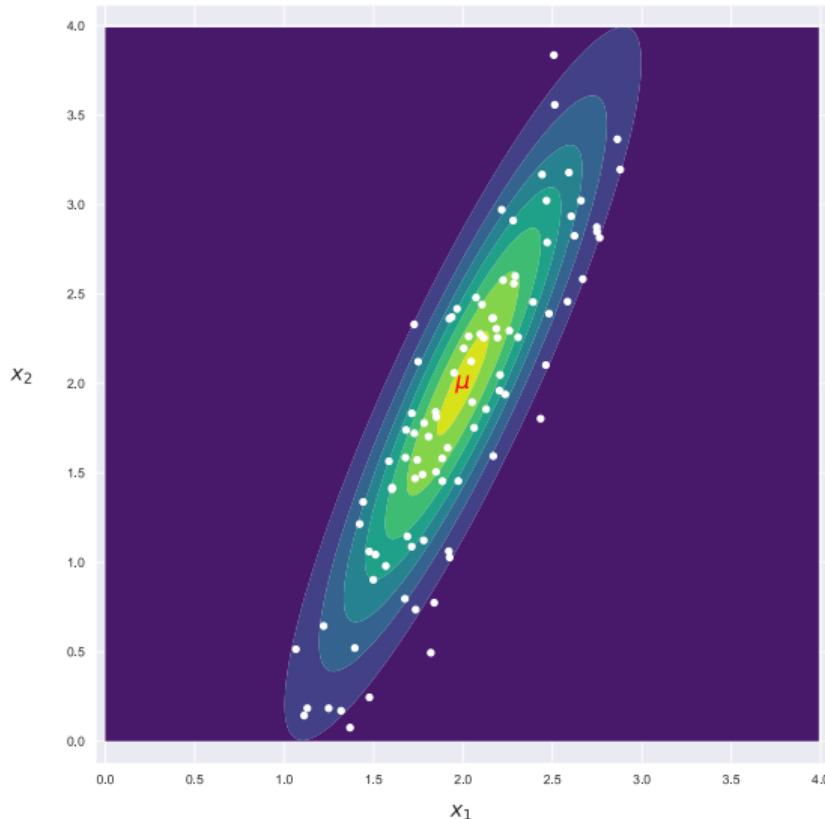
- Matrice de covariance Σ

- Matrice $D \times D$ symétrique ($\sigma_{i,j} = \sigma_{j,i}$)
- Valeurs positives sur la diagonale ($\sigma_{i,i} = \sigma_i^2$)

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \cdots & \sigma_{1,D} \\ \sigma_{1,2} & \sigma_2^2 & \cdots & \sigma_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1,D} & \sigma_{2,D} & \cdots & \sigma_D^2 \end{bmatrix}$$

$$\begin{aligned}\Sigma \equiv \text{Cov}(\mathbf{X}) &= \mathbb{E}[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^\top] \\ &= \mathbb{E}[\mathbf{X}\mathbf{X}^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top\end{aligned}$$

Moyenne et covariance d'échantillons



Estimateur de moyennes et variances, cas multivarié

- Estimateur de la moyenne selon un maximum de vraisemblance

$$\mathbf{m} = \frac{\sum_{t=1}^N \mathbf{x}^t}{N}, \text{ où } m_i = \frac{\sum_{t=1}^N x_i^t}{N}, i = 1, \dots, D$$

- Soit \mathbf{S} , l'estimateur de la matrice de covariance Σ

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{1,2} & \cdots & s_{1,D} \\ s_{1,2} & s_2^2 & \cdots & s_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ s_{1,D} & s_{2,D} & \cdots & s_D^2 \end{bmatrix} \quad \begin{aligned} s_i^2 &= \frac{\sum_{t=1}^N (x_i^t - m_i)^2}{N} \\ s_{i,j} &= \frac{\sum_{t=1}^N (x_i^t - m_i)(x_j^t - m_j)}{N} \end{aligned}$$

- Développement des équations pour \mathbf{S} est complexe, requiert l'application du théorème spectral

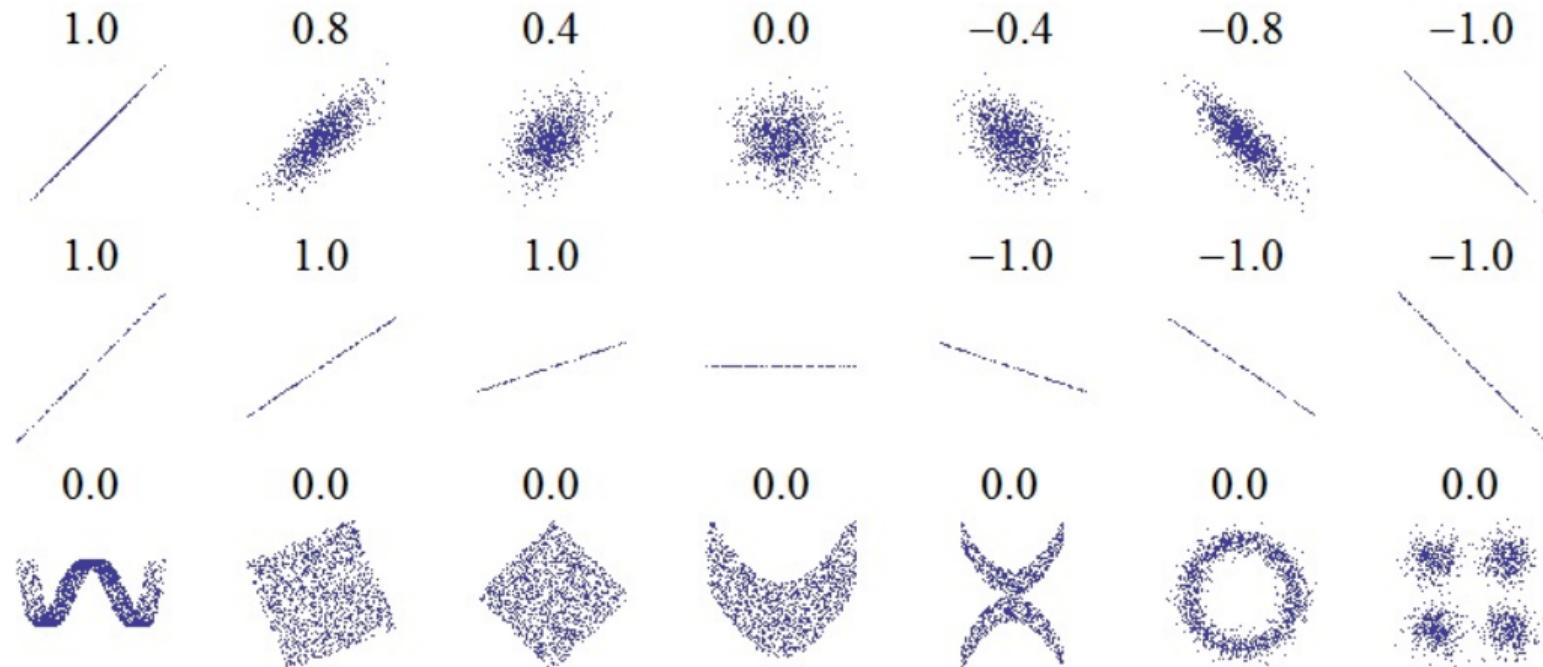
Corrélation

- Corrélation entre variables X_i et X_j

$$\text{Corr}(X_i, X_j) \equiv \rho_{i,j} = \frac{\sigma_{i,j}}{\sigma_i \sigma_j}$$

