Avant-propos

Je suis diplômée d’un master 2 Données, Apprentissage et Connaissances (DAC) à Sorbonne Université Sciences (ex-UPMC). Lors de ma dernière année de master, j’ai pris l’initiative de collecter mes connaissances et de les partager. Cet ouvrage, est le fruit des résumés de cours que j’ai suivis à l’université, des livres tels que « le Machine Learning avec Python », « Machine Learning avec Scikit Learn », Deep Learning avec TensorFlow », mes propres recherches sur internet ainsi que des cours sur Youtube et Udemy.

Knowledge is power

Le temps est d’or

Le savoir se partage

<!> Cet ouvrage est en cours de mise à jour

**Celia KHERFALLAH**

**2017 - Now**

# **Introduction au Machine Learning (ML)**

## Qu’est-ce que le ML ?

L’apprentissage automatique concerne l’extraction de connaissance à partir de données. C’est un champ de recherche qui se situe entre les statistiques, l’informatique et l’intelligence artificielle. On parle aussi d’analyse prédictive ou apprentissage statistique.

« *Etant donné une tache T est une mesure de performance P, on dit qu’un programme informatique apprend à partir de l’expérience E si les résultats obtenus sur T, mesurés par P, s’améliorent avec l’expérience E.* » TOM MITCHELL, 1997

Exemple : E-mail spam

|  |
| --- |
| Tâche T consiste à identifier parmi les nouveaux courriels ceux qui sont frauduleux. |
| L’expérience E est constituée par les données d’entrainement. |
| La mesure P doit être définie, fonction de coût / hypothèse. |

## **Pour quoi le ML ?**

De nombreux systèmes utilisent des règles en dur (figées dans le code), basées sur des décisions de type Si/Sinon ~ Règles décisionnelles ciblées à la main par un expert.

* Système de règles expert

Points négatifs :

* La modification peut impliquer la réécriture de l’ensemble du système ;
* La présence d’un expert humain ;
* Nécessite énormément de règles.

La collecte de données :

* Processus différents à chaque fois ;
* Peu onéreuse ;
* Nécessite des matériaux couteux (image médicale) ;
* Expertise rare et chère ;
* Problème d’éthique et protection de la vie privée.

## **Comprendre le but et les données**

1. Comprendre les données sur lesquelles on travaille.
2. Comprendre comment elles se relient au problème qu’on veut résoudre afin de bâtir le modèle (choix de l’algorithme).

Exemples de régression : Prédire le salaire annuel en fonction de l’éducation, de l’âge, du sexe.

Différence entre la classification et la régression

Une régression Prédire une valeur

La performance P MSE

Classification Prédire une classe

La performance P Accuracy

La différence réside dans la sortie.

## Sur ou sous apprentissage ?

|  |  |
| --- | --- |
| Contexte | |
| Si le modèle est **trop simple** alors on **ne capture pas** tous les aspects de nos données ainsi que leurs variabilités | Construire un modèle qui soit capable d’effectuer une généralisation avec l’exactitude la plus élevée possible, revient à construire un modèle parfois **très complexe** pour une quantité de données.  Il suit trop précisément les particularités du jeu d’apprentissage.  Il est incapable de généraliser |
| Evaluation | |
| Train score <= 50 % et test score <= 50 %  | training score – test score| < ɛ et score < 100 % | 99 % |
| Solution | |
| Choisir un modèle plus puissant ou de diminuer le degré de régularisation | Choisir un modèle plus simple, contraindre le modèle « Régularisation » ou ajouter plus de données. |

## **+Complexité ?**

La complexité d’un modèle est infiniment liée à la variance des éléments contenus dans votre jeu de données d’apprentissage. Plus les éléments sont variés, plus le modèle qu’on utilisera sera complexe pour éviter le sur-apprentissage.

# **Probabilité et théorie de l’information :**

**La théorie des probabilités** permet l’étude de phénomènes aléatoires et de **raisonner en présence de l’incertitude**.

**La théorie de l’information** permet de **quantifier** ces **incertitudes** avec **des lois de probabilité**.

L’informatique est un environnement relativement **déterministe et certains**. Néanmoins, le ML utilise la théorie de la l’information et les probabilités pour faire face à des **quantités incertaines et stochastiques (non déterministes).**

Il existe 3 sources d’incertitude :

1. La stochasticité inhérente ou système modélisé : Exemple mécanique quantique
2. Observabilité incomplète : Les systèmes déterministes peuvent paraître stochastique lorsque nous ne pouvons pas observer toutes les variables qui déterminent le comportement du système.
3. Modélisation incomplète : Lorsqu’un modèle doit rejeter une information qui a été observée.

En pratique, il est plus pratique d’utiliser des règles simples mais incertaine plutôt qu’une règle complexe mais certaines. Car, elles sont difficiles à développer, à entretenir, à communiquer et reste sujettes à l’échec.

**A partir de 1980, on quantifie l’incertitude à l’aide de notion de probabilité.**

|  |  |  |  |
| --- | --- | --- | --- |
| Type de probabilité | L’événement β | Expérience |  |
| Probabilité fréquentiste | A une probabilité de se produire | Répétable plusieurs fois et même conditions | Lié au taux de réalisation des événements |
| Probabilité bayésienne | A 50% de chance de se produire | Non répétable et conditions différents | On représente un degré de croyance |

Différence entre logique et probabilité :

|  |  |
| --- | --- |
| Logique | Ensemble de règle formelle pour déterminer quelles propositions sont vraies ou fausses étant donné l’hypothèse selon laquelle un autre ensemble de proposition soit vrai ou faux. |
| Probabilité | Fournit des règles formelles pour déterminer la probabilité qu’une proposition soit vraie compte tenu de la vraisemblance d’autres propositions. |

## Variable aléatoire :

Une application définie sur l’ensemble des résultats possibles d’une expérience aléatoire (Prend différentes valeurs de manières aléatoires/description des états possibles). Elle doit être couplée à une loi de probabilité qui spécifie la probabilité de chacun de ces états. Les variables aléatoires peuvent être discrètes ou continues (à densité).

Lois de probabilité :

**Une loi de probabilité** est une description de la vraisemblance selon laquelle une VA, ou un ensemble de VA, prenne chacun de ses états (valeurs possibles).

## Espérance, variance et covariance

|  |  |  |
| --- | --- | --- |
| Variable discrète | Variable continue | |
| Fonction de masse (FM) | Densité de probabilité (DP) | |
| Une loi de probabilité sur des variables discrètes peut-être décrite à l’aide d’une fonction de masse P(x = x).  Elle peut également agir sur plusieurs variables en même temps **« probabilité jointe – P(x = x, y = y)»**  **P doit vérifier :**   * Le domaine de P doit être l’ensemble des états possibles de x. | Une loi de probabilité sur des variables continues peut-être décrite à l’aide d’une densité de probabilité (DP).  Théorie de la mesure  Une DP ne donne pas la probabilité d’un état spécifique, elle donne la probabilité d’une région infinitésimale de volume  **P doit vérifier :**   * Le domaine de P doit être l’ensemble des états possibles de x. | |
| La loi de probabilité marginale : On connait la loi de P(x, y) et on souhaite savoir la loi de probabilité d’un sous ensemble de celle-ci. | | |
|  | |  |
| Loi de probabilité conditionnelle | | |
| Expérience  Une fonction f(x) par rapport à une loi de probabilité P(x) est la moyenne que f prend lorsque x est tirée de P.  L’espérance est linéaire. | | |
|  | |  |
| Variance  Donne une mesure de dispersion sur les valeurs d’une fonction d’une variable aléatoire x lorsque nous échantillons différentes valeurs de x de sa loi de probabilité. | | |
| Covariance  Donne une idée de la relation linéaire entre deux valeurs, ainsi que l’échelle de ces variables.   * Covariance nulle => variables indépendantes (pas de dépendance lin éaire). * Covariance positif => les valeurs changent bcp et sont toues deux très éloignées de leurs moyennes respectives en même temps. * Covariance négatif => une augmente et l’autre décroit (vice versa).   La covariance et l’indépendance sont des concepts liés mais distincts. Car l’indépendance est une exigence plus forte que la covariance nulle.  Il se peut que deux Vas soient dépendantes mais avec une covariance nulle. | | |

Loi de probabilités usuelles.

1. Loi normale :

Un bon choix par défaut pour deux raisons :

1. Le théorème centrale limite montre que la somme d’un grand nombre de Vas indépendantes suit approximativement une loi normale. En pratique, ce signifie que de nombreux systèmes peuvent être modélisés avec succès comme des bruits normalement distribués.
2. De toutes les lois de probabilités possibles ayant la même variance, la loi normale code la quantité maximale d’incertitude sur les nombres réels. Elle insère donc le moins de connaissances préalables dans un modèles.

La loi normale se généralise sur Rn. Cette généralisation est connue sous le nom de la loi normale multivariée.

# **Calcul numérique**

La codification des nombres sur un ordinateur numérique est indispensable pour que celui-ci puisse les stocker et les manipuler. Toutefois, le problème est qu’un nombre mathématique peut être infini, mais la représentation d’un nombre dans un ordinateur doit être faite sur un nombre prédéfini de bits.

Cela signifie que pour presque tous les nombres réels, nous commettons une erreur d’approximation lorsque nous représentons le nombre dans l’ordinateur. Ces erreurs sont dans de nombreux cas des erreurs d’arrondi et cette dernière est problématique lorsqu’elle se propage au cours de nombreuses opérations.

Une forme d’erreur d’arrondi sont : l’underflow et l’overflow.

|  |  |
| --- | --- |
| L’underflow | Se produit lorsque les chiffres proches de zéros sont arrondis à zéro |
| L’overflow | Se produit lorsque les nombre de grande magnitude sont approximés comme |

Les opérations courantes en ML comprennent l’optimisation (trouver la valeur d’un argument qui minimise ou maximise une fonction) et la résolution de systèmes d’équations linéaires.

L’évaluation d’une fonction mathématique sur un ordinateur numérique est difficile lorsque la fonction implique des nombres réels ne pouvant être représentés avec précision en utilisant une quantité finie de mémoire. Ce qui en pratique fait échouer des algorithmes qui fonctionnent en théories.

# **Modèles supervisés**

# :

* Possède **deux paramètres**: et le choix de la mesure de distance
* Algorithme simple
* Une prédiction sur un nouveau point de données, consiste à trouver les éléments qui s’en rapprochent le plus ‘ses voisins’ et faire un vote
* Utilisé pour la classification binaire et multi-tâche
* A utiliser *avant* d’envisager des techniques plus avancées
* Nécessite **un prétraitement sur les données**
* **Il n’est pas très utilisé en pratique, du fait de sa lenteur et de son incapacité à gérer de nombreuses caractéristiques**

Distance :

|  |  |
| --- | --- |
| Nombre de voisins | |
| 1 voisin | N voisins |
| Utiliser un seul voisin produit une frontière de décision qui suit de près les données d’apprentissage => modèle complexe. | Lorsque que le nombre de voisin augmente, le modèle devient plus simple, la frontière tend à se lisser.  Utiliser plus de voisin correspond à une moindre complexité du modèle. |

Avantages et inconvénients

* Donne des performances raisonnables sans à avoir à effectuer un tas d’ajustement
* Facile à comprendre, la construction du modèle est rapide
* Entraînement et inférence lent (nombre de données et/ou de caractéristiques > 100 ou )

## Les distances

Une distance vérifie les propriétés suivantes :

Une image contenant table

Description générée automatiquement

|  |  |  |
| --- | --- | --- |
| Continue | [0.512, 8435, 3545] | Euclidienne / Manhattan |
| Ordinal (catégorielle et ordonné) | [Petit, moyen, grand] | Gower distance |
|  | ] | Jaccard distance |
|  | [PC or MAX] | Simple Matching Distance |
|  |  |  |

### Distance d’euclidienne / Manhattan / Tchebychev

Une image contenant table

Description générée automatiquement

### Indice et distance de Jaccard :

L'indice et la distance de Jaccard sont deux métriques utilisées en statistiques pour comparer la similarité et la diversité entre des échantillons, constitués d'attributs binaires.

L’indice de Jaccard :

Une image contenant texte

Description générée automatiquement

La distance de Jaccard :

Une image contenant texte

Description générée automatiquement

Soit deux séquences *A* et *B*, chacune avec *n* attributs binaires (0 ou 1). On a ainsi :

 A = (a_1, a_2, ..., a_n)~

 B = (b_1, b_2, ..., b_n)~

On définit plusieurs quantités qui caractérisent les deux ensembles :

M_{11}~ Représente le nombre d'attributs qui valent 1 dans *A* et dans *B*

M_{01}~ Représente le nombre d'attributs qui valent 0 dans *A* et 1 dans *B*

M_{10}~ Représente le nombre d'attributs qui valent 1 dans *A* et 0 dans *B*

M_{00}~ Représente le nombre d'attributs qui valent 0 dans *A* et dans *B*

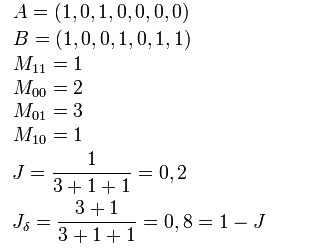
Chaque paire d'attributs doit nécessairement appartenir à l'une des quatre catégories, de telle sorte que :

M_{11} + M_{01} + M_{10} + M_{00} = n ~.

L'indice de Jaccard devient : J = {M_{11} \over M_{01} + M_{10} + M_{11}} 

La distance de Jaccard devient : J_{\delta} = {M_{01} + M_{10} \over M_{01} + M_{10} + M_{11}} 

Exemple :



### Distance de Gower

L’indice et la distance de Gower consistent à mesurer dans quelle mesure deux individus sont semblables.

L’indice de Gower varie entre 0 et 1 :

* Si l’indice vaut 1, les deux individus sont identiques
* S’il vaut 0, les deux individus considérés n’ont pas de point commun.

Soit Sg l’indice de similarité de Gower, et Dg la distance de Gower : Dg = 1 - Sg

Une image contenant texte, montre, horloge

Description générée automatiquement

représente le nombre total de caractères (ou de variables) descriptifs utilisés pour comparer les deux individus.

# Modèles linéaires

Largement utilisés en pratique, ils effectuent une prédiction en utilisant une fonction linéaire des caractéristiques présentées en entrée. Ils ont des scores de 95% en Test.

**La prédiction est une somme pondérée des paramètres, plus un terme constant (biais ou intercept).** Le biais est un paramètre, il varie tout au long de l’entrainement, une fois l’apprentissage terminé il devient bel et bien constant.

Sous-apprentissage 🡪 Modèle simple ou contraindre le modèle (régularisation)

Sur-apprentissage 🡪 modèle complexe ou diminuer le degré de régularisation

## **Avantages et inconvénients**

* Sujet au sur-apprentissage
* Difficile d’interpréter les coefficients si les caractéristiques sont corrélées.
* Rapide
* Puissant si beaucoup d'échantillons et beaucoup de dimension contrairement à une unique caractéristique
* Facile à comprendre le processus de prédiction en partant des formules

La complexité de l’inférence d’un modèle entraîné (par équation normale, décente de gradient ou …) est linéaire par rapport au nombre d’observation et au nombre de variable.

Pour entraîner un modèle, il s’agit de trouver le vecteur  **qui minimise la fonction la fonction de coût.**

## Equation Normale

Forme analytique (Formule mathématique)

L’inverse de , matrice , le nombre de variable. La complexité algorithmique est entre O () et O ().

Elle prend beaucoup de temps lorsque n >>>

Elle est linéaire par rapport au nombre d’observation du jeu d’entrainement ce qui lui permet de traiter efficacement des jeux de données >> .

L’inférence

Une image contenant texte

Description générée automatiquement

|  |  |
| --- | --- |
| Equation Normale |  |
| Une image contenant texte  Description générée automatiquement  Une image contenant texte  Description générée automatiquement |  |

## Descente de gradient

**L’idée :** corriger petit à petit les paramètres dans le but de minimiser une fonction de coût.

Si gradient = 0 Alors on atteint le minimum, meilleur quand n >>>

**Initialisation aléatoire :**

On fait décroitre la fonction de coût pas à pas jusqu’à convergence vers un minimum.

Taux d’apprentissage : Un hyperparamètre de l’algorithme d’apprentissage et non du modèle.

Ce n’est pas un paramètre qu’on cherche à optimiser pendant l’apprentissage.

Si <<< l’algorithme devra effectuer un grand nombre d’itération pour converger et prendra beaucoup de temps.

Si >>> risque de dépasser le point le plus bas et on se retrouver de l’autre côté ou plus haut qu’avant.

==> diverger l’algorithme avec des valeurs de plus en plus grandes.

|  |  |  |
| --- | --- | --- |
| Descente de gradient | | |
| DG Ordinaire (Batch) | DG Stochastique | DG par mini lot |
| Calcule le gradient de la fonction de cout par rapport à chaque paramètre du modèle (la modification de la fonction de cout lorsqu’on modifie un petit peu dérivé partielle))  = = (X.  A chaque époque le calcule se fait sur toute la base  -Lent sur des jeux de données volumineux.  Si données normalisées, taux d’apprentissage = 0.01 | Le calcul du gradient se base uniquement sur un exemple tiré aléatoirement  Moins régulier qu’une descente de gradient batch, la fonction cout effectue des sauts vers le haut et vers le bas et ne décroitra qu’en moyenne.  Les valeurs finales des paramètres sont bonnes mais pas optimales (=> solution réduire progressivement le taux d’apprentissage (grand au début pour échapper des minimas locaux puis petit à la fin pour rester dans le minima globale - **Learning Schedule**)  + Arrive à s’échapper des minimums locaux  + Converge vite  + Bon pour des gros jeux de données  + Certains exemples peuvent être vus plusieurs fois (🡺 mélange le jeu à chaque époque) | Calcule le gradient d’un sous ensemble de données sélectionnées aléatoirement.  Idem que la DS la solution est au tour de la solution optimale. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Algorithme | M (exemples) >>> | N (variables) >>> | Normalisation requise ? | Scikit - Learn |
| Equation normale | Rapide | Lent | Non | linearRegression |
| DG Batch | Lent | Rapide | Oui | - |
| DG Stochastique | Rapide | Rapide | Oui | SGDRegressor |
| DG Mini Batch | Rapide | Rapide | Oui | SGDRegressor |

## Modèle linéaire pour la régression

Dans le cas de la régression, la formule de prédiction générale pour le modèle linéaire se présente comme suit :

Produit scalaire

Sous forme vectorielle.

X = x1, x2 … (x0 = 1)

Si caractéristique, alors la prédiction est une ligne droite

Si caractéristiques, alors la prédiction est un plan

Si caractéristiques, alors la prédiction est un hyperplan

### MSE / LASSO / RIDGE

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *MSE*(LinearRegression) | *LASSO* (Least Absolute Shrinkage and Selection Operator, *Alternative du Ridge*) | *RIDGE* | | |
| La plus simple et la plus classique.  Elle trouve W et b, qui minimisent la valeur de l’erreur quadratique moyenne. | Même formule que MSE + une contrainte. | Même formule que MSE + une contrainte supplémentaire pour le choix de W.  Il faut qu’il soit le plus petit possible ~ 0. Cela signifie que chaque caractéristique devrait avoir aussi peu d’impact que possible sur la sortie *(donc présenter une pente faible)* tout en fournissant une bonne prédiction. | | |
| **Régulation : Une quantité est ajouté à la fonction coût que durant l’entraînement**  **La fonction de performance est non régularisée**   * Quantité de régularisation est réglable à l’aide du paramètre * Revient à force davantage les coefficients à se rapprocher de 0 => ce qui fait décroitre les performances sur le jeu d’apprentissage mais peut aider à obtenir une meilleure généralisation * Le modèle ressemble de plus en plus à une régression linéaire (les deux finissent par se confondre) | | | | |
| **Fonction coût** | | | | |
| Aucune régulation | Régulation L1 (Manhattan) | | Régulation L2 (Euclidienne) | |
|  | | | | |
| Aucun paramètre alpha | Paramètres : Alpha 1.0 par défaut   * Si alpha = 0 alors Ridge/Lasso = LinearRegression * Si Alpha >> alors :   + On autorise plus de caractéristique à être non nul (conseil : augmenter max\_iter : 100000).   + Permet d’avoir un modèle plus complexe. * Si Alpha << alors :   + Tous les coefficients de pondération finiront par avoir des valeurs très proches de 0 et le résultat sera une ligne horizontale passant par la moyenne des données. De plus, on perd l’effet de régularisation et on finit par retourner au problème de base *🡺 sur-apprentissage*, même résultat que celui de LR. | | |
| Aucune contrainte. | Contraindre les coefficients à se rapprocher de zéro (plus que Ridge), mais avec une régularisation L1, effectue la somme des valeurs absolues des coefficients au lieu de la moitié du carré de la norme L2 | | Contrainte de régularisation : restreindre explicitement un modèle pour éviter le sur-apprentissage.  Avec la norme L2, d’un point de vue mathématique, la méthode pénalise le carré de la norme L2, ou encore la longueur euclidienne de W. | |
| - Sur-apprentissage est probable.  + Bonne performance avec des dataset de grande taille.  +/- N’a pas d’autres paramètres ce qui est un avantage mais implique aussi qu’elle ne permet pas de contrôler la complexité du modèle. | + Idem que Ridge.  + A choisir si on a beaucoup de caractéristique et qu’on pense que certains sont inutile.  + Facile à interpréter du fait qu’elle sélectionne un sous ensemble de caractéristique. | | + Modèle restrictif, donc le sur-apprentissage est moins probable.  - En pratique, elle est la plus utilisée. | |
| LR vs Ridge | LASSO vs Ridge | | Ridge | |
| Pour des dataset **<<** 400 exemples :  La **régression linéaire** **n’apprend rien** mais le Ridge a de bon score dans le test du fait de la régularisation.  Au fur et à mesure que la taille de du dataset augmente (0-400), le Ridge s’améliore et le score de la RL sur le test décroit (1-.96).  A 400 les deux modèles ont les mêmes performances.  - La régularisation devient moins importante.  - Il devient plus difficile pour un modèle linéaire de sur-apprendre. | Avec une régularisation L1, certains coefficients valent exactement 0, cela signifie que certaines caractéristiques sont totalement ignorées par le modèle, cela peut être vu comme une sélection automatique des caractéristiques, avoir des coefficients nuls peut le rendre plus facile à interpréter et met en évidence les caractéristiques les plus importantes. | | Ridge établit un compromis entre la simplicité du modèle (coefficient proche de 0) et les performances sur le jeu de données. | |
| Ex :  From sklearn.linear\_model import LinearRegression  % jeu de donnée simple, une caractéristique  X, y =gleran.datestes.make\_wave(n\_samples = 60)  X\_train, X\_test, y\_train y\_test =  train\_test\_split(X,y, randome\_state = 42)  Lr = LinearRegression.fit( X\_train, y\_train)  Print(« w : {} ».format(lr.coef\_)) %.67  Print(« b : {} ».format(lr.intercept\_)) %.66  % pas tres bon, mais pas de grande difference  % jeu de donnée complexe, 104 caractéristiques  X, y =gleran.datestes.load\_extended\_boston()  X\_train, X\_test, y\_train y\_test =  train\_test\_split(X,y, randome\_state = 0)  Lr = LinearRegression.fit( X\_train, y\_train)  Print(« w : {} ».format(lr.coef\_)) %.95  Print(« b : {} ».format(lr.intercept\_)) %.61  % clairement un signe de sur-apprentissage, donc mieux vaux trouver un autre modèle qui contrôle sa complexité | From sklearn.linear\_model import Lasso  Lass = Lasso().fit(X\_tain, y\_train)  Print(« Training set score {} ».format(lasso.score(X\_tain, y\_train) ))  Print(« Test set score {} ».format(lasso.score(X\_test, y\_test) ))  % .29 vs .21  Print(“Number of features used {}”.format(np.sum(lasso.coef\_ != 0)))  % 4 parmi 104 | | EX :  From sklearn.linear\_model import Ridge  ridge = ridge().fit(X\_train,y\_train)  Print(« training set score : {} ».format(ride.score(X\_train,y\_train))  Print(« test set score : {} ».format(ride.score(X\_train,y\_train))  % .89 vs .75 | |

### ELASTIC NET

Mélanges des termes de régularisation de ces deux régressions, contrôlé par le ratio de mélange R.

Si R=0, régression Ridge (Ridge mieux mais si variables inutiles alors Lasso ou Elastique Net).

Si R=1, régression Lasso (Mieux que Lasso si variables >>> exemples ou variables corrélés).

### Régression polynomiale

Si sur-ajuster, alors :

* 1. Augmente la taille des données.
  2. Appliquer une régularisation l1 ou l2 (=> réduira aussi les degrés de liberté du module).
  3. Diminuer le nombre de degré de liberté.

Est-ce que tous les algorithmes de DG aboutissent-ils au même modèle si vous les laissez s’exécuter suffisamment longtemps ?

Si la fonction à optimiser est convexe et que le taux d’apprentissage n’est pas trop élevé 🡺 tous les algorithmes de DG s’approcheront du minimum global ~ modèles similaires.

Si on ne réduit pas graduellement le taux d’apprentissage, les DG stochastique et par mini-bach ne convergeront jamais (autour de la solution même si itération >>>)

Remarques :

* ***Variable\_*** : scikit-learn enregistre tout ce qui est dérivé des données d’apprentissage dans des attributs qui se terminent par un trait.
* **ELASTICNET** : Scikit-learn propose une classe ELASTICNET qui est un mélange du Ridge et Lasso. En pratique, elle fonctionne bien mais au prix d’un ajustement à effectuer entre deux paramètres (L1 et L2).

### Fonction de coût

L’erreur quadratique moyenne :

: La fonction avec les paramètres

## Modèle linéaire pour la classification

|  |
| --- |
| Estimation de la probabilité :  Fonction sigmoide [0 :1]  Retourne la logistique (la somme pondérée + intercept) |
| Entrainement et fonction de cout :  Perte logistique (Log Loss) :  N’existe pas de solution analytique pour la régression logistique.  La fonction de cout est convexe => DG est assuré de trouver un minima global |
| Prédiction :  0 si < 0.5, 1 sinon |

Forme d'intersection de pente () :

La prédiction est effectuée en utilisant la formule :

Seuillage de la valeur prédite par rapport à 0

Si la fonction est négative 🡪 classe -1

Si la fonction est positive 🡪 classe +1

* Il existe plusieurs façons d’approximer les coefficients 𝑊 et l’intercept 𝑏, ainsi que différents types de régularisation.
* La frontière de décision est une fonction linéaire de l’entrée.
* Un classifier linéaire binaire est un classifier qui sépare deux classes en utilisant une ligne, un plan ou un hyperplan.
* Comme pour la régression, les modèles linéaires peuvent sembler restrictifs dans les espaces à petites dimensions, car les frontières de décision tels que les droites, plans et hyperplans ont une flexibilité limitée. Une façon de s’y prendre un modèle linéaire plus souple consiste à étendre l’ensemble des caractéristiques en y ajoutant des interactions non linéaires, variables polynomiales, des carrés… 🡪 ajouter de nouvelle dimension
* Les modèles linéaires peuvent devenir très puissants avec des espaces à grandes de dimensions.

|  |  |
| --- | --- |
| Algorithmes | |
| Régression logistique (RL) | Support Vector Machine (SVM) |
| ≠ régression linéaire | Séparateur à vaste marge |
| `sklearn.linear\_model.LogisticRegression` | `sklearn.svm.LinearSVC` |
| * + - Modèle simple et interprétable ;     - Estime la probabilité qu’une observation appartienne à une classe particulière.   Considérons une entrée X= {, , …  }, la RL a pour objectif de trouver une fonction h telle que nous puissions calculer :  =  La fonction est une probabilité , paramétrée par {1, 2, 3, …, n} à optimiser, et que le seuil correspond au critère de classification, généralement 0.5. | * + - Pendant la phase d’entraînement, SVM apprend quelle importance chaque point de donnée a pour la représentation de la frontière de décision être les deux classes.     - Seul un sous-ensemble de données compte vraiment pour définir la frontière de décision 🡪 vecteur support     - La décision de classification est effectuée à partir des distances aux vecteurs support et de l’importance de ces vecteurs support *duel\_coef\_*     - Les SVM à noyau constituent une extension qui permet de former des modèles plus complexes (la frontière de décision n’est pas que des lignes ou des hyperplans mais plutôt une ellipse)     - L’astuce du noyau consiste à ajouter des caractéristiques non linéaires à la représentation des données d’origines pour rendre les modèles linéaires bien plus puissants. Cependant, nous ne savons souvent pas quelles caractéristiques ajouter, mais ajouter toutes les interactions possibles pourrait avoir une incidence considérable en coût de traitement. Heureusement l’astuce du noyau permet de d’entraîner un classifieur dans un espace de grandes dimensions sans calculer réellement la nouvelle représentation (qui est très volumineuse).   L’astuce du noyau consiste à calculer la distance (produit scalaire) des points de données pour la représentation étendue des caractéristiques sans jamais calculer réellement cette extension. |
|  | Hyperparamètre choix du noyau |
|  | * + - Noyau polynomiale : calcule tous les polynômes possibles jusqu’à un certain degré des caractéristiques.     - Noyau radian gaussien : considère tous les polynômes possibles à tous les degrés, mais avec une décroissance de l’importance des caractéristiques pour les degrés les plus élevés.   La distance entre les points de données est mesurée par le noyau gaussien :  Détermine la portée de l’influence d’un certain échantillon de donnée.   * + - : correspond à une grande portée. Plus le rayon du noyau est large, plus on considère de points et plus l’influence de ce sous-ensemble de points est grande.   Cela se traduit par des frontières de décision plus lisses et des modèles linéaires simples.   * + - : correspond à une portée limitée, des frontières de décision qui se concentrent sur les points individuels et des modèles linéaires plus complexes.   Par défaut |
| Paramètres communs de régularisation | |
| * LinearSVC et LogisticRegression appliquent une régularisation (par défaut). Mais, peuvent aussi appliquer une régularisation , si on souhaite réduire le nombre de caractéristiques ou que le modèle soit facilement interprétable. | |
| * + - Pénalité | * + - Pénalité |
| * LinearSVC et LogisticRegression possède un paramètre d’équilibrage ou de compromis 𝐶. * Il limite l’importance de chaque point (de la valeur duel\_coef\_ dans le cas du SVM. | |
| C >> | C << |
| * + - Ajuster le jeu d’apprentissage le mieux possible     - Le modèle met l’accent sur l’importance pour chaque point de donnée individuel d’être classifié correctement     - Moins y'a de régularisation, moins on met l'accent sur la bonne classification des points | * + - Les points exercent une influence limitée sur le modèle.     - Forte régularisation, conduit à de nombreux coefficients nuls     - Le modèle met l’accent sur le fait de trouver des coefficients 𝑊, b proches de 0   Le modèle essaye d’ajuster la majorité des points de données |
|  |  |
|  | * + - SVM est sensible au paramétrage     - SVM est sensible au calibrage des données, il exige que toutes les caractéristiques varient selon échelle similaires.   Pour vérifier les ordres de grandeurs on peut tracer la magnitude logarithmique (boite à moustaches, log(min), log(max))  Un recalibrage des données pour qu’elles se trouvent approximativement dans la même échelle (comprise entre 0 et 1) est indispensable.   * SVM est lent avec les grandes quantités de données * Difficile à interpréter et à expliquer |
|  | * + - SVM est efficace avec données possédant peu ou beaucoup de dimension     - SVM est efficace avec des caractéristiques qui ont des unités semblables (eg. Pixel) |

Régression SoftMax (exponentielle normalisée) : LogisticRegression (multi\_class = ‘multimonial’, solver = ‘lbfgs’, C=10)

Prend en compte plusieurs classes, sans avoir à entrainer plusieurs classifieurs binaires puis à les combiner (comme la stratégie OvR). Elle ne prédit qu’une seule classe (=> utilisé avec des classes mutuellement exclusives), est multi-classe mais pas multi-sortie.

Objectif : est d’avoir un modèle qui estime une forte probabilité pour la classe ciblée => minimiser la fonction de cout « entropie croisée » (pour mesurer l’écart entre deux distributions de probabilités).

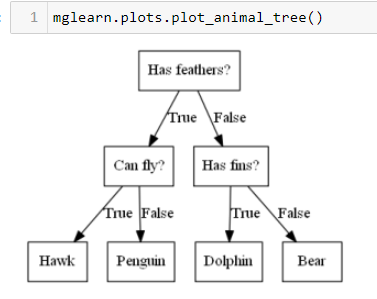
# **Classifieurs bayésiens**

* Une famille de classifieurs
* Plus rapides mais généralisation moins bonne que les classifieurs linéaires
* Ils déterminent les paramètres en analysant individuellement les caractéristiques et en élaborant leurs règles de classification à partir de ces caractéristiques individuelles
* Pour effectuer une prédiction, un point de donnée est comparé aux statistiques obtenues pour chacune des classes trouvées

|  |  |  |
| --- | --- | --- |
|  | Classification textuelle | |
|  |  |  |
| Appliquée à des données continues quelconques.  Enregistre la valeur moyenne ainsi que la déviation standard de chaque caractéristique pour chaque classe. | Appliquée a des données qui sont dénombrements de quelque chose (chaque caractéristique représente un entier qui compte une certaine propriété, comme le nombre d’occurrences d’un mot dans une phrase).  Prend en compte la valeur moyenne de chaque caractéristique pour chaque classe | Suppose que les données sont binaires. |

# Arbre de décision

Des modèles largement employés pour des taches de classification et de régression. Ils se construisent sous la forme d’une hiérarchie de question de type si/sinon qui amène à une décision.



## Construction

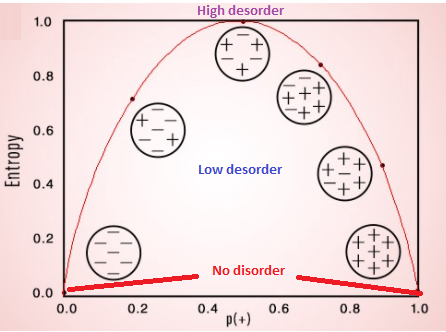
Entrainer un arbre de décision consiste à former la séquence de questions si/sinon (chaque nœud fait un *test*) qui nous conduise le plus rapidement possible à la bonne répondre.

L’index de Gini ou le gain d'information sont des algorithmes pour construire un arbre de décision, pour déterminer les nœuds et la meilleure façon de se diviser.

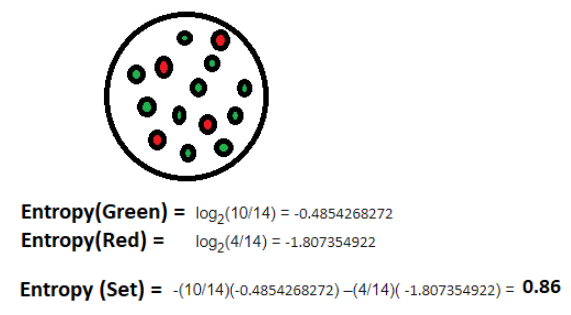
### Méthode 1 : Le gain d'information (IG)

#### L’entropie

Mesure l’homogénéité d’un nœud



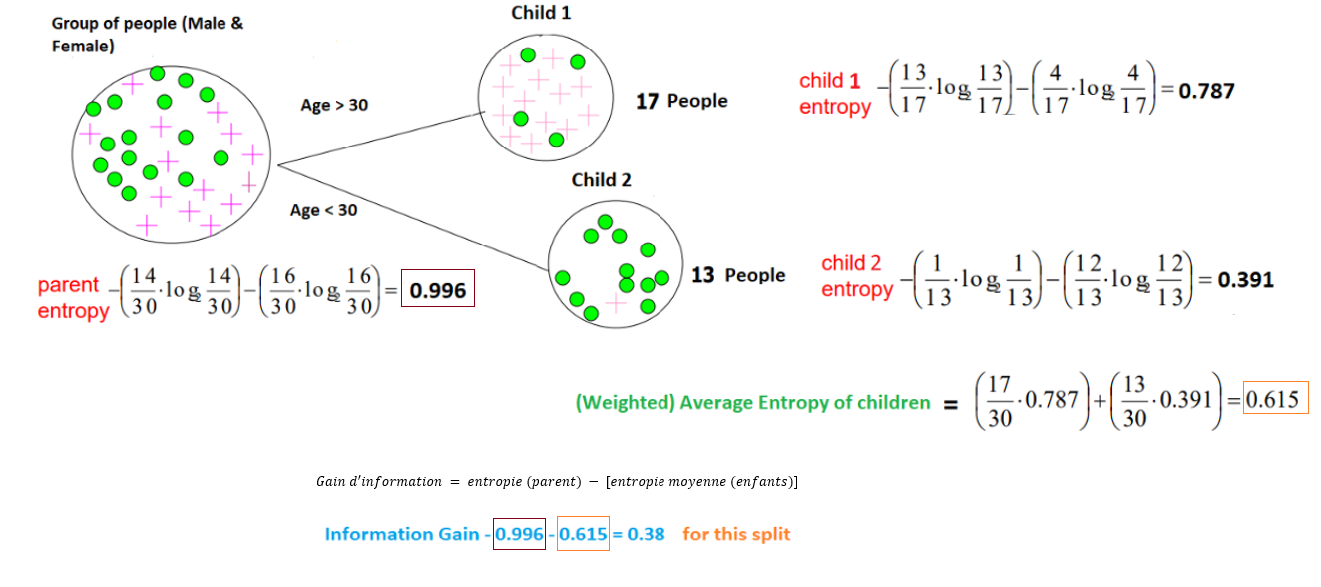
Example :



Le gain d'information (IG) mesure la quantité « d’informations » qu'une caractéristique nous donne sur la classe.

Le gain d'information (IG) est utilisé pour décider de l'ordre des attributs dans les nœuds d'un arbre de décision.

Utilisé par les algorithmes : ID3, C4.5



### Méthode 2 : L’indice de Gini

Une image contenant texte

Description générée automatiquement

* Utilisé par les
* Fait des splits binaire

## Contrôler la complexité

Construire un arbre jusqu’à ce que toutes les feuilles soient purs conduits à des modèles complexes et sujet au surapprentissage. Deux stratégies servent à éviter le surapprentissage :

1. ou pré-élagage :

Stopper en amont la création de l’arbre, par la limitation :

- De la profondeur maximale de l’arbre *‘*

- Du nombre maximal de feuilles

- Du nombre minimum de points dans un nœud pour qu’il soit partageable

1. ou post-élagage :

Construire un arbre puis supprimer ou regrouper les nœuds qui contiennent peu d’information.

Scikit-learn : utilise l’algorithme CAR, qui produit des arbres binaires. L’algothime sépare les données en deux sous-ensembles les plus purs en utilisant une seule caractéristique k et un seuil (exemple : pétal <= 20). Il minimiise la fonction de cout J(k, et ainsi de suite jusqu’à ce qu’il n’est plus de donnée a partagé ou atteindre max\_depth.

CART est un algorithme glouton, approche la meilleure solution mais pas forcément la plus optimale, c’est un problème Np-Complet O(exp(n))

Citerion = geni ou entropy (mesure l’impureté)

Modifier max\_depth va contraindre le modèle et le régulariser en cas de sur-ajustement.

|  |  |
| --- | --- |
| Gini | Entropie |
| Gini rapide à calculer, lorsqu’elle diffère. Gini à tendance à isoler la classe la plus fréquente dans une branche de l’arbre, | entropy d’un jeu de donnée est nulle lorsque les observations appartiennent à une seule classe tandis que l’entropie produit en general des arbres légèrement plus équilibré |

Bagging (avec remise) pasting (sans remise)

Bootstrap : mécanisme d’échantillonnage pour construire des arbres diversifiés afin d’améliorer les performances, variantes :

Bagging : bootsrap aggregating

Après la création de base aléatoires (avec/sans remise) …

Agrégation des prédictions.

+ réduit la variance du modèle.

FA

Un sous ensemble des attributs pris aléatoirement, le partitionnement sur les valeurs des attributs.

Extrem tree (variante des FA)

Attributs sont aléatoires et la partition sur ces valeurs est prise aléatoirement (rien à apprendre).

Boosting : combine des arbres très simples (faibles), chaque arbre essaye de corriger son prédécesseur. 2 variantes :

AdatBoost et Gradient Boosting

1. adatBoost : pondérer les exemples mal classé et … (modifie un peu les poids)
2. Gradient Boosting

Adaboost : technique d’apprentissage séquentielle ~ DG, sauf qu’ajuster les paramètres d’un seul prédicteur pour minimiser les poids d’une fonction de cout, Ada ajuste des prédicteurs à l’ensemble en les améliorant progressivement.

Corrige son prédécesseur en considérant les points qu’il a sur-ajuster, les suivant se concentrent de plus en plus sur les cas difficiles.

Arbre 1 = un arbre de décision, entrainé puis effectue des prédictions, les poids relatifs aux observations mal classées sont alors accrus.

Le second classifieur utilise mes poids modifiés, entrainé puis effectue des prédictions, poids modifiés ….

-Non parallélisable

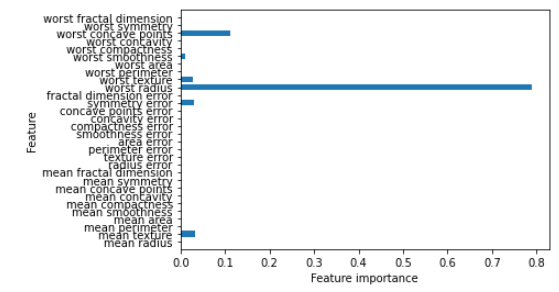
-Non adapté aux données de grande taille.

## Importances des caractéristiques

Au lieu de regarder un graphe dans sa totalité, on peut regarder l’importance [0 ; 1] des caractéristiques dans la décision à laquelle conduit un arbre.

Une image contenant texte

Description générée automatiquement



L’arbre a choisi comme étant la caractéristique la plus importante (niveau 1 de l’arbre) mais il ne dit pas si cette mesure est indicative d’un échantillon correspondant à la classe Y ou la classe X, il est possible qu’il n’existe pas de relation entre la caractéristique et la classe.

Concernant les caractérisés à faible importance, cela ne signifie pas pour autant qu’elle ne véhicule pas d’information. Cela veut dire qu’elle n’a pas été choisie par l’arbre, probablement car une autre caractéristique encode la même information.

## Forces et faiblesses

* Facile à visualiser et compris par des non-experts (arbre de taille raisonnable).
* Invariants quant au recalibrage des données, du fait que chaque caractéristique est traitée séparément. Donc pas de normalisation ou standardisation, un mélange de caractéristiques (binaire ou continues) est accepté.

Normalization typically means rescales the values into a range of [0,1].

Standardization typically means rescales data to have a mean of 0 and a standard deviation of 1 (unit variance).

* + - Malgré le pré-élagage ils tendent au sur-apprentissage, c’est pour quoi on applique des méthodes d’ensemble à la place d’arbre de décision uniques.

## **Tâche de régression**

* Effectuer une prédiction revient à traverser l’arbre en suivant le résultat des tests sur chaque nœud afin de trouver la feuille où va venir se poser le nouveau point de donnée (idem pour la classification). La sortie obtenue pour ce point est la cible moyenne des points d’entraînement pour cette feuille.
* Tous les modèles de régression basés sur les arbres ne se sont pas capable d’extrapolation (de faire des prédictions en dehors du champ des données d’apprentissage).
* On utilisera la au lieu de l’indice de Gini ou le gain d’information.

# **Ensemble d’arbre de décision**

Les ensembles sont des méthodes qui combinent de multiples modèles d’apprentissage automatique pour créer des modèles plus puissants, aussi bien pour la classification que pour la régression.

E

## Construction

1. Choisir le nombre d’arbre
2. Les arbres doivent être différents (indépendants les uns des autres), comment ?
   * On prend un **échantillon** Bootstrap (**tirage aléatoire** avec remise parmi

< ! > => **création d’un jeu de donnée de même taille**.

1. Chaque nœud **sélectionne aléatoirement** un sous-**ensemble de caractéristiques** , puis il recherche le meilleur test possible :
   * *Chaque partage peut prendre en compte* ***toutes les caractéristiques*** *du jeu de données 🡪 A****rbres très semblables, hormis l’aspect aléatoire du tirage avec remise***
   * ***Les partages n’ont pas*** *le* ***choix*** *quant à* ***la caractéristique à tester****, et que la seule possibilité est* ***d’affecter des recherches sur certains seuils de la caractéristique*** *sélectionnée de manière aléatoire*
   * ***Arbres assez semblables***
   * ***Arbres assez différents*** *et qu’ils* ***pourraient avoir besoin d’être très profond******pour*** *bien ajuster les données*

***Valeur par défaut :***

Chacun arbre commet plusieurs erreurs car certains des points ne se trouvent pas dans le jeu de d’apprentissage à cause de la méthode de tirage avec remise.

### Importances des caractéristiques

La forêt aléatoire donne une importance non nulle à bien plus de caractéristique de l’arbre seul car le mode de construction de l’algorithme le force à prendre en considération de nombreuses explications possibles car elle capture une image des données bien plus large qu’un seul arbre.

### Forces et faiblesses

* Parallélisable
* Invariants quant au recalibrage des données
  + - **Mauvais comportement** avec les **données éparses et/ou multiples dimensions** (données textuelles)
    - **Entraînement et inférence lente par rapport aux modèles linaires.**

# Gradient Boosting :

## Construction :

* **Construit des arbres** de manière **sérielle** où chaque arbre **essaye de corriger les erreurs faites par le précédent**.
* Il n’y a pas de part aléatoire mais l’utilisation d’un fort pré-élagage (arbre très peu profond 🡪 arbres simples 🡪
* Les principaux paramètres sont et (contrôle le degré de correction des erreurs faites par les arbres précédents).  Nécessite plus un plus grand nombre d’arbre

## Importances des caractéristiques

Résultat assez semblable à ceux des forêts aléatoires, à ceci près que plusieurs caractéristiques sont ignorées.

## Forces et faiblesses

* **Arbres peