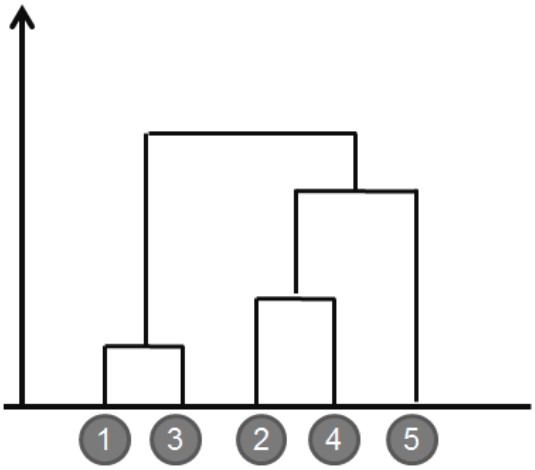
**3.** Représenter le **dendrogramme** obtenu **[1pt]** :




## Exercice2 : Séparateur à Vaste Marge SVM [4pts]

**1.** Proposer une démarche **mathématique** pour obtenir la **meilleure séparation linéaire sans individus mal classés**, en se basant sur le schéma ci-contre : **[2pts]**

$\overrightarrow{x_2} = \overrightarrow{x_1} + \vec{d}$ et $\overrightarrow{x_2} - \overrightarrow{x_1} = \vec{d}$

$w.(\overrightarrow{x_2} - \overrightarrow{x_1}) = w.\vec{d}$

$or\ wx_2 + b = +1\ et\ wx_1 + b = -1$

$donc$

$wx_2 + b - wx_1 - b = +1 - (-1) = 2$

$wx_2 - wx_1 = 2$

$w(x_2 - x_1) = 2$

$2 = w.\vec{d}$

$\|2\| = \|w\|.\|\overrightarrow{d}\|$

$2 = \|w\|.d$

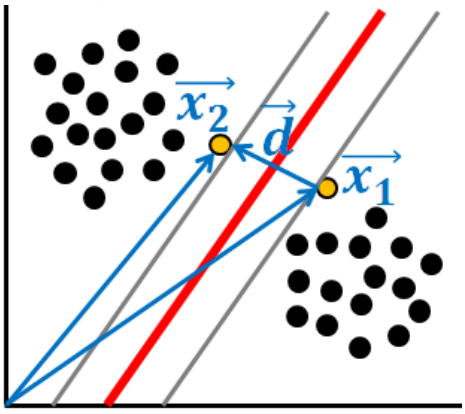
$$d = \frac{2}{\|w\|}$$

Classer correctement :

$$y_i\,(wx_i + b) \geq +1\ \forall i$$

**Maximiser la marge :**

$$minimiser\ \frac{\|w\|^2}{2}$$




**2.** En déduire la formulation SVM dans la **présence d'individus mal classés**. (possibilité de s'appuyer sur un schéma) **[1pt]**

Classer correctement :

$$y_i\,(wx_i + b) \geq +1\ - \varepsilon_i\ \ \forall i, \varepsilon_i$$

Maximiser la marge :

$$\frac{1}{2}\,\|w\|^2 + C.\sum_1^N \varepsilon_i$$

**3.** Pour remédier au problème de classes non séparables linéairement, SVM offre d'autres fonctions séparatrices. Citer deux exemples de **séparateurs Non Linéaires**, expliciter leurs **formules**. **[1pt]**

- Polynomiale : $K(x_i,x_j) = (1 + x_i{}^T x_j)^p$
- Gaussienne :

$$K(\mathbf{x_i}, \mathbf{x_j}) = \exp\left(-\frac{\|\mathbf{x_i} - \mathbf{x_j}\|^2}{2\sigma^2}\right)$$

## Exercice3 : Régression [4pts]

1. La régression linéaire permet d'étudier les individus atypiques. Décrire les différents cas possibles d'anormalité. **[1pt]**

Atypique aberrant : valeur inhabituelle sur une variable ou combinaison de variables

Influent : pèse de manière exagérée dans la régression

Atypique régression : cible très mal expliquée, erreur très élevée

2. Quelles sont les méthodes pas à pas pour la construction de modèles réduits ? Donner brièvement l'idée de chacune. **[1pt]**

- méthode descendante ou élimination en arrière lorsqu'on élimine des variables
- méthode ascendante ou sélection en avant lorsque on ajoute des variables
- méthode mixte est une combinaison de ces deux méthodes