- Mesure statistique normalisée, $-1 \leq \rho_{i,j} \leq 1$
 - Deux variables X_i et X_j indépendantes \Rightarrow corrélation nulle
 - L'inverse n'est cependant pas vrai, même si $\rho_{i,j} = 0$, variables X_i et X_j ne sont pas nécessairement indépendantes (relation non-linéaire entre variables)
 - Estimation de la corrélation
- $$r_{i,j} = \frac{s_{i,j}}{s_i s_j}$$
- Matrice \mathbf{R} est la matrice de l'estimateur de corrélation contenant les $r_{i,j}$

Corrélation et non-linéarités



Source : http://en.wikipedia.org/wiki/File:Correlation_examples.png

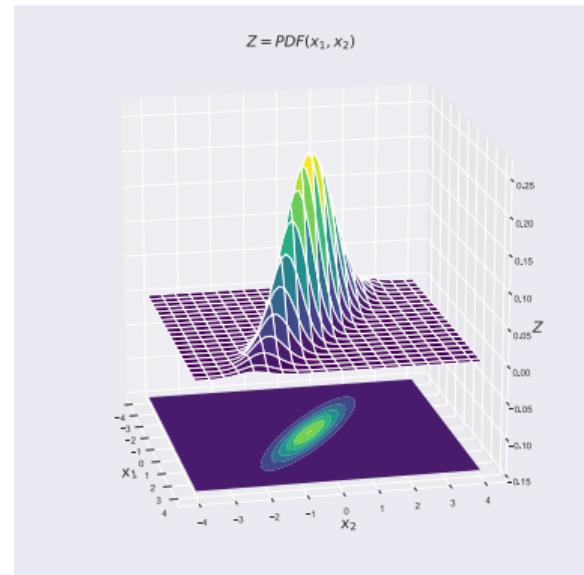
3.2 Loi normale multivariée

Loi normale multivariée

- Loi normale à plusieurs dimensions $\mathcal{N}_D(\mu, \Sigma)$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{0,5D} |\Sigma|^{0,5}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

- Vecteur moyen $\boldsymbol{\mu}$: centre de la distribution
- Normalisation par l'inverse de la matrice de covariance $\boldsymbol{\Sigma}$



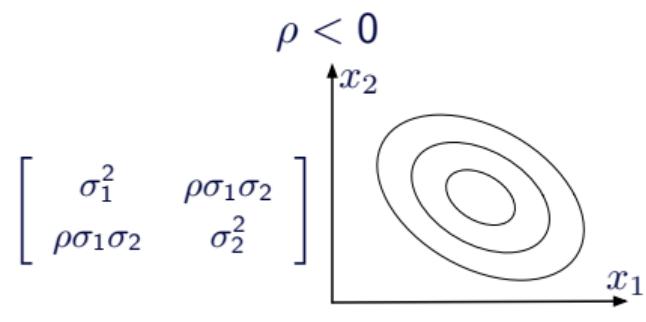
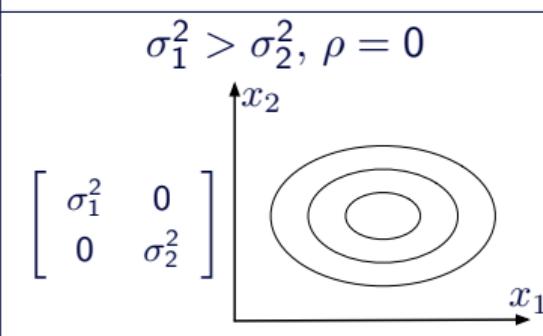
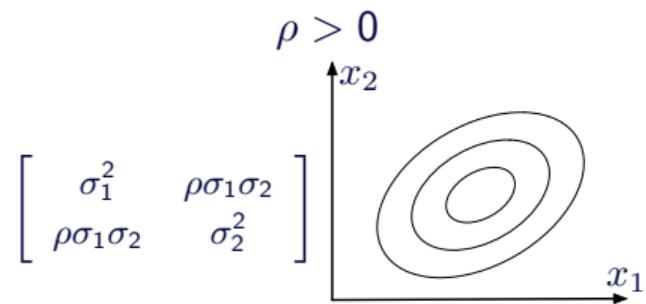
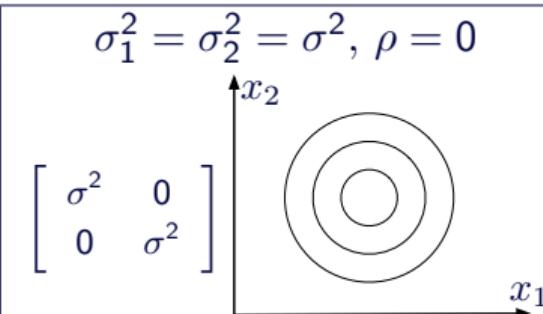
Exemple en deux dimensions

- Loi normale en deux dimensions ($\sigma_{i,j} = \rho\sigma_i\sigma_j$) :

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- Quatre cas de figure selon Σ
 1. Σ diagonale ($\rho = 0$) et variance égale pour les deux dimensions (isotropique),
 $\sigma_1^2 = \sigma_2^2 = \sigma^2$
 2. Σ diagonale ($\rho = 0$) et variances différentes selon les deux dimensions, $\sigma_1^2 \neq \sigma_2^2$
 3. Corrélation positive entre les variables, $\rho > 0$
 4. Corrélation négative entre les variables, $\rho < 0$

