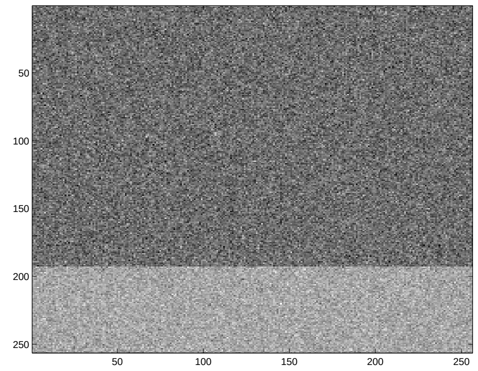
Théorie bayesienne de la décision

A. Rakotomamonjy, R. Flamary

8 avril 2013

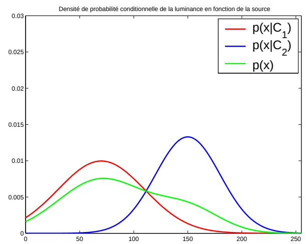
Introduction par l'exemple

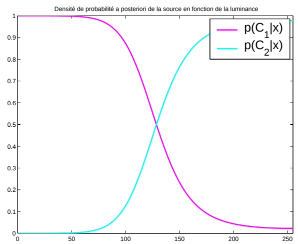
- On tire un pixel de l'image au hasard et on cherche à prédire si il provient de la zone 1 (sombre) ou de la zone 2(clair) avec le moins d'erreur possible
- comment décider si on a aucune information ?
- comment décider si on a une information de luminance du pixel ?



Introduction par l'exemple

- On a l'information de luminance $\mathbf{x}$ une v.a de densité de probabilité $p(\mathbf{x})$
- on suppose que l'on connait les densités de probabilité conditionnelle de x sachant les classes ω_j : $p(\mathbf{x}|\omega_1)$ et $p(\mathbf{x}|\omega_2)$.
- Règle de Bayes : $p(\mathbf{x}, \omega_i) = p(\mathbf{x}|\omega_i)p(\omega_i) = p(\omega_i|\mathbf{x})p(\mathbf{x})$
- on affecte $\mathbf{x}$ à la classe de plus forte probabilité a posteriori





Test du Rapport de vraisemblance

Cadre

- On veut catégoriser un objet à partir d'un ensemble de caractéristiques $\mathbf{x}$ et à l'aide de la vraisemblance de ce vecteur vis à vis des classes ω_j.
- une règle de décision intuitif et raisonnable est : "choisir la classe la plus probable étant donné l'observation $\mathbf{x}$.
  - Formellement : on évalue les probabilités $P(w_j|\mathbf{x})$ pour tout j et on choisit la classe qui maximise $P(w_j|\mathbf{x})$.

Analyse dans le cas où on a 2 classes

- Cette règle de décision s'écrit :

$$\text{si } P(w_1|\mathbf{x}) > P(w_2|\mathbf{x}) \text{ choisir } w_1 \text{ sinon } w_2 \quad \text{ou} \quad P(\omega_1|\mathbf{x}) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2|\mathbf{x})$$

Test du Rapport de Vraisemblance

En appliquant le théorème de Bayes, on a

$$\frac{p(\mathbf{x}|\omega_1)p(\omega_1)}{p(x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(\mathbf{x}|\omega_1)p(\omega_2)}{p(x)}$$

Comme $p(x)$ est toujours positif et n'influe pas dans le test, il peut etre éliminé, on a alors

$$\Lambda(x) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(\omega_2)}{p(\omega_1)}$$

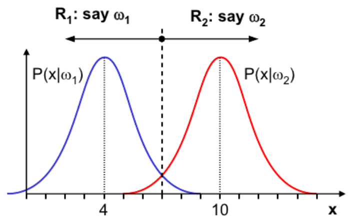
$\Lambda(\mathbf{x})$ est appelé le rapport de vraisemblance et la règle de décision le test de rapport de vraisemblance

