Apprentissage Automatique

Dr. Abdelhalim hadjadj

Centre Université abdelhafid boulssouf Mila

Institut mathématique et informatique

Département informatique

Email : a.hadjadj@centre-univ-mila.dz

5.0 Février 2025

Table des matières

Chapitre2 : Apprentissage supervisé

1. Objectifs Spécifiques

À l'issue de ce chapitre, l'étudiant sera capable de :

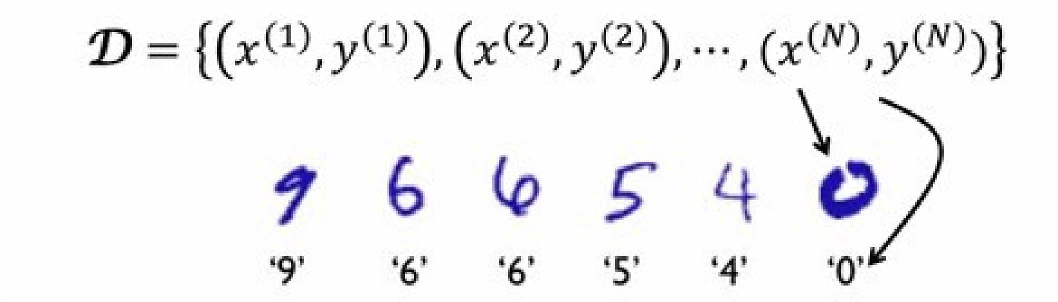
- Décrire en détail le principe de l'apprentissage supervisé, incluant les notions d'ensemble d'entraînement, d'attributs, d'étiquettes, et d'objectif de généralisation.
- Expliquer le fonctionnement et l'objectif de l'apprentissage supervisé pour les tâches de régression.
- Comprendre le concept de modèle linéaire et son application à la régression.
- Expliquer le fonctionnement et l'objectif de l'apprentissage supervisé pour les tâches de classification.
- Comprendre comment les modèles linéaires peuvent être adaptés et utilisés pour résoudre des problèmes de classification.
- Distinguer clairement les problèmes de régression et de classification et identifier le type de tâche approprié pour un scénario donné

2. Principe de l'apprentissage supervisé

2.1. Principe

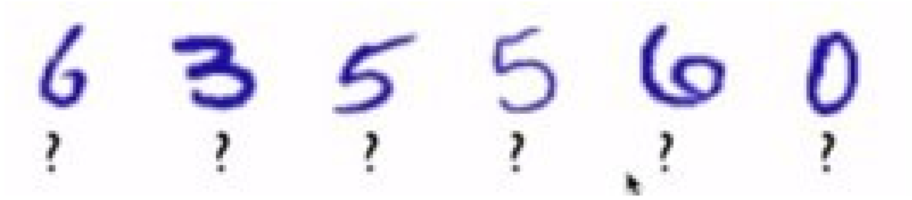
Les algorithmes d'apprentissage supervisé procèdent comme suit:

1. On fournit à l'algorithme des données d’entraînement:
   - On appelle $x(i)$ l'entrée et $y(i)$ la cible du i-iéme exemple.
   - Un élément de $D*$ est appelé exemple d'apprentissage ou une instance de données.
2. L'algorithme retourne un «programme» capable de se généraliser à de nouvelles données:



apprentissage supervisé entrainement

- On note le «programme» généré par l'algorithme d'apprentissage f(x).
- On appelle f(x) un modèle ou une hypothèse.
- On utilise souvent un ensemble de test **Dtest** pour mesurer la performance du modèle f(x).



Apprentissage supervisé test

Motivation — ***Exemple***

Des cellules cancéreuses sont prises de tumeurs de cancer du sein avant la chirurgie et elles sont photographiées.

Les tumeurs sont excisées. Les patients sont suivis pour voir s'il y a récurrence du cancer, on mesure le temps avant que la récurrence du cancer ou que le patient est déclaré sans la maladie.

Ensuite on utilise 30 caractéristiques par tumeur.

Deux variables sont prédites:

1. Résultat (R: récurrence, N: non-récurrence).
2. Temps (Jusqu'a récurrence, pour R, et en santé, pour N).

L'ensemble de données est représenté par une matrice X :

	Tumor size	texture	perimeter	....	outcome	time
	18,02	25,25	122,23		N	31
X	17,99	14,22	112,11		N	25
	26,32	31,22	142,011		R	58
			.....			

- Les colonnes de matrice sont appelées variables d'entrée, **attributs ou caractéristiques**.
- Le résultat et le temps (que nous essayons de prédire) sont appelés l**es variables résultats ou les cibles**.
- Une ligne du tableau est appelée **un exemple d'entraînement ou instance**.