Exemple en deux dimensions

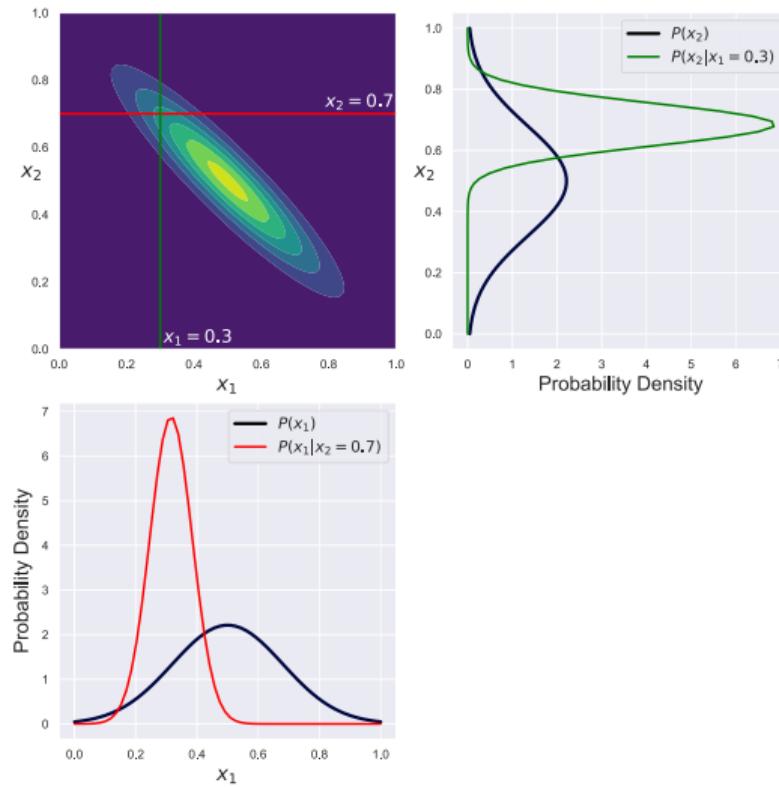


Propriétés de la loi normale multivariée

- Valeur du déterminant $|\Sigma|$ indique la proximité des échantillons autour de μ
 - Valeur faible peut indiquer une forte corrélation entre variables
- Généralement, Σ est une matrice symétrique définie positive
 - Sinon, Σ est singulière et $|\Sigma| = 0$
 - ⇒ Dépendance linéaire entre variables
 - ⇒ Variance nulle d'une variable
- Si $\mathbf{x} \sim \mathcal{N}_D(\mu, \Sigma)$ alors $x_i \sim \mathcal{N}(\mu_i, \tilde{\sigma}_i^2)$
 - Si x_i indépendants ($\sigma_{i,j} = 0, \forall i \neq j$), alors $x_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$
- Une projection linéaire définie par \mathbf{W} dans un espace à K dimensions ($K < D$) suit également une loi normale multivariée

$$\mathbf{W}^\top \mathbf{x} \sim \mathcal{N}_K \left(\mathbf{W}^\top \mu, \mathbf{W}^\top \Sigma \mathbf{W} \right)$$

Loi conditionnelle de loi normale multivariée



Distance de Mahalanobis

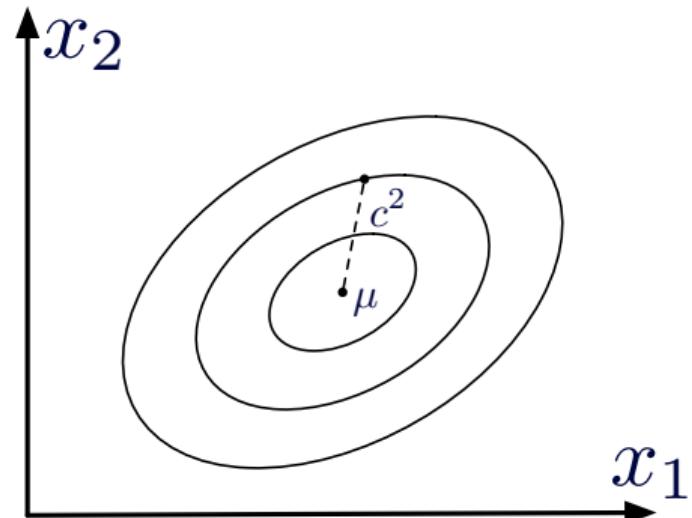
- Distance de Mahalanobis

$$D_M(\mathbf{x}) = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Distance entre le vecteur moyen $\boldsymbol{\mu}$ et un point \mathbf{x} , pondéré par la matrice de covariance $\boldsymbol{\Sigma}$
- Courbe de niveau correspond à distance constante c^2

- Cas 1D

$$\frac{(x - \mu)^2}{\sigma^2} = (x - \mu)(\sigma^2)^{-1}(x - \mu)$$



3.3 Classement multivarié

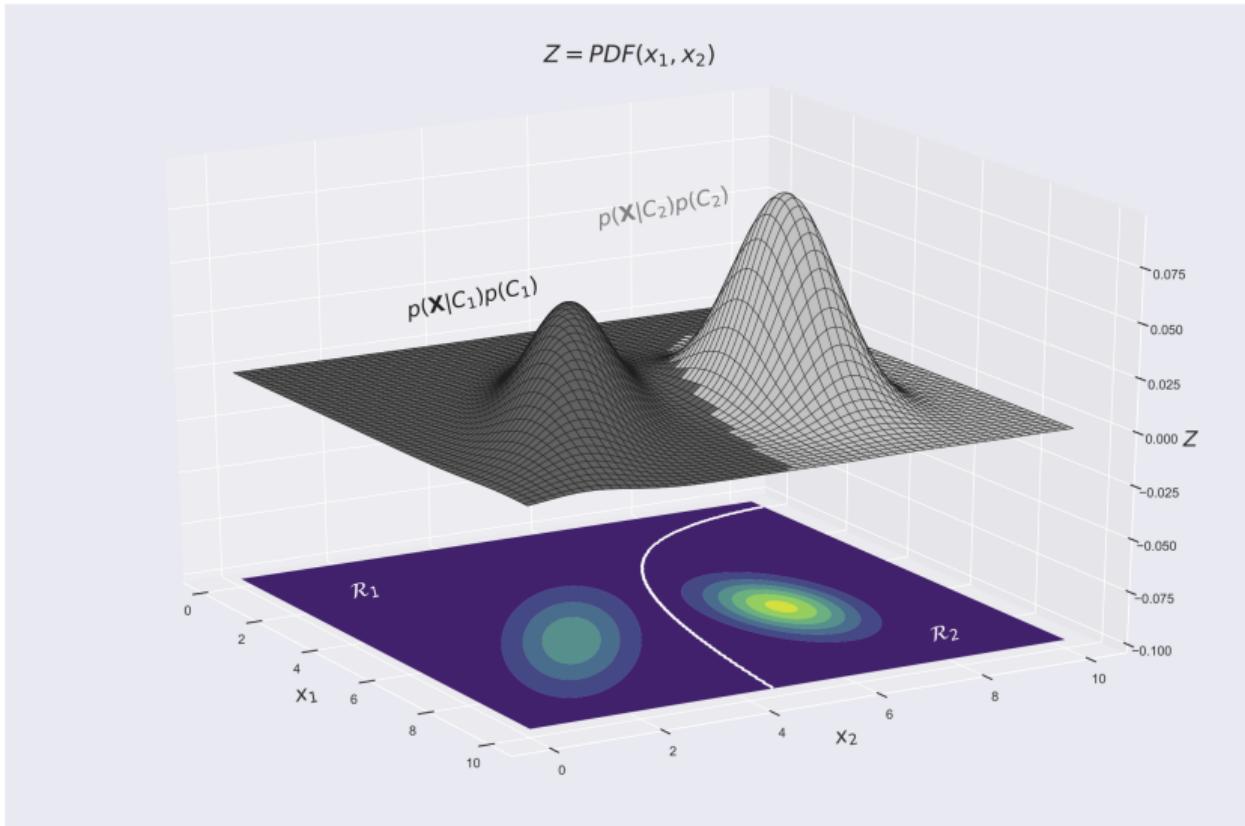
Classement multivarié

- Densité de probabilité conditionnelle aux classes $p(\mathbf{x}|C_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{0,5D} |\boldsymbol{\Sigma}_i|^{0,5}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right]$$

- Raisons pour l'utilisation de loi normale en classement multivarié
 - Simplicité de l'équation pour développements analytiques
 - Modèle de nombreux phénomènes naturels
 - Les observations sont en général des variations légères ($\boldsymbol{\Sigma}$) d'une observation moyenne ($\boldsymbol{\mu}$)
 - Modèle robuste, permet de bonnes approximations
 - Nécessite cependant que les données soient groupées
 - Avec plusieurs groupes, on doit faire des *densités-mélanges*, c'est-à-dire des combinaisons linéaires de densités (présenté plus loin)