Exemple de Test de RV

- On a un problème de catégorisation à deux classes et on connait les lois conditionnelles des caractéristiques sachant les classes

$$p(\mathbf{x}|\omega_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu_i)^2}$$

  avec $\mu_1 = 4$ et $\mu_2 = 10$. On suppose que les loi a priori de chaque classe sont égales *i.e* $p(\omega_1) = p(\omega_2)$
- quel est la règle de décision définie par le test de RV ?
  - On décide ω_1 dans la région $R_1 : \mathbf{x} \geq 7$ et ω_2 dans la région $R_2 : \mathbf{x} < 7$
  - résultat intuitif étant donné les lois



- que devient cette règle si $p(\omega_1) = 2p(\omega_2)$

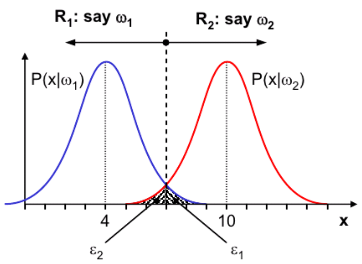
Probabilité d'erreur

- La performance d'une règle de décision donnée peut être évalué par sa probabilité d'erreur $P(Err)$, qui peut s'écrire étant donné, le théorème des probabilités totales $P(Err) = \sum_{i=1}^{C} P(Err|\omega_i)P(\omega_i)$
- La probabilité conditionnelle d'erreur sachant la classe peut également s'exprimer comme

$$P(Err|\omega_i) = P(choisir\ \omega_j|\omega_i) = \int_{R_j} p(\mathbf{x}|\omega_i)d\mathbf{x}$$

- Dans le cas à deux classes, on a donc

$$P(Err) = P(\omega_1)\int_{R_2} p(\mathbf{x}|\omega_1)d\mathbf{x} + P(\omega_2)\int_{R_1} p(\mathbf{x}|\omega_2)d\mathbf{x}$$

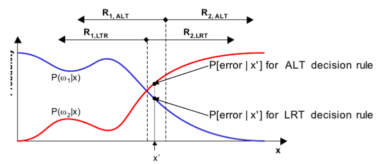


Probabilité d'erreur et Test RV

- Est que ce que le test de Rapport de Vraisemblance est une bonne règle de décision ?
  - Pour répondre à cette question, on s'interesse à la probabilité conditionnelle d'erreur $P(Err|\mathbf{x})$

$$P(Err) = \int_{\mathbf{x}} P(Err|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

  - la règle optimale minimise $P(Err|\mathbf{x})$ pour tout $\mathbf{x}$
  - en un point $\mathbf{x}'$, $P(Err|\mathbf{x}')$ est égale à $P(\omega_i|\mathbf{x}')$ quand la classe ω_j est choisie.



- De la figure, on devine que pour tout $\mathbf{x}'$, la règle de décision issue du Test de RV admet un $P(Err|\mathbf{x}')$ plus faible que celle d'une règle alternative
- Pour tout problème de catégorisation, la règle de décision issue du Test de RV admet la plus faible probabilité d'erreur. Cette probabilité d'erreur est appelé Erreur de Bayes.

Risque de Bayes

- Pour le calcul de l'erreur de Bayes, on fait l'hypothèse qu'une erreur de catégorisation admet le même coût.
  - Cette hypothèse n'est pas verifiée dans beaucoup de situations.Par exemple, le cas des diagnostic médical
- On formalise ce concept en introduisant la notion de coût $C_{i,j}$
  - $C_{i,j}$ représente le coût de choisir ω_i alors que la vraie classe est ω_j.
- On définit le risque de Bayes comme étant l'espérance du coût $\mathcal{R} = E(C)$

$$E(C) = \sum_i \sum_j C_{i,j} P(\text{choisir } \omega_i \text{ et } \mathbf{x} \in \omega_j) = \sum_i \sum_j C_{i,j} P(\mathbf{x} \in R_i | \omega_j) P(\omega_j)$$