2.2. Types de prédiction

- Le problème de prédiction des résultats de la maladie est appelé **une classification**.
- Le problème de prédiction du temps est appelé **régression**.
- Un exemple d'entraînement à la forme x(i), y(i) , où:

$$X^{(i)}=(X_1^{(i)}, X_2^{(i)}, ..., X_D^{(i)})$$

D est le nombre d'attributs (dans notre cas 30).

- L'ensemble d'entrée D contient N exemples.
- On dénote par X l'espace des variables d'entrées (ex. $\mathbb{R}$D)
- On dénote par Y l'espace des variables de sortie (ex. $\mathbb{R}$)

La fonction hypothèse — ***Définition***

Ayant un ensemble de données D ⊂ X × Y, trouver une fonction

f: X →Y

- Telle que f est un bon prédicteur de la valeur de y. La fonction f est appelée une hypothèse.
- Les problèmes sont classés par type de domaine de sortie:
    1. Si Y= $\mathbb{R}*$, on parle alors de régression.
    2. Si Y est un ensemble discret fini, on parle de classification.
    3. Si Y a 2 éléments, on parle de classification binaire.

2.3. Étapes de résolution d'un problème d'apprentissage supervisé

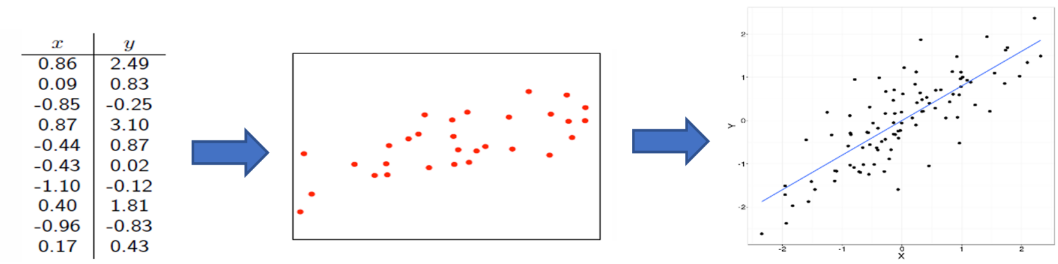
- Identifier les variables d'entrée et de sortie.
- Déterminer le **codage des variables** d'entrée (X) et de sortie (Y).
- Sélectionner **la classe d'hypothèses** ou les représentations possibles (H).
- Trouver l'hypothèse **f optimale** pour effectuer la prédiction.

3. Apprentissage supervisé pour la régression

3.1. Principe de la régression

La régression ***Définition***

Méthode statistique pour modéliser la relation entre variables qui permet de prédire une variable dépendante (Y) à partir de variable(s) indépendante(s) (X)



x	y
0.86	2.49
0.09	0.83
-0.85	-0.25
0.87	3.10
-0.44	0.87
-0.43	0.02
-1.10	-0.12
0.40	1.81
-0.96	-0.83
0.17	0.43

Exemple de régression linéaire

En termes algébriques, le modèle serait défini comme suit :

$$Y = Wx + b$$

Ou :

- Y: est la valeur que nous voulons prédire.
- W: est la pente de la droite.
- x: correspond à notre valeur d'entrée.
- b: est l'ordonnée à l'origine.

L'hypothèse d'une régression linéaire s'écrit sous la forme suivante :

Pour faciliter l'écriture, on introduit un attribut supplémentaire x_0=1, ce qui permet de réécrire la formule sous forme de produit scalaire :

$$y = f(x) = \sum_{d=0}^{D} w_d x_d = W^{\sim T} x^{\sim}$$

L'objectif de la régression linéaire est de trouver les paramètres w qui permettent au modèle f(x) de prédire des valeurs y aussi proches que possible des valeurs réelles.

Pour ajuster correctement w, nous devons :

1. Définir une fonction d'erreur ou de perte qui mesure l'écart entre les prédictions du modèle et les valeurs réelles.
2. Minimiser cette fonction d'erreur afin de trouver les meilleurs paramètres w.

3.2. Erreur des moindres carrés

- Une fonction d'erreur couramment utilisée est l'erreur quadratique moyenne (*MSE*∗), aussi appelée erreur des moindres carrés
- Essayer de rendre f(x) très proche des valeurs des y dans tous les exemples d'apprentissage dans D,

$$E(W^{\sim})=\frac{1}{2N}\sum_{i=1}^{N}(f(x)^{(i)}-y^{(i)})^2$$

Ou :

- N est le nombre d'exemples d'entraînement,
- y_i est la valeur réelle associée à 'entrée x_i
- $f(x_i)=w^T x$ est la prédiction du modèle.