Exemple de classement multivarié



Fonction discriminante

- Fonction discriminante avec modèle multivarié

$$h_i(\mathbf{x}) = \log p(\mathbf{x}|C_i) + \log P(C_i)$$

- Pour une loi normale, $p(\mathbf{x}|C_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$\begin{aligned} h_i(\mathbf{x}) &= -\frac{D}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \log P(C_i) \\ &= -\frac{1}{2} \log |\boldsymbol{\Sigma}_i| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) + \log P(C_i) \end{aligned}$$

Estimation des paramètres

- Estimation des paramètres selon un maximum de vraisemblance

- Ensemble $\mathcal{X} = \{\mathbf{x}^t, \mathbf{r}^t\}_{t=1}^N$, avec $r_i^t = \begin{cases} 1 & \text{si } \mathbf{x}^t \in C_i \\ 0 & \text{autrement} \end{cases}$

$$\begin{aligned}\hat{P}(C_i) &= \frac{\sum_t r_i^t}{N} \\ \mathbf{m}_i &= \frac{\sum_t r_i^t \mathbf{x}^t}{\sum_t r_i^t} \\ \mathbf{S}_i &= \frac{\sum_t r_i^t (\mathbf{x}^t - \mathbf{m}_i)(\mathbf{x}^t - \mathbf{m}_i)^\top}{\sum_t r_i^t}\end{aligned}$$

Fonction discriminante quadratique

- Intégrer $\hat{P}(C_i)$, \mathbf{m}_i et \mathbf{S}_i dans la formule de $h_i(\mathbf{x})$

$$h_i(\mathbf{x}) = -\frac{1}{2} \log |\mathbf{S}_i| - \frac{1}{2} (\mathbf{x} - \mathbf{m}_i)^\top \mathbf{S}_i^{-1} (\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

- Formulation équivalente

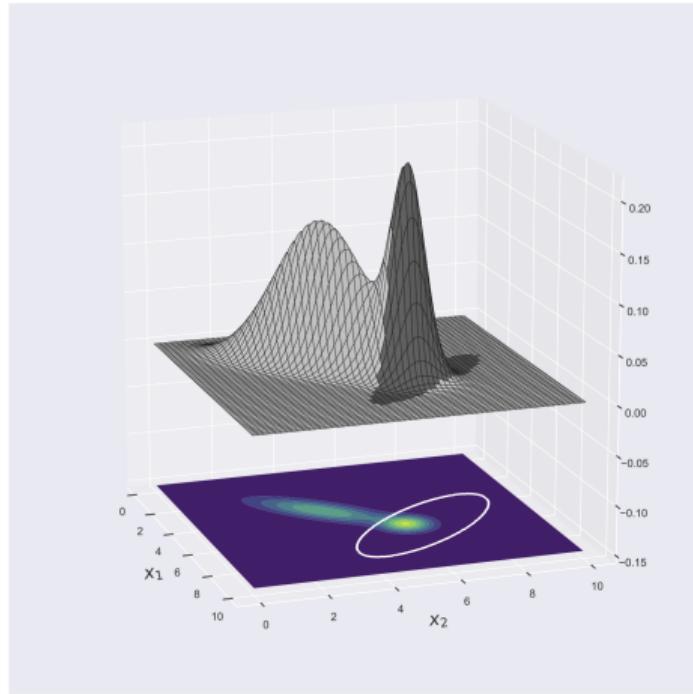
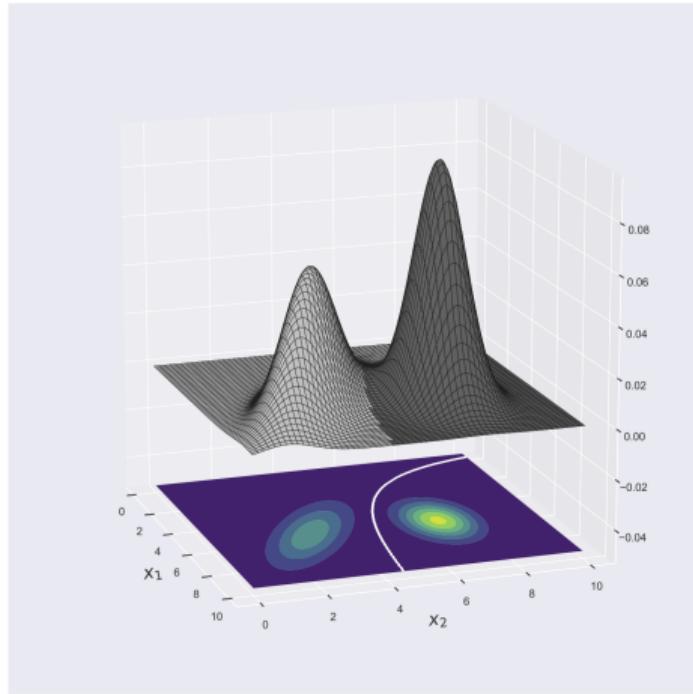
$$h_i(\mathbf{x}) = \mathbf{x}^\top \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^\top \mathbf{x} + w_i^0$$

$$\mathbf{W}_i = -\frac{1}{2} \mathbf{S}_i^{-1}$$

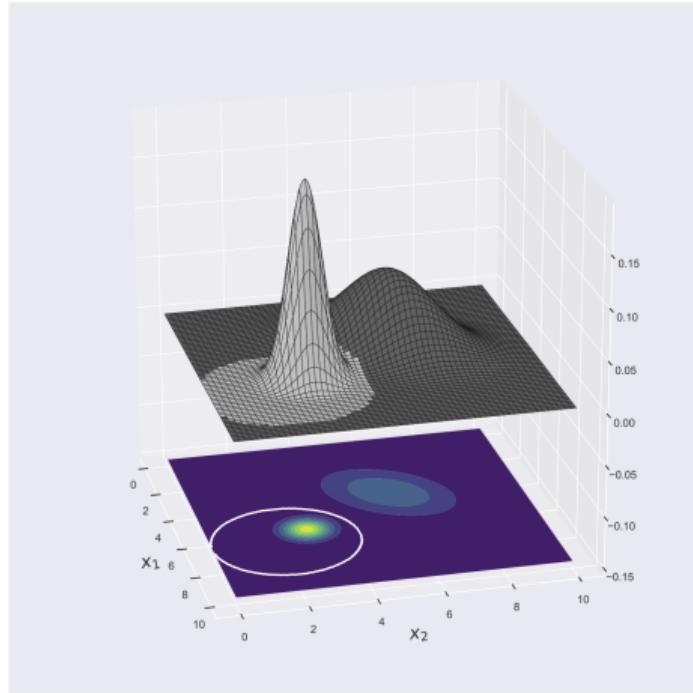
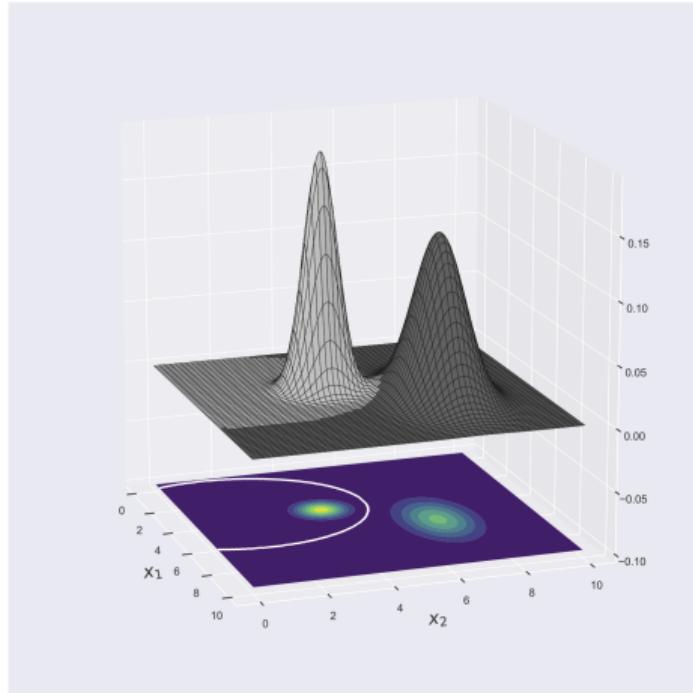
$$\mathbf{w}_i = \mathbf{S}_i^{-1} \mathbf{m}_i$$