- Quelle est la règle de décision qui minimise le risque de Bayes dans le cas 2 classes ?
  - On note tout d'abord que

$$P(\mathbf{x} \in R_i | \omega_j) = \int_{R_i} p(\mathbf{x}|\omega_j) d\mathbf{x}$$

Risque de Bayes (2)

- Le risque s'écrit

$$\begin{aligned} \mathcal{R} &= \int_{R_1} [C_{11} P(\omega_1) p(\mathbf{x}|\omega_1) + C_{12} P(\omega_2) p(\mathbf{x}|\omega_2)] d\mathbf{x} + \\ & \int_{R_2} [C_{21} P(\omega_1) p(\mathbf{x}|\omega_1) + C_{22} P(\omega_2) p(\mathbf{x}|\omega_2)] d\mathbf{x} \end{aligned}$$

- et comme $\int_{R_2} p(\mathbf{x}|\omega_i) d\mathbf{x} = 1 - \int_{R_1} p(\mathbf{x}|\omega_i) d\mathbf{x}$, on a

$$\begin{aligned} \mathcal{R} &= C_{21} P(\omega_1) + C_{22} P(\omega_2) \\ &+ (C_{12} - C_{22}) P(\omega_2) \int_{R_1} P(\mathbf{x}|\omega_2) d\mathbf{x} - (C_{21} - C_{11}) P(\omega_1) \int_{R_1} P(\mathbf{x}|\omega_1) d\mathbf{x} \end{aligned}$$

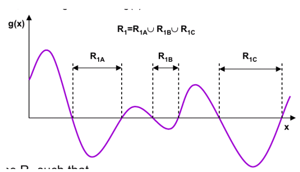
- Si on cherche à minimiser le risque, les deux premiers termes sont sans influence et on cherche donc une règle de décision telle que la région R_1 définie par cette règle minimise

$$\begin{aligned} R_1 &= \arg\min \int_{R_1} [(C_{12} - C_{22}) P(\omega_2) P(\mathbf{x}|\omega_2) - (C_{21} - C_{11}) P(\omega_1) P(\mathbf{x}|\omega_1)]\, d \\ &= \arg\min \int_{R_1} g(\mathbf{x}) d\mathbf{x} \end{aligned}$$

Risque de Bayes (3)

Intuition sur $g(x)$ et la région R_1

- On cherche une région R_1 qui minimise l'intégrale $\int_{R_1} g(\mathbf{x}) d\mathbf{x}$. On cherche donc les régions telles que $g(x) < 0$



- On choisit donc R_1 tel que

$$(C_{21} - C_{11}) P(\omega_1) P(\mathbf{x}|\omega_1) > (C_{12} - C_{22}) P(\omega_2) P(\mathbf{x}|\omega_2)$$

- En ré-arrangeant les termes, la règle de décision minimisant le risque de Bayes est

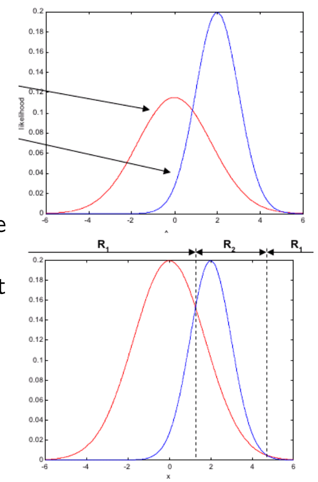
$$\Lambda(x) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{C_{12} - C_{22}}{C_{21} - C_{11}} \frac{p(\omega_2)}{p(\omega_1)}$$