Donc, L'objectif est de trouver w qui minimise cette erreur.

3.3. Méthodes d'optimisation des paramètres w

En notation matricielle, l'erreur quadratique moyenne est écrite comme :

$$E(x)=\frac{1}{2N}(Xw-y)^t(Xw-y)$$

Sa dérivée par rapport à w est donnée par :

$$\nabla E(x)=X^t(Xw-y)$$

En utilisant la formule de Taylor, on obtient :

$$\nabla E(w)=2X^t Xw-2X^t y$$

En posant ∇E(w)=0 pour minimiser l'erreur, on obtient :

$$X^t Xw=X^t y$$

Si la matrice XTXest inversible (ce qui signifie que les colonnes de X sont **linéairement indépendantes**), on peut isoler w :

$$w=(X^t X)^{-1} X^t y$$

Exemple

Calcule les paramètres optimaux

x	y
0,86	2,49
0,09	0,83
-0,85	-0,25
0,87	3,1
-0,44	0,87
-0,43	0,02
-1,1	-0,12
0,4	1,81
-0,96	-0,83
0,17	0,43

x=

0,86	1
0,09	1
-0,85	2
0,87	3
-0,44	4
-0,43	5
-1,1	6
0,4	7
-0,96	8
0,17	9

$X^TX=$

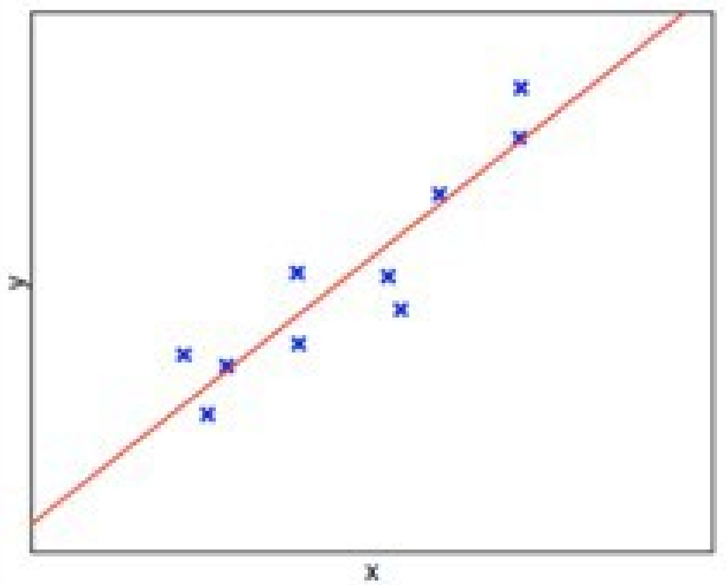
4,95	-1,39
-1,39	10

$X^ty=$

6,49
85,34

$(X^TX)^{-1}X^ty=$

1,6
1,05



Résultat de l'exemple de régression linéaire

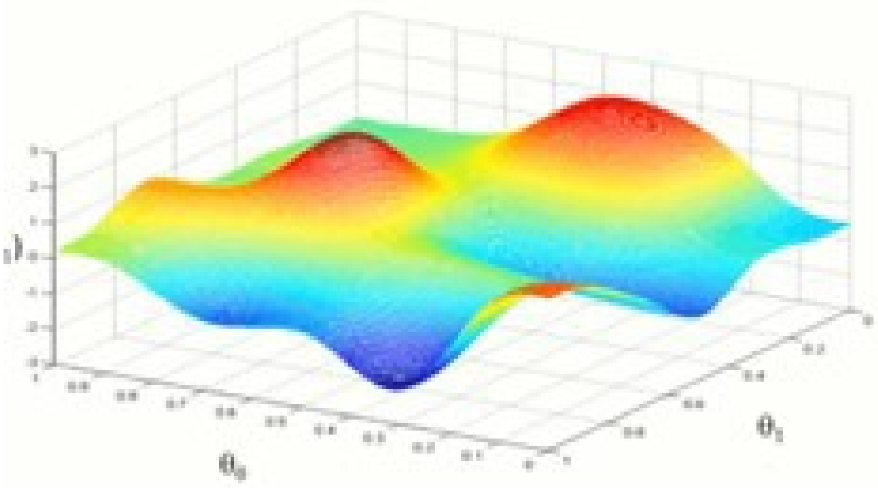
3.4. La descente de gradient

La descente de gradient

Définition

*La descente de gradient** est une méthode d'optimisation utilisée pour minimiser une fonction, généralement appelée fonction objectif ou fonction de coût .