3. Expliciter une méthode pour l'obtention des coefficients estimateurs dans la régression linéaire. **[1pt]**

$Y = X\,a \Rightarrow {}^tX.Y = {}^tX.\,X.\,a \Rightarrow [{}^tX.\,X]^{-1}.{}^tX.Y = a$

4. Peut-on avoir le même modèle pour prédire une variable qualitative ? Justifier votre réponse en s'appuyant sur des arguments mathématiques. **[1pt]**

On peut construire un modèle sous forme de coefficients classifieurs si on désire prédire une variable qualitative binaire, et ce en se basant sur un modèle de régression qui prédit la probabilité d'une modalité. Cependant, prédire des probabilités risque de donner un modèle absurde car on se limite à l'intervalle [0,1], d'où le besoin d'une fonction bijective logit = ln (P/(1-P)) qui associe à chaque valeur de probabilité une seule valeur dans]-∞, +∞ [ et inversement

## Exercice4 : Règles Associatives – Apriori [3pts]

On considère les données suivantes :

| *Id-ticket* | *Transaction* |
|---|---|
| Ticket1 | coca, lait, pain, pâtes, œufs, café |
| Ticket2 | pâtes, pain, coca, thé, riz, tomate, fromage |
| Ticket3 | olive, pain, poulet, pâtes, farine |
| Ticket4 | viande, coca, jus, poivron |
| Ticket5 | pain, poisson, laitue, pâtes, yaourt |

Comment juger la mesure d'intérêt de la règle associative suivante : **pain, pâtes → coca**

Support (pain, coca, pâtes) = 2/5

Confiance (pain, pâtes → coca) = (2/5)/(4/5) = 2/4

Lift(pain, pâtes → coca) = (2/4)/(3/5) = 5/6 <1 corrélation négative => règle non intéressante

## Exercice5 : Arbre de Décision [4pts]

Une banque souhaite promouvoir une offre commerciale via les adresses mails de ses clients.

Pour cela, elle fait appel à vous et à vos connaissances en fouille de donnée pour sélectionner ceux qui sont potentiellement intéressés. Trois attributs descriptifs sont à votre disposition :

- ✓ L'âge en deux tranches : [18; 35] et [36 et plus]
- ✓ Le sexe H : Homme ou F : Femme
- ✓ Propriétaire O : oui ou N : non

L'attribut cible qui prend deux valeurs : O (intéressé) et N (non intéressé).

Le résultat d'une enquête préliminaire sur un échantillon représentatif de clients donne :

| Age | Sexe | Propriétaire | Intéressé |
|---|---|---|---|
| 20 | H | N | N |
| 25 | F | N | N |
| 32 | H | O | O |
| 34 | H | O | O |
| 37 | H | N | O |
| 41 | F | O | N |
| 45 | H | O | O |
| 45 | F | O | N |
| 52 | H | O | N |
| 60 | F | O | N |

En utilisant l'algorithme de **Cart** avec le gain en information basé sur **l'indice de Gini**, construire un arbre de décision sur ces données. (## pour le calcul et ## pour le schéma de l'arbre de décision)

| | |
|---|---|
| IG [age, 18 ;35]=1-((2/4)² + (2/4)²) = 0,5<br>IG [age, 36 ; +]= 1- ((2/6)² + (4/6)²) = 0,44<br>IG[age] = 0,94 | IG[sexe, H]= 1- ((4/6)² + (2/6)²) =0,44<br>IG[sexe, F]= 1-((0/4)² + (4/4)²) = 0<br>IG[sexe]=0,44 |
| IG [Propriétaire, N] = 1 – ((1/3)² + (2/3)²) =0,44<br>IG [Propriétaire, O] = 1 – ((3/7)² + (4/7)²) =0,49<br>IG[Propriétaire] = 0,93 | ->La racine de l'arbre est la variable SEXE |
| IG [age, 18 ;35, sexe=H]=1-((2/3)² + (1/3)²) = 0,44<br>IG [age, 36; +, sexe=H]=1-((2/3)² + (1/3)²) = 0,44<br>IG[age, sexe=H]=0,888 | IG [Propriétaire, N, sexe=H] = 1 – ((1/2)² + (1/2)²) =0,5<br>IG [Propriétaire, O, sexe=H] = 1 – ((3/4)² + (1/4)²) =0,375<br>IG [Propriétaire, sexe=H] =0,875 |