$$w_i^0 = -\frac{1}{2} \mathbf{m}_i^\top \mathbf{S}_i^{-1} \mathbf{m}_i - \frac{1}{2} \log |\mathbf{S}_i| + \log \hat{P}(C_i)$$

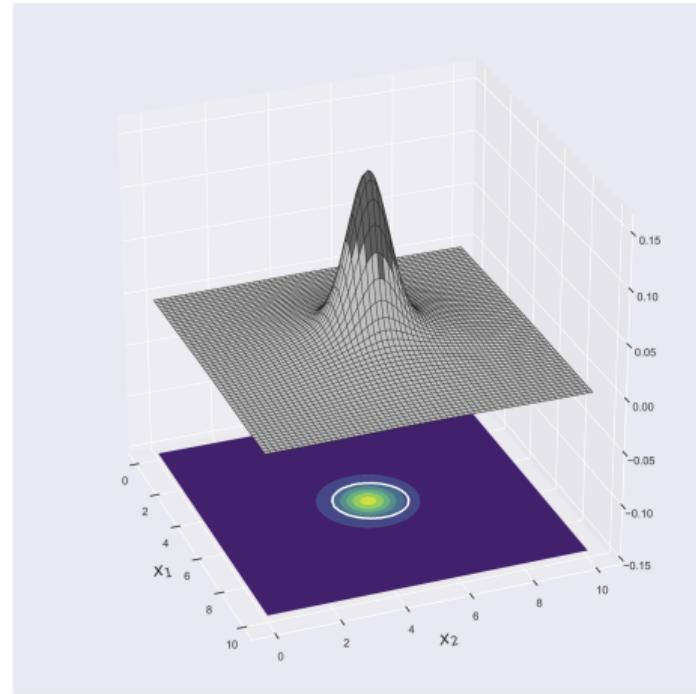
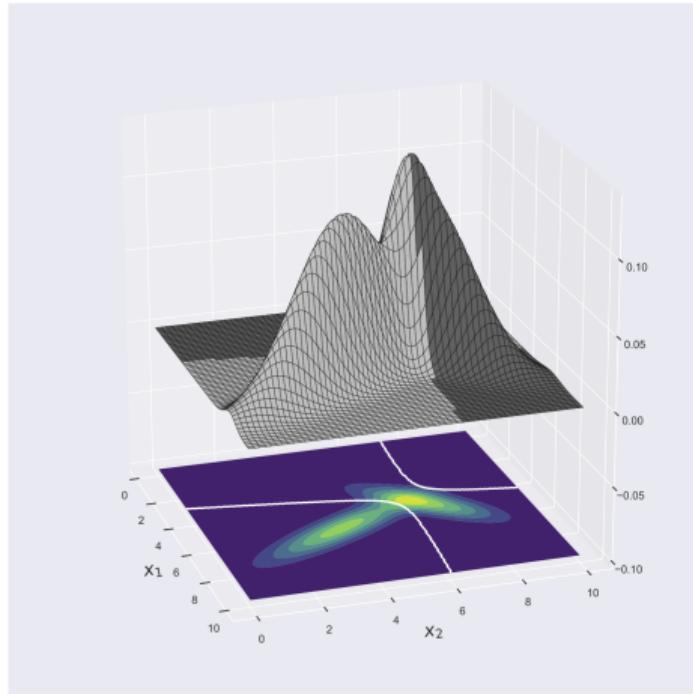
Exemples de fonction discriminante quadratique (1/3)



Exemples de fonction discriminante quadratique (2/3)



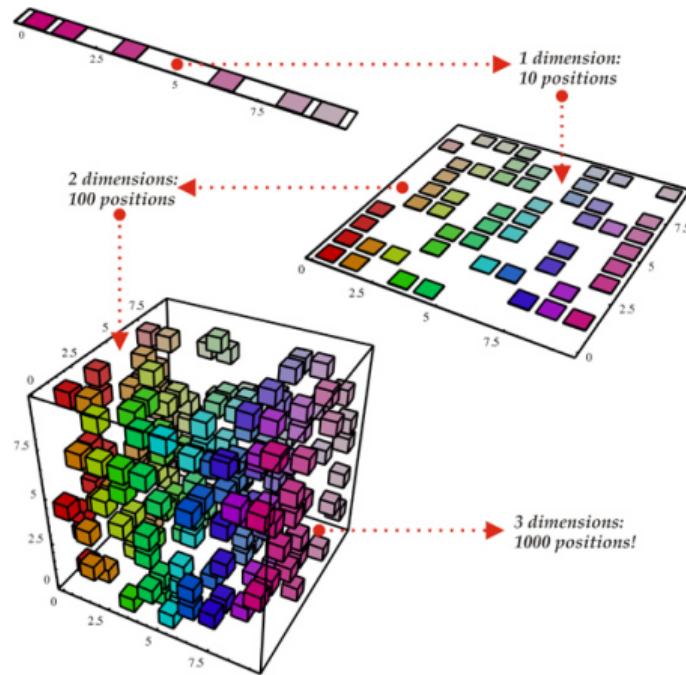
Exemples de fonction discriminante quadratique (3/3)



Malédiction de la dimensionnalité

- Malédiction de la dimensionnalité
 - Ajout d'une dimension crée une augmentation exponentielle de l'espace mathématique
 - 100 points équidistants de 0,01 en une dimension $\Rightarrow 10^{20}$ points nécessaires en 10 dimensions pour conserver la même densité d'échantillonnage
- Nombre élevé de paramètres à estimer avec fonction discriminante quadratique
 - $K \times D$ pour les moyennes et $K \times \frac{D(D+1)}{2}$ pour les matrices de covariance
- À grande dimensionnalité (D grand) et peu de données (N petit), risque élevé de matrices \mathbf{S}_i singulières
 - Même si $|\mathbf{S}_i| \neq 0$, un petit changement peut causer une importante variation de \mathbf{S}_i^{-1}
 \Rightarrow instabilités
- Solution : réduction de la dimensionnalité par traitement des caractéristiques (vu à la fin de la session)

Malédiction de la dimensionnalité



Source : Y. Bengio, http://www.iro.umontreal.ca/~bengioy/yoshua_en/research_files/CurseDimensionality.jpg, accédé le 2 octobre 2016.

3.4 Simplifications du modèle pour le classement

Partage de la matrice de covariance

- Simplification 1 : partage de la matrice de covariance

$$\mathbf{S} = \sum_t \hat{P}(C_i) \mathbf{S}_i$$

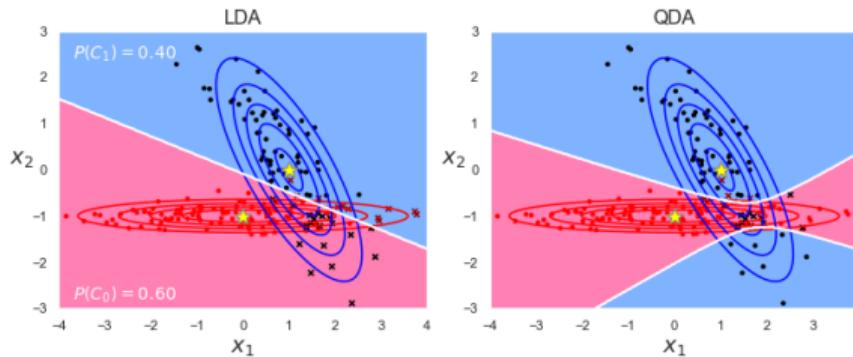
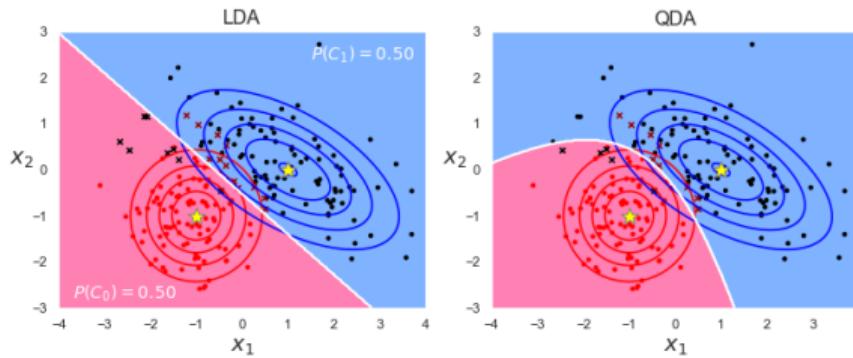
- $K \times D$ paramètres pour les moyennes
- $\frac{D(D+1)}{2}$ paramètres pour la matrice de covariance partagée
- Fonction discriminante correspondante

$$h_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mathbf{m}_i)^\top \mathbf{S}^{-1}(\mathbf{x} - \mathbf{m}_i) + \log \hat{P}(C_i)$$

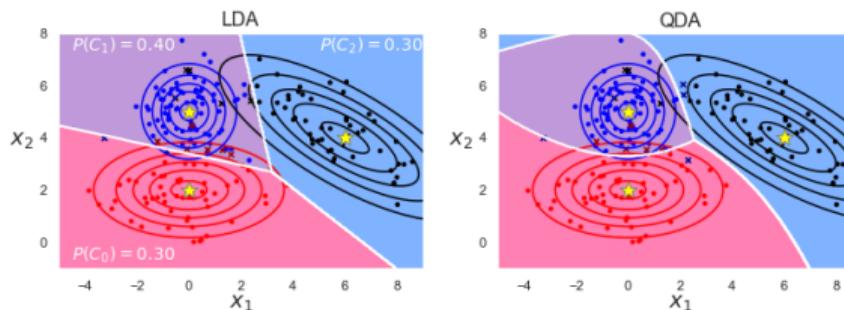
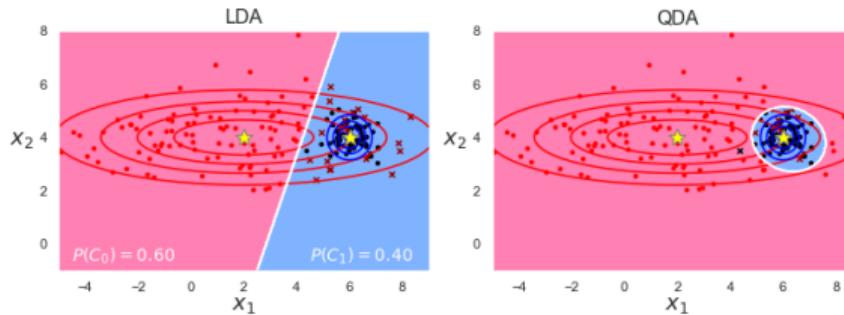
- $\mathbf{x}^\top \mathbf{S}^{-1} \mathbf{x}$ commun pour tout $h_i(\mathbf{x})$
- Reformulation comme une fonction discriminante linéaire

$$h_i(\mathbf{x}) = \mathbf{w}_i^\top \mathbf{x} + w_i^0, \quad \mathbf{w}_i = \mathbf{S}^{-1} \mathbf{m}_i, \quad w_i^0 = -\frac{1}{2} \mathbf{m}_i^\top \mathbf{S}^{-1} \mathbf{m}_i + \log \hat{P}(C_i)$$

Fonctions discriminantes linéaires et quadratiques (1/2)



Fonctions discriminantes linéaires et quadratiques (2/2)



Classifieur bayésien naïf

- Simplification 2 : éléments hors diagonale de \mathbf{S} nuls

$$\mathbf{S} = \begin{bmatrix} s_1^2 & 0 & \cdots & 0 \\ 0 & s_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & s_D^2 \end{bmatrix}$$

- Fonction discriminante correspondante (classifieur bayésien naïf)

$$h_i(\mathbf{x}) = -\frac{1}{2} \sum_{j=1}^D \left(\frac{x_j - m_{i,j}}{s_j} \right)^2 + \log \hat{P}(C_i)$$

- Nombre de paramètres pour la matrice de covariance : D
 - Réduction d'un ordre quadratique à un ordre linéaire

Classifieur à la plus proche moyenne

- Simplification 3 : matrice de covariance isotropique, avec toutes variances égales ($\sigma_i = \sigma, \forall i$)
- Réduction d'une distance de Mahalanobis à une distance Euclidienne

$$(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = \sigma^{-2} (\mathbf{x} - \boldsymbol{\mu})^\top (\mathbf{x} - \boldsymbol{\mu}) = \frac{\|\mathbf{x} - \boldsymbol{\mu}\|^2}{\sigma^2}$$

- Fonction discriminante correspondante

$$h_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2s^2} + \log \hat{P}(C_i) = -\frac{1}{2s^2} \sum_{j=1}^D (x_j - m_{i,j})^2 + \log \hat{P}(C_i)$$

- Simplification 4 : probabilités a priori égales ($P(C_i) = P(C_j), \forall i, j$)
 - Classifieur à la plus proche moyenne

$$h_i(\mathbf{x}) = -\|\mathbf{x} - \mathbf{m}_i\|^2$$

Classifieur à la plus proche moyenne

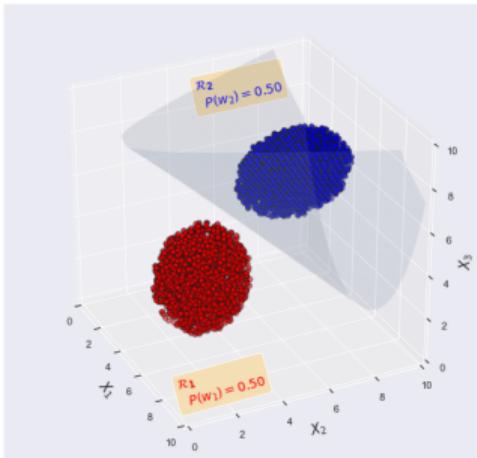
$$\begin{aligned} h_i(\mathbf{x}) &= -\|\mathbf{x} - \mathbf{m}_i\|^2 \\ &= -(\mathbf{x} - \mathbf{m}_i)^\top (\mathbf{x} - \mathbf{m}_i) \\ &= -(\mathbf{x}^\top \mathbf{x} - 2\mathbf{m}_i^\top \mathbf{x} + \mathbf{m}_i^\top \mathbf{m}_i) \end{aligned}$$

Comme $\mathbf{x}^\top \mathbf{x}$ commun $\forall h_i(\mathbf{x})$

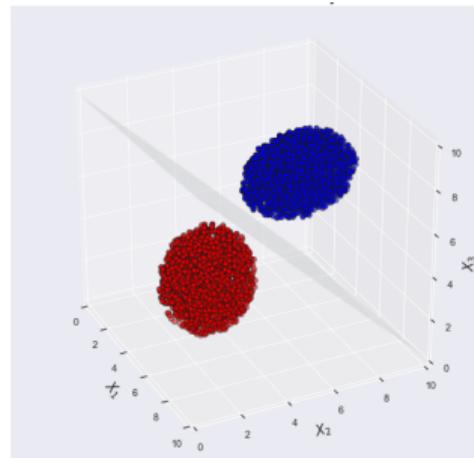
$$\begin{aligned} h_i(\mathbf{x}) &= \mathbf{w}_i^\top \mathbf{x} + w_i^0 \\ \mathbf{w}_i &= \mathbf{m}_i \\ w_i^0 &= -\frac{1}{2}\|\mathbf{m}_i\|^2 \end{aligned}$$

Exemple 3D avec simplifications

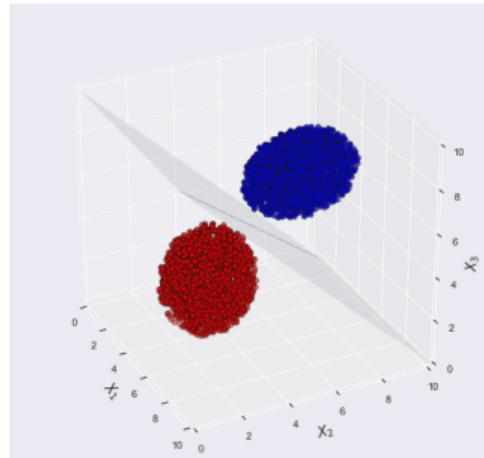
Quadratique



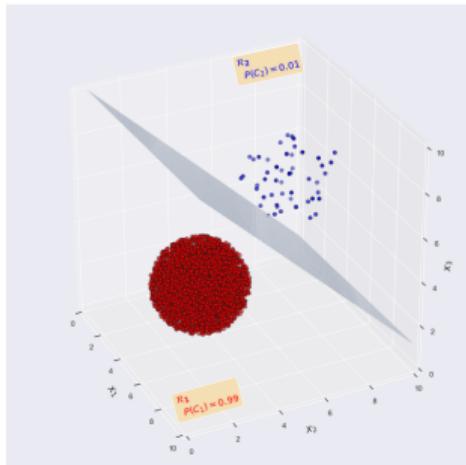
Linéaire



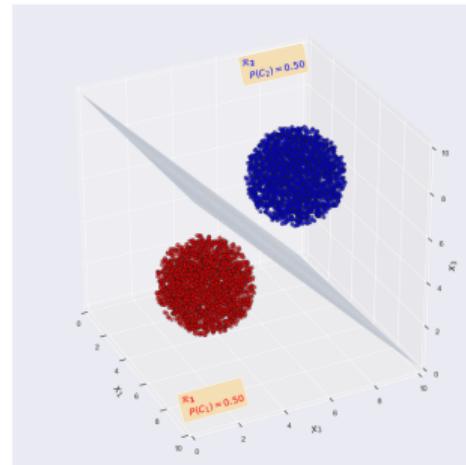
Plus proche moyenne



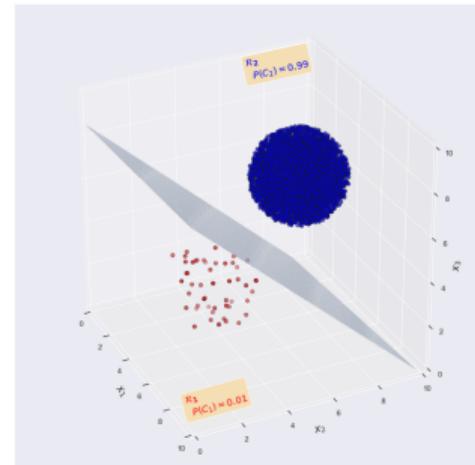
Effet des probabilités a priori (cas linéaire)



$$P(C_1) = 0.99, P(C_2) = 0.01$$



$$P(C_1) = 0.50, P(C_2) = 0.50$$

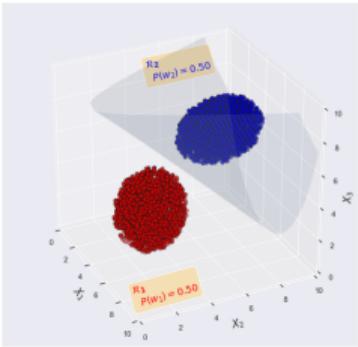


$$P(C_1) = 0.01, P(C_2) = 0.99$$

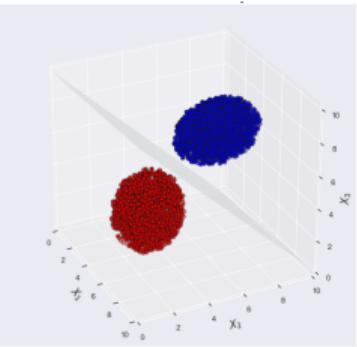
Effet des probabilités a priori

$$\begin{aligned}P(C_1) &= 0,99 \\P(C_2) &= 0,01\end{aligned}$$

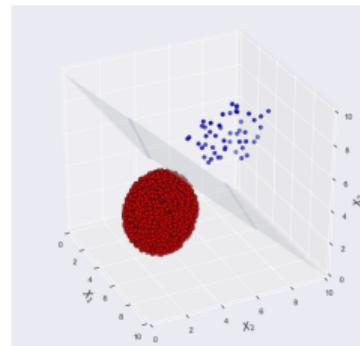
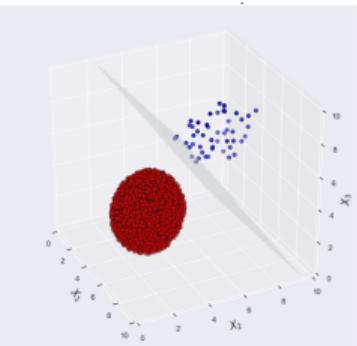
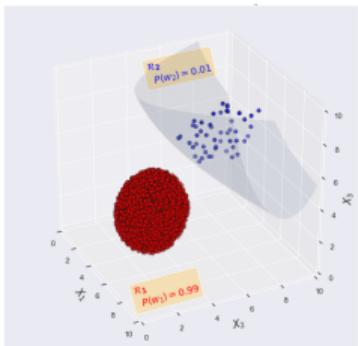
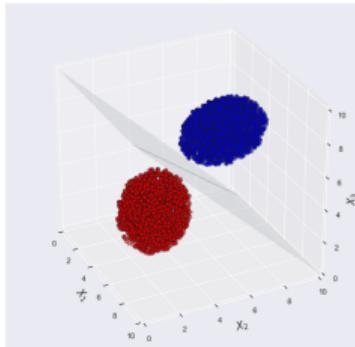
Quadratique



Linéaire



Plus proche moyenne



Résumé des variantes

Densités	Matrices de covariance	Nombre de paramètres
Σ partagée, densités hypersphères (isotropiques)	$\Sigma_i = \Sigma = \sigma^2 \mathbf{I}$	1
Σ partagée, densités alignées sur les axes	$\Sigma_i = \Sigma$ et $\sigma_{i,j} = 0$	D
Σ partagée, densités hyperellipsoïdales	$\Sigma_i = \Sigma$	$\frac{D(D+1)}{2}$
Σ différentes, densités hyperellipsoïdales	Σ_i	$K \frac{D(D+1)}{2}$

Analyse discriminante avec régularisation

- Ré-écriture de la matrice de covariance

$$\Sigma'_i = \alpha\sigma^2\mathbf{I} + \beta\Sigma + (1 - \alpha - \beta)\Sigma_i$$

- $\alpha = \beta = 0 \Rightarrow$ discriminant quadratique
- $\alpha = 0$ et $\beta = 1 \Rightarrow$ discriminant linéaire avec matrice de covariance partagée
- $\alpha = 1$ et $\beta = 0 \Rightarrow$ discriminant linéaire avec matrice de covariance isotropique partagée (classifieur à la plus proche moyenne si probabilités a priori égales)
- Variété de classifieurs avec α et β entre ces valeurs extrêmes
- Régularisation possible par un critère d'optimisation tenant compte des valeurs de α et β

3.5 Densité-mélange

Densité-mélange

- Classement paramétrique avec loi normale : un groupe par classe
 - Avec plusieurs modes dans une classe, modèle de loi normale s'applique difficilement
- Densité-mélange : combinaison linéaire de lois de densité associées à plusieurs groupes

$$p(\mathbf{x}) = \sum_{i=1}^K p(\mathbf{x}|\mathcal{G}_i)P(\mathcal{G}_i)$$

- Les groupes doivent être connus et identifiés dans les données
- Alternative : utiliser une approche non supervisée (*clustering*) pour apprendre les groupes
- Densité-mélange de composantes suivant une loi normale multivariée
 - Densité de composantes : $(\mathbf{x}|\mathcal{G}_i) \sim \mathcal{N}_D(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$
 - Paramétrisation : $\Phi = \{P(\mathcal{G}_i), \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^K$

Probabilités de la densité-mélange

- Densité-mélange

$$p(\mathbf{x}) = \sum_{i=1}^K p(\mathbf{x}|\mathcal{G}_i)P(\mathcal{G}_i)$$

- Proportion du groupe \mathcal{G}_i dans le mélange, $P(\mathcal{G}_i)$

$$\sum_i P(\mathcal{G}_i) = 1$$

- Probabilité que \mathbf{x} appartient au groupe \mathcal{G}_i , $P(\mathcal{G}_i|\mathbf{x})$

$$P(\mathcal{G}_i|\mathbf{x}) = \frac{P(\mathcal{G}_i)p(\mathbf{x}|\mathcal{G}_i)}{\sum_j P(\mathcal{G}_j)p(\mathbf{x}|\mathcal{G}_j)}$$

3.6 Régression multivariée

Régression multivariée

- Modèle de fonction linéaire en régression multivariée

$$r^t = h(\mathbf{x}|w_0, w_1, \dots, w_D) + \epsilon = w_0 + w_1 x_1^t + w_2 x_2^t + \dots + w_D x_D^t + \epsilon$$

- Bruit blanc gaussien de moyenne nulle, $\epsilon \sim \mathcal{N}(0, \sigma^2)$
- Minimisation de l'erreur quadratique (maximum de vraisemblance)

$$E(w_0, w_1, \dots, w_D | \mathcal{X}) = \frac{1}{2} \sum_t (r^t - w_0 - w_1 x_1^t - w_2 x_2^t - \dots - w_D x_D^t)^2$$

- Solution par dérivées partielles

$$\frac{\partial E}{\partial w_j} = 0, \forall j$$

Équations normales pour régression multivariée

$$\begin{aligned}\sum_t r^t &= Nw_0 + w_1 \sum_t x_1^t + w_2 \sum_t x_2^t + \cdots + w_D \sum_t x_D^t \\ \sum_t x_1^t r^t &= Nw_0 \sum_t x_1^t + w_1 \sum_t (x_1^t)^2 + w_2 \sum_t x_1^t x_2^t + \cdots + w_D \sum_t x_1^t x_D^t \\ \sum_t x_2^t r^t &= Nw_0 \sum_t x_2^t + w_1 \sum_t x_1^t x_2^t + w_2 \sum_t (x_2^t)^2 + \cdots + w_D \sum_t x_2^t x_D^t \\ &\vdots \\ &\vdots \\ \sum_t x_D^t r^t &= Nw_0 \sum_t x_D^t + w_1 \sum_t x_1^t x_D^t + w_2 \sum_t x_2^t x_D^t + \cdots + w_D \sum_t (x_D^t)^2\end{aligned}$$

- Version matricielle : $\mathbf{X}^\top \mathbf{r} = \mathbf{X}^\top \mathbf{X} \mathbf{w}$

$$\mathbf{X} = \begin{bmatrix} 1 & x_1^1 & x_2^1 & \cdots & x_D^1 \\ 1 & x_1^2 & x_2^2 & \cdots & x_D^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^N & x_2^N & \cdots & x_D^N \end{bmatrix}, \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix}, \mathbf{r} = \begin{bmatrix} r^1 \\ r^2 \\ \vdots \\ r^N \end{bmatrix}$$

- Résolution du système d'équations linéaires

$$\mathbf{w} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{r}$$

Notes sur la régression multivariée

- Équations normales : polynômes d'ordre 1
 - Résolution avec polynômes d'ordre supérieur rare, excepté pour D faible
- Analyse par inspection des valeurs w_i
 - w_i donne l'importance de la variable X_i , permet de classer les variables par importance
 - Retirer les variables dont $w_i \rightarrow 0$
 - Intéressant pour la réduction de la dimensionnalité (vu en fin de session)
 - Signe de w_i donne idée sur l'effet de la variable X_i
- Plusieurs valeurs de sortie \Rightarrow ensemble de problèmes de régression indépendants