Le gradient donne la direction (vecteur) ayant le taux d'accroissement de la fonction le plus élevé



Descente de gradient

L'idée principale derrière la descente de gradient repose sur deux concepts fondamentaux :

1. Il s'agit d'un vecteur indiquant la direction de la plus forte augmentation d'une fonction.
2. Mise à jour des paramètres : Les paramètres du modèle sont ajustés de manière itérative dans la direction opposée au gradient, ce qui permet de minimiser la fonction de coût.

$$\theta := \theta - \alpha * \nabla J(\theta)$$

θ représente les paramètres du modèle,

α est le taux d'apprentissage (learning rate), qui contrôle la taille des pas lors de la descente,

∇*J(θ) est le gradient de la fonction de coût par rapport aux paramètres.

3.5. Modèles avec plusieurs caractéristiques

- Bien que l'exemple de cette section n'utilise qu'une seule caractéristique, un modèle plus sophistiqué peut reposer sur plusieurs caractéristiques, chacune étant associée à une pondération distincte ($w_1, w_2, \ldots, w_n$).
- un modèle basé sur **n caractéristiques** peut être représenté sous la forme générale suivante :

$$y = f(x) = w_0 + w_1 x + \ldots + w_n x_n$$

La prédiction de la consommation de carburant ***Exemple***

Dans le cas d'un modèle de prédiction de la consommation de carburant, les caractéristiques utilisées peuvent inclure : Cylindrée (x_1), Accélération (x_2), Nombre de cylindres (x_3), Cheval-vapeur (x_4)

L'équation du modèle prend alors la forme :

consommation=w_0+w_1(Cylindrée)+w_2(Accélération)+w_3(Nombre de cylindres)+w_4(Cheval-vapeur)

Complément

Nous pouvons reformuler ce modèle sous forme matricielle en définissant une matrice de conception X qui contient les différentes puissances des observations x(i),où chaque ligne correspond à une observation, et chaque colonne représente une puissance de x.

La régression quadratique — *Exemple*

Pour notre exemple, une régression quadratique (m=2) aura la forme:

x=

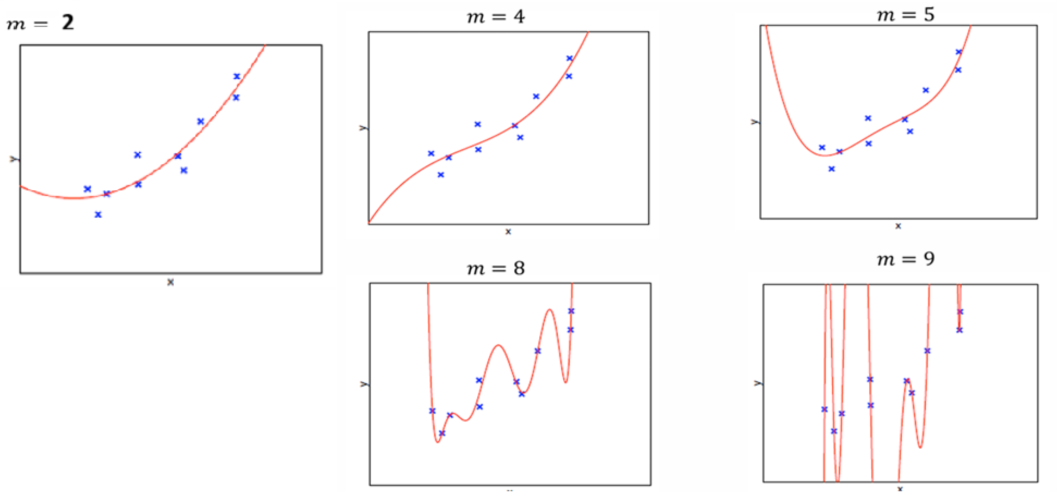
0,7396	0,86	1
0,0081	0,09	1
0,7225	-0,85	1
0,7569	0,87	1
0,1936	-0,44	1
0,1849	-0,43	1
1,21	-1,1	1
0,16	0,4	1
0,9216	-0,96	1
0,0289	0,17	1
0,7396	0,86	1

y=

y
2,49
0,83
-0,25
3,1
0,87
0,02
-0,12
1,81
-0,83
0,43

$(X^TX)^{-1}X^ty=$

3,6
1,74
0,73



Exemple des résultats de modèle régression linéaire ordre superieur

3.6. Sur apprentissage (Overfitting)

Le sur apprentissage — *Définition*

Le sur apprentissage est un défi majeur dans les techniques d'apprentissage automatique représente par un modèle peut apprendre trop spécifiquement les données d'entraînement, obtenant de très bonnes performances sur ces données, mais échouant à bien généraliser sur de nouvelles données.

3.7. Exercice : Auto-évaluation

a) Exercice

L'apprentissage supervisé repose sur

- ☐ Des données d'entraînement étiquetées pour permettre l'apprentissage d'un modèle
- ☐ L'auto-apprentissage des systèmes sans intervention humaine
- ☐ Des algorithmes d'apprentissage qui ne nécessitent aucune supervision
- ☐ L'utilisation de règles pré-définies pour ajuster un modèle

b) Exercice

La prédiction de tendances dans le domaine médical est souvent basée sur :

- ☐ Des modèles de régression statistique
- ☐ Des systèmes de classification supervisée
- ☐ Des réseaux neuronaux profonds

c) Exercice

Calculer les paramètres de la régression linéaire qui permet de modéliser ces données. x={0,86 0,09 -0,85 0,87 -0,44 -0,43 -1,1 0,4 -0,9 0,1 } y={ 2,49 0,83 -0,25 3,1 0,87 0,02 -0,12 1,81 -0,83 0,43}

4. Modèles linéaires pour la classification

4.1. Théorie de la décision pour la classification

a) Fonction discriminante

Définition

Dans un problème de classification, nous avons plusieurs classes C_1,C_2,...,C_K et notre but est d'attribuer une entrée x à l'une de ces classes. Fonction discriminante prend une entrée x et lui attribue la classe qui lui correspond parmi K classes. Elle pour but : Prédire la valeur cible y à partir de l'entrée x.

Image de rayon-X ***Exemple***

Imaginons une *image de rayon-X* . Notre tâche est de prédire si l'image montre la présence d'une malaIdie (classe C_1) ou son absence (classe C_2).

b) La théorie des probabilités pour la décision

- *La théorie des probabilités*∗ nous aide à traiter l'incertitude associée à un phénomène aléatoire, comme la classification d'un objet dans une certaine classe. Elle nous permet de quantifier cette incertitude et de prendre des décisions sur cette base. Elle quantifie les chances qu'un exemple **x** appartienne à une classe donnée $\mathbf{C_k}$.
- Lors de la classification d'une image, la question est : **À quelle classe appartient x** ?

c) Calcul de la Probabilité : Théorème de Bayes

La probabilité d'une classe C_k pour une donnée x est donnée par la formule de Bayes :

$$p(C_k|x)=\frac{p(x|C_k)\,p(C_k)}{p(x)}$$

- p(x|C_k) probabilité a posteriori, qui nous donne la probabilité que x appartienne à la classe C_k, après avoir vu x.
- P(C_k)est la probabilité a priori d'observer la classe C_k.
- p(x) est la probabilité de l'observation x.

d) Diviser l'Espace des Données

La règle minimisant l'erreur de classification consiste à diviser l'espace des données D en différentes **régions de décision**. Chaque région correspond à une classe donnée, et un exemple x sera classé dans la région à laquelle il appartient. Chaque région R_k correspond à une classe C_k, une région R1 correspond à la classe C1, et ainsi de suite.

e) Minimiser l'Erreur de Classification

Erreur de classification — ***Définition***

L'erreur se produit lorsqu'une donnée qui appartient à C1 est assignée à C2, ou vice versa.

Calcul de l'erreur

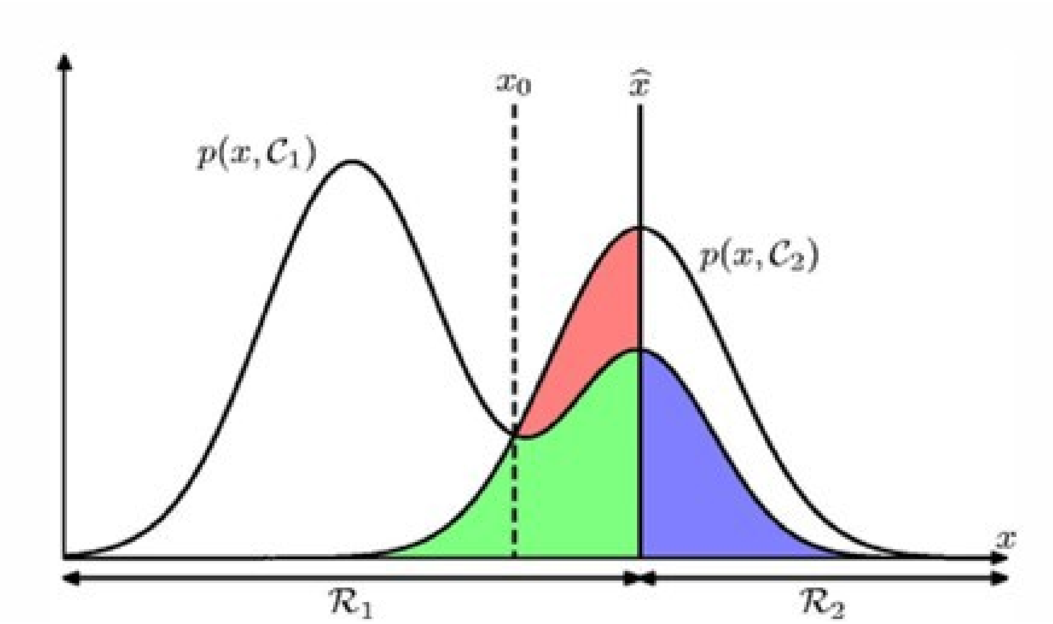
L'erreur totale est la somme des probabilités des erreurs commises dans chaque région de décision :

$$Erreyr = p(x \in R_1, C_2) + p(x \in R_2, C_1) = \int_{R1} p(x, c2)dx + \int_{R2} p(x, c1)dx$$

Cette somme représente **la probabilité d'attribuer une mauvaise classe à** x.

f) Règle de Décision pour Minimiser l'Erreur

- Afin de minimiser l'erreur de classification, nous devons assigner chaque exemple x à la classe C_k ayant la probabilité a posteriori la plus élevée $p(C_k|x)$
- Puisque p(x) est un facteur constant entre toutes les classes, le critère de minimisation revient à sélectionner la classe C_k qui maximise $p(C_k|x)$



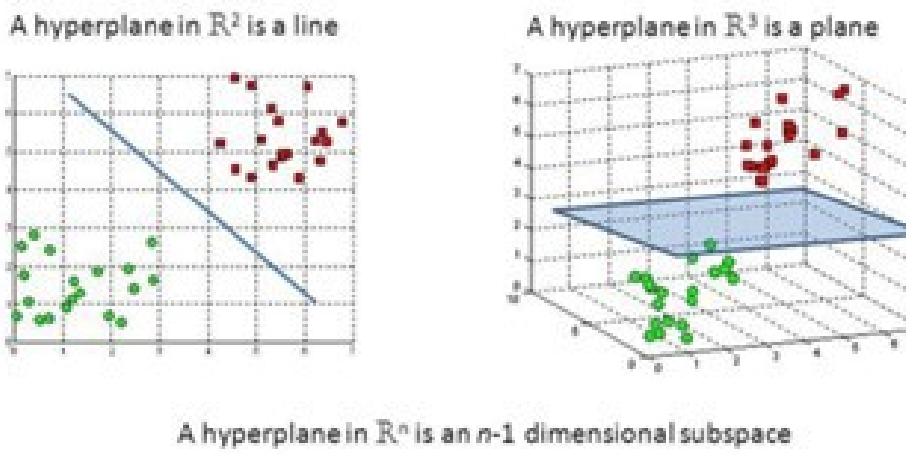
Erreur de classification

4.2. Modèles linéaires pour la classification

a) La frontière linéaire

- Pour la classification, on possède K classes $\{C_1, C_2, ..., C_K\}$. Dans la plupart des scénarios, les classes sont disjointes.
- L'espace d'entré est alors divisé en régions séparées par des frontières (ou surface) de décision.
- Dans un problème de classification, l'objectif est de prendre une entrée x et de lui attribuer une classe parmi un ensemble de K classes existantes. Cette tâche est effectuée à l'aide d'une ***fonction discriminante**** qui attribue la classe appropriée à x.

- Pour les modèles linéaire pour la classification, la frontière de décision est une fonction linéaire de la variable d'entrée x.
- La frontière linéaire est définie dans l'espace à $(D-1)$ dimensions dans l'espace de données à D dimensions.
  - Pour D = 2, la frontière sera une droite.
  - Pour D = 3, la frontière sera un plan.
  - Pour D > 3, la frontière sera appelée *hyperplan*.



La frontière linéaire

b) Régression vs. Classification

Dans la régression, la variable cible y prend des valeurs réelles continues.

Dans la classification, la variable cible y représente des classes discrètes. Pour un problème de classification binaire, y peut prendre les valeurs 1 ou 0 :

- y=1 pour la classe C1
- y=0 pour la classe C2

c) Frontière de décision linéaire: codage pour K>2

Pour des problèmes de classification avec plus de deux classes (K>2), un codage vectoriel est utilisé.

Soit $y=(y_1,y_2,\ldots,y_K)$, où :

- Si $x \in C_k$, alors $y_k=1$
- Pour toutes les autres classes $j \neq k$, $y_j=0$

Cela permet de représenter de manière binaire la classe à laquelle appartient x.