qui est également définit par un test de rapport de vraisemblance

Exemple de Risque de Bayes

- Problème de catégorisation à deux classes avec $p(\mathbf{x}|\omega_1) = \frac{1}{\sqrt{2\pi}\sqrt{3}} e^{\frac{1}{2}\frac{\mathbf{x}^2}{3}}$, $p(\mathbf{x}|\omega_2) = \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(\mathbf{x}-2)^2}$
- tracer les densités
- que vaut le rapport de vraisemblance ?
- quelle est la règle de décision qui minimise le risque de Bayes ? si $P(\omega_1) = P(\omega_2), C_{11} = C_{22} = 0,\ C_{12} = 1$ et $C_{21} = \sqrt{3}$
-

$$2\mathbf{x}^2 - 12\mathbf{x} + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0$$



Variations sur un même thème : Test du RV

- Le classifieur basé sur le test du RV qui minimise le risque de Bayes s'écrit

$$\Lambda(x) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{C_{12}-C_{22}}{C_{21}-C_{11}} \frac{p(\omega_2)}{p(\omega_1)}$$

et est appelé Classifieur de Bayes
- Le classifieur basé sur le test du RV qui minimise la probabilité d'erreur ($C_{ij} = 1$ si $i = j$ et 0 sinon) s'écrit

$$\Lambda(x) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(\omega_2)}{p(\omega_1)} \Leftrightarrow \Lambda(x) = \frac{p(\omega_1|\mathbf{x})}{p(\omega_2|\mathbf{x})} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

et est appelé Classifieur du maximum à posteriori
- Le classifieur basé sur le test du RV qui minimise la probabilité d'erreur sachant que les classes ont une probabilité égale $P(\omega_1) = P(\omega_2) = 0.5$

$$\Lambda(x) = \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

et est appelé Classifieur du maximum de vraisemblance

Probabilité d'erreur dans le cas de plusieurs classes

- La règle de décision qui minimise la probabilité d'erreur se généralise facilement au cas multi-classe.
- La probabilité d'erreur s'obtient, dans ce cas, par la probabilité de bonne catégorisation

$$P(Err) = 1 - P(Corr) = 1 - \sum_{i=1}^{C} P(\omega_i) \int_{R_i} p(\mathbf{x}|\omega_i) d\mathbf{x}$$

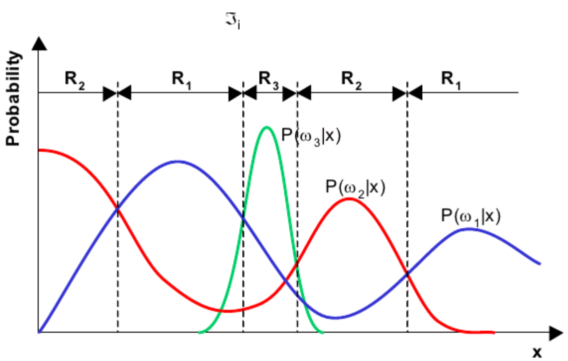
- On cherche de manière équivalente à maximiser la probabilité de bonne catégorisation

$$P(Corr) = \sum_{i=1}^{C} P(\omega_i) \int_{R_i} p(\mathbf{x}|\omega_i) d\mathbf{x} = \sum_{i=1}^{C} \int_{R_i} p(\mathbf{x}|\omega_i) P(\omega_i) d\mathbf{x} = \sum_{i=1}^{C} \int_{R_i} p(\omega_i|\mathbf{x})$$

- Pour maximiser $P(Corr)$,il s'agit donc de maximiser chaque intégrale. Cela s'obtient en choisissant, pour un x donnée, la classe qui maximise $P(\omega_i|\mathbf{x})$.

Probabilité d'erreur dans le cas de plusieurs classes

- Pour maximiser $P(Corr)$,il s'agit donc de maximiser chaque intégrale. Cela s'obtient en choisissant, pour un x donnée, la classe qui maximise $P(\omega_i|\mathbf{x})$.
- La région R_i est définie comme la région où $P(\omega_i|\mathbf{x})$ est maximum.
- La règle de décision qui minimise la probabilité d'erreur est la règle qui maximise le critère de maximum a posteriori.



Risque de Bayes dans le cas multi-classe

- Pour déterminer la règle de décision qui minimise le risque de Bayes dans le cas multi-classe, on appelle
  - α_i, la décision de choisir la classe ω_i
  - $\alpha(\mathbf{x})$ la règle de décision qui associe à un vecteur de caractéristique $\mathbf{x}$ une classe ω_i
- le risque conditionnel associé à α_i, étant donné l'observation $\mathbf{x}$ est :

$$R(\alpha_i|x) = \sum_{j=1}^{C} C_{i,j} P(\omega_j|\mathbf{x})$$

- et le risque associé à une règle de décision $\alpha(\mathbf{x})$ est

$$R(\alpha(\mathbf{x})) = \int R(\alpha(\mathbf{x})|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- si on veut minimiser ce risque, on minimise le risque conditionnel pour tout $\mathbf{x}$. Cela consiste à choisir la classe ω_i pour lequel $R(\alpha_i|x)$ est minimum, *i.e* la classe i telle que $\sum_{j=1}^{C} C_{i,j} P(\omega_j|\mathbf{x}) < \sum_{j=1}^{C} C_{k,j} P(\omega_j|\mathbf{x})\ \forall k$

Fonctions discriminantes

- L'ensemble des règles de décision que nous avons vu ont la même structure
  - un vecteur de caractéristiques $\mathbf{x}$ est assignée à la classe qui maximise (ou minimise) une certaine mesure $g_i(\mathbf{x})$
- Cette structure peut se formaliser à l'aide d'un ensemble de fonctions discriminantes $\{g_i(\mathbf{x})\}_{i=1}^{C}$ et la règle de décision définit comme

$$\text{assigner} \mathbf{x} \text{ à la classe } \omega_i \text{ telle que } g_i(\mathbf{x}) > g_j(\mathbf{x}),\ \forall j \neq i$$

- Pour les trois règles de décision qui nous interessent, nous avons :

Classifieur	Fonctions discriminantes
Bayes	$g_i(\mathbf{x}) = -R(\alpha_i\|\mathbf{x})$
Maximum a Posteriori	$g_i(\mathbf{x}) = P(\omega_i\|\mathbf{x})$
Maximum de Vraisemblance	$g_i(\mathbf{x}) = p(\mathbf{x}\|\omega_i)$

Classifieur à minimum d'erreur pour des classes de lois Gaussiennes

- Dans ce cas, on s'interesse aux fonctions discriminantes de la forme $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$
- On montre dans ce qui suit que pour des données suivant une loi Gaussienne ces fonctions discriminantes peuvent prendre une expression très simple
- Expression générale
  - $p(\mathbf{x}|\omega_i)$ est une loi normale

$$p(\mathbf{x}|\omega_i) = \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i)}$$

  - le théorème de Bayes donne

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(x)} = \frac{1}{2\pi^{d/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i)} P(\omega_i)\frac{1}{p(x)}$$

  - En éliminant les termes constants en i et en prenant le log (fonction croissante qui ne modifie pas la relation d'ordre), on a

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))$$

Cas 1 : $\Sigma_i = \sigma^2 I)$

- les caractéristiques sont statistiquement indépendants et de même variance pour toutes les classes
- la fonction discriminante quadratique $g_i(\mathbf{x})$ devient

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i) - \frac{1}{2}d\log(\sigma^2) + \log(P(\omega_i))$$

  d étant la dimensionalité du problème.
- En laissant de coté les termes constants et en développant, on a

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}(\mathbf{x}^T\mathbf{x} - 2\mu_i^T\mathbf{x} + \mu_i^T\mu_i) + \log(P(\omega_i))$$

- En eliminant les termes $\mathbf{x}^T\mathbf{x}$ qui sont constant pour toutes les fonctions discriminantes, on a

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_0 \quad \text{avec } \mathbf{w}_i = \frac{\mu_i}{\sigma^2} \text{ et } w_0 = -\frac{1}{2\sigma^2}\mu_i^T\mu_i + \log P(\omega_i)$$

- La fonction de décision est linéaire
- si en plus, on fait l'hypothèse que les $P(\omega_i)$ sont égaux, alors

$$g_i(x) = -\frac{1}{2\sigma^2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i)$$

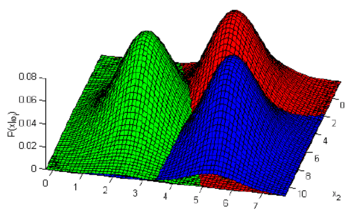
  on obtient un classifieur basé sur la distance minimum à la moyenne

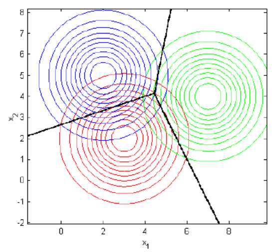
Cas 1 : $\Sigma_i = \sigma^2 I$ - Exemple

- illustration pour un cas à trois classes et 2 dimensions :

$\mu_1 = (3\ 2)^T \quad \mu_2 = (7\ 4)^T \quad \mu_1 = (2$

et $\Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$





Cas 2 : $\Sigma_i = \Sigma$ avec Σ diagonal

- Les classes ont la meme matrice de covariance mais les caractéristiques ont des variances différentes
- la fonction discriminante quadratique $g_i(\mathbf{x})$ devient

$$g_i(\mathbf{x}) = -\frac{1}{2}\sum_{k=1}^{d}\frac{(\mathbf{x}_k - \mu_{i,k})^2}{\sigma_k^2} - \frac{1}{2}\log(\Pi_{k=1}^{d}\sigma_k^2) + \log(P(\omega_i))$$

- en développant et en eliminant les termes $\mathbf{x}^T\mathbf{x}$ qui sont constant pour toutes les fonctions discriminantes, on a

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_0 \quad \text{avec } \mathbf{w}_{i,k} = -\frac{\mu_{i,k}}{\sigma_k^2} \text{ et } w_{0,k} = -\sum_k \frac{\mu_{i,k}}{2\sigma_k^2} + \log P(\omega_i)$$

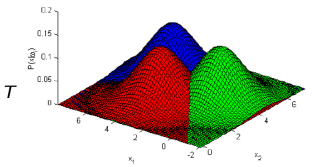
- On obtient encore une fonction de décision linéaire

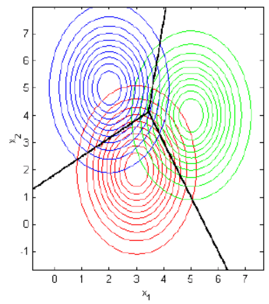
Cas 2 : $\Sigma_i = \Sigma$ avec Σ diagonal

- illustration pour un cas à trois classes et 2 dimensions :

  $\mu_1 = (3\ 2)^T \ \mu_2 = (5\ 4)^T \ \mu_3 = (2\ 5)^T$

  et $\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$





Cas 3 : $\Sigma_i = \Sigma$ avec Σ quelconque

- Les classes ont la meme matrice de covariance
- la fonction discriminante quadratique $g_i(\mathbf{x})$ devient

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T\Sigma_i^{-1}(\mathbf{x} - \mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))$$

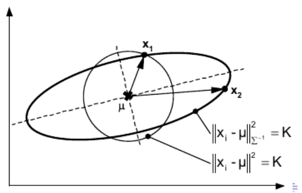
- En remplaçant Σ_i par Σ et en éliminant les termes constantes pour toutes les classes, on a :

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T\Sigma_i^{-1}(\mathbf{x} - \mu_i) + \log(P(\omega_i))$$

- le terme quadratique est un terme important en reconnaissance de formes, il se nomme distance de Mahalanobis

$$\|x - y\|^2_{\Sigma^{-1}} = (\mathbf{x} - \mathbf{y})^T\Sigma_i^{-1}(\mathbf{x} - \mathbf{y})$$

- La distance de Mahalanobis utilise une métrique en Σ^{-1} qui a tendance à étirer et "tordre" l'espace



Cas 3 : $\Sigma_i = \Sigma$ avec Σ quelconque

- En développant le terme quadratique et en supprimant les termes constantes pour toutes les classes, on obtient

$$g_i(\mathbf{x}) = -\frac{1}{2}\left(-2\mu_i^T\Sigma^{-1}\mathbf{x} + \mu_i\Sigma^{-1}\mu_i\right) + \log(P(\omega_i))$$

- en réorganisant, on a

$$g_i(\mathbf{x}) = \mathbf{w}_i^T\mathbf{x} + w_{i,0} \quad \text{avec } \mathbf{w}_i = \Sigma^{-1}\mu_i \text{ et } w_{i,0} = -\frac{1}{2}\mu_i\Sigma^{-1}\mu_i + \log(P(\omega_i))$$

- Les fonctions discriminantes sont toujours linéaire
- Si on suppose que les classes sont équiprobables, on a

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \mu_i)^T\Sigma^{-1}(\mathbf{x} - \mu_i)$$

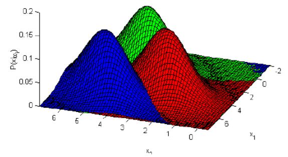
et le classifieur devient un classifieur basé sur la distance de Mahalanobis minimum à la moyenne

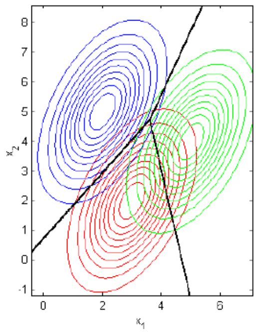
Cas 3 : $\Sigma_i = \Sigma$ avec Σ quelconque - Exemple

- illustration pour un cas à trois classes et 2 dimensions :

$\mu_1 = (3\ 2)^T \ \mu_2 = (5\ 4)^T \ \mu_3 = (2\ 5)^T$

et $\Sigma = \begin{pmatrix} 1 & 0.7 \\ 0.7 & 2 \end{pmatrix}$





Cas 4 : $\Sigma_i = \sigma_i^2 I$

- Chaque classe a une matrice de covariance différente mais proportionnelle à la matrice identité
- La fonction discriminante devient

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma_i^2}(\mathbf{x}-\mu_i)^T(\mathbf{x}-\mu_i) - \frac{1}{2}N\log(\sigma_i^2) + \log(P(\omega_i))$$

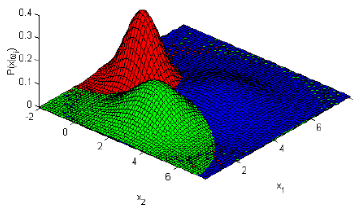
- elle ne peut pas etre simplifier ultérieurement.
- Les frontières de décision sont des fonctions quadriques : des hyper-ellipses

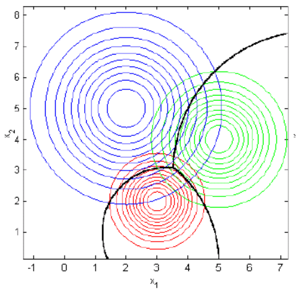
Cas 4 : $\Sigma_i = \sigma_i^2 I$ -Exemple

- illustration pour un cas à trois classes et 2 dimensions :

$\mu_1 = (3\ 2)^T \ \mu_2 = (5\ 4)^T \ \mu_3 = (2\ 5)^T$

$\Sigma_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ et $\Sigma_3 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$





cas 5 : le cas général $\Sigma_i \neq \Sigma_j$

- la fonction discrimante est

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)\Sigma_i^{-1}(\mathbf{x}-\mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i))$$

- que l'on peut ré-organiser sous une forme quadratique

$$g_i(\mathbf{x}) = \mathbf{x}^T\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^T\mathbf{x} + w_{i,0}$$

avec

$$\begin{aligned} \mathbf{W}_i &= -\tfrac{1}{2}\Sigma_i^{-1} \\ \mathbf{w}_i &= \Sigma_i^{-1}\mu_i \\ w_{i,0} &= -\tfrac{1}{2}\mu_i^T\Sigma_i^{-1}\mu_i - \tfrac{1}{2}\log(|\Sigma_i|) + \log(P(\omega_i)) \end{aligned}$$

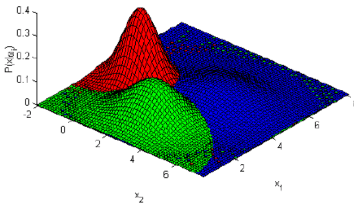
- les frontières de décision sont des quadriques

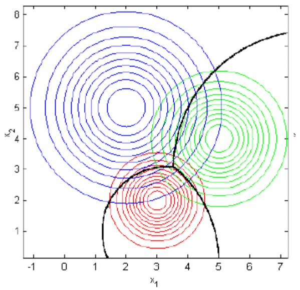
Cas 5 : $\Sigma_i \neq \Sigma_j$ -Exemple

- illustration pour un cas à trois classes et 2 dimensions :

$$\mu_1 = (3\ 2)^T\ \ \mu_2 = (5\ 4)^T\ \ \mu_3 = (2\ 5)^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix},\ \Sigma_2 = \begin{pmatrix} 1 & -1 \\ -1 & 7 \end{pmatrix} \text{ et}$$
$$\Sigma_3 = \begin{pmatrix} 0.5 & 0.5 \\ 0.5 & 3 \end{pmatrix}$$





Exemple numérique

- Calculer la fonction de décision linéaire du problème à 2 classes définis par des lois normales

$$\mu_1 = (0\ 0\ 0)^T\ \ \mu_2 = (1\ 1\ 1)^T\quad \Sigma_1 = \Sigma_2 = \begin{pmatrix} 1/4 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/4 \end{pmatrix}$$

avec $P(\omega_1) = 2P(\omega_2)$

$$g_i(x) = -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log P(\omega_i) = -\frac{1}{2}\begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z-\mu_z \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \\ z--\mu_z \end{bmatrix} + \log P(\omega_i)$$

$$g_1(x) = -\frac{1}{2}\begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-0 \\ y-0 \\ z-0 \end{bmatrix} + \log\frac{1}{3};\quad g_2(x) = -\frac{1}{2}\begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix}^T \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} x-1 \\ y-1 \\ z-1 \end{bmatrix} + \log\frac{2}{3}$$

$$g_1(x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} g_2(x) \Rightarrow -2(x^2+y^2+z^2) + \log\frac{1}{3} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} -2\left((x-1)^2+(y-1)^2+(z-1)^2\right) + \log\frac{2}{3}$$

$$\boxed{x+y+z \underset{\omega_1}{\overset{\omega_2}{\gtrless}} \frac{6-\log 2}{4} = 1.32}$$

- catégoriser l'exemple $x = [0.1\ 0.7\ 0.8]^T$

Conclusions

- pour des classes suivant une loi normale, en général la fonction de décision est quadratique
- pour des classes suivant une loi normale de matrice de covariance identique, la fonction de décision est linéaire
- des classifieurs très simples comme le classifieur à distance à la moyenne minimum est optimale dans certaines situations.
- Les hypothèses pour lesquels ces classifieurs sont optimales sont parfois difficile à vérifier. La question que l'on se pose alors est : " est ce que ce classifieur résoud mon problème ?"