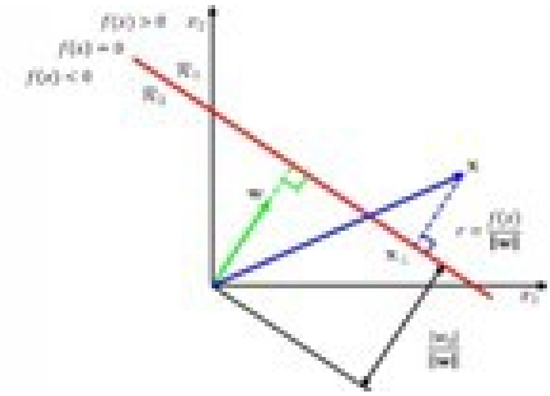
Exemple de Codage avec un Vecteur — ***Exemple***

Prenons un exemple où K=5, pour un problème de classification avec cinq classes différentes. Les données et leur codage respectif seraient les suivants :

Donnée(X)	Classe	Codage(Y)
	C1	(1,0,0,0,0)
	C2	(0,1,0,0,0)
	C3	(0,0,1,0,0)
	C4	(0,0,0,1,0)
	C5	(0,0,0,0,1)

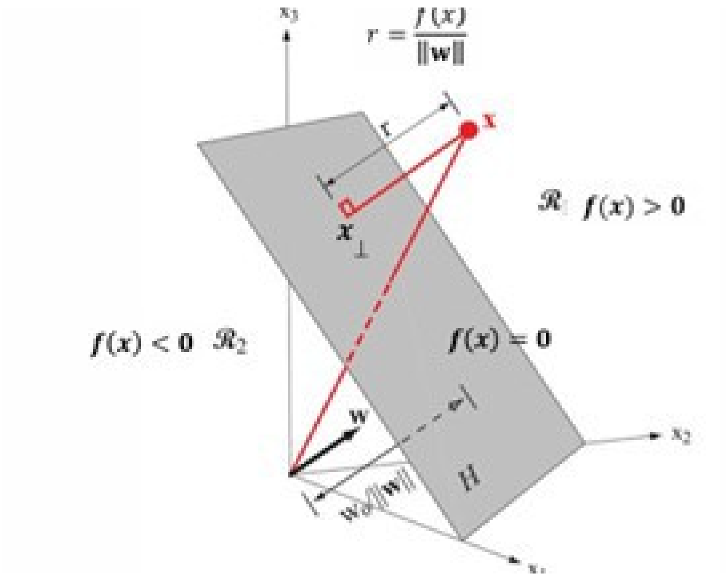
d) Cas de classification binaire (K = 2)

- Dans une **classification binaire**, la fonction discriminante est une fonction linéaire $w^T x + w_0$, qui permet de séparer les deux classes C_1 et C_2.
- ***La frontière de décision**** est définie par les points où f(x)=0, et elle sépare les deux classes.
- Le vecteur **w est orthogonal** à cette frontière de décision, ce qui signifie que la direction du vecteur w détermine **la position de la frontière**.
- Soit un problème de classification à 2 classes (classification binaire). Dans le cas de d = 2, on a la géométrie suivante:



Frontière de décision linéaire(K=2)

Dans le cas de d = 3, on a la géométrie suivante:



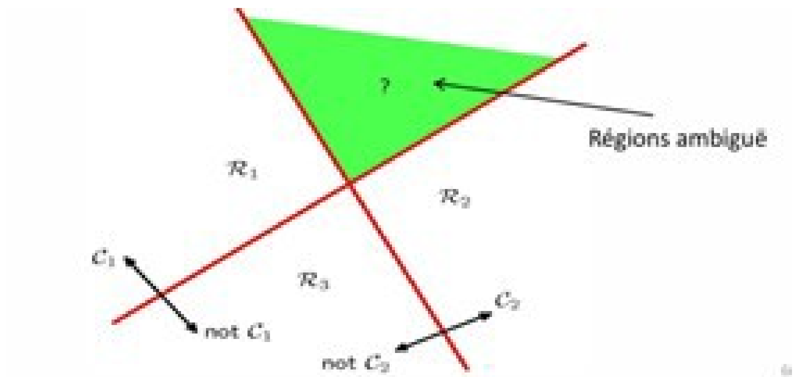
Frontière de décision linéaire(K=2)3d

e) Cas de classification multiple (K > 2)

Une des façons de construire un classificateur à K classes est de combiner plusieurs classificateurs binaires.

• Soit (K-1) classificateurs binaires chacun séparant une des classes C_k, k = 1..., K-1 (classification un-versus-tous).

Une autre alternative est de construire K(K-1)/2 classificateurs binaires, un pour chaque deux classes (classification un versus-un).



Cas de classification multiple

f) Classification par les Moindres Carrés

- Bien que la méthode des moindres carrés fournisse une solution exacte pour la classification, elle présente certaines limitations :
- Elle fonctionne bien lorsque les classes sont séparables et que les données suivent des lois gaussiennes.
- Elle n'est pas robuste aux données aberrantes.
- Les éléments de f ont une somme égale à 1, mais ils ne sont pas nécessairement des probabilités.

5. Exercice ; Auto-évaluationDivisions

5.1. Exercice

La prédiction de tendances dans le domaine médical est souvent basée sur :

- ☐ Des modèles de régression statistique
- ☐ Des systèmes de classification supervisée
- ☐ Des réseaux neuronaux profonds

5.2. Exercice

L'apprentissage supervisé repose sur

- ☐ Des données d'entraînement étiquetées pour permettre l'apprentissage d'un modèle
- ☐ L'auto-apprentissage des systèmes sans intervention humaine
- ☐ Des algorithmes d'apprentissage qui ne nécessitent aucune supervision
- ☐ L'utilisation de règles pré-définies pour ajuster un modèle

5.3. Exercice

Calculer les paramètres de la régression linéaire qui permet de modéliser ces données. x={0,86 0,09 -0,85 0,87 -0,44 -0,43 -1,1 0,4 -0,9 0,1 } y={ 2,49 0,83 -0,25 3,1 0,87 0,02 -0,12 1,81 -0,83 0,43}

5.4. Exercice

3. Qu'est-ce qu'une "frontière de décision linéaire" dans le contexte d'un problème de classification à deux classes ?

Exercice : Test Final

1. instructions

1. Lisez attentivement chaque question avant de répondre.
2. Répondez directement sur cette feuille ou sur une feuille séparée en indiquant clairement le numéro de la question.
3. Soyez clair et concis dans vos réponses. Justifiez vos réponses lorsque cela est demandé.
4. Le barème est indicatif.

2. Exercices

2.1. Exercice : Concepts Fondamentaux et Comparaisons

a) Exercice : Cochez la meilleure réponse pour chaque question

Lequel des problèmes suivants est typiquement un problème de régression ?

- ○ détecter des transactions bancaires frauduleuses.
- ○ Regrouper des articles de presse par sujet.
- ○ estimer le prix d'une maison en fonction de sa surface.
- ○ Prédire si un client va cliquer sur une publicité (Oui/Non).

b) Exercice

La validation croisée (cross-validation) est principalement utilisée pour :

- ○ Accélérer l'entraînement des modèles.
- ○ Sélectionner les meilleurs attributs (features).
- ○ Obtenir une estimation plus fiable de la performance du modèle sur des données inconnues.
- ○ Augmenter la taille de l'ensemble d'entraînement.

c) Exercice

L'algorithme K-Moyennes (K-Means) appartient à quelle catégorie d'apprentissage ?

- ○ Supervisé - Classification
- ○ Supervisé - Régression
- ○ Non supervisé
- ○ Par renforcement

d) Exercice

Expliquez pourquoi la réduction de la dimensionnalité peut être bénéfique avant d'entraîner un modèle d'apprentissage automatique

e) Exercice

Citez les deux méthodes de réduction de dimensionnalité vues en cours

f) Exercice

Discutez brièvement de l'importance de l'étape de prétraitement des données (nettoyage, normalisation, encodage des variables catégorielles, etc.) dans un projet d'apprentissage automatique typique?

2.2. Application et Évaluation

a) Exercice

Problèmes résolus ou usages de la Régression Logistique :

- [ ] Prédire la probabilité d'appartenance à une classe.
- [ ] Classification binaire.
- [ ] Classification multi-classe (par ex. via des approches One-vs-Rest ou multinomiale).

b) Exercice

Laquelle des méthodes suivantes est une technique de réduction de dimensionnalité très courante ?

- ○ K-Moyennes (K-Means).
- ○ Machine à Vecteurs de Support (SVM).
- ○ Analyse en Composantes Principales (PCA).
- ○ Régression Logistique.

c) Exercice

(Scénario) Attribut(s) pertinent(s) pour prédire si un utilisateur aimera une vidéo :

- [ ] L'historique des notes données par l'utilisateur à d'autres vidéos.
- [ ] Le genre principal de la vidéo proposée.
- [ ] La durée moyenne de visionnage de l'utilisateur sur la plateforme.

Abréviations

D : Ensemble de données (Dataset en anglais).
MSE : Mean Squared Error(Erreur Quadratique Moyenne)
R : Ensemble des nombres réels (utilisé pour les problèmes de régression).
∇ : Symbole du gradient

Bibliographie

[4] Christopher M. Bishop Pattern Recognition and Machine Learning

Webographie

[1] https://developers.google.com/machine-learning/crash-course?hl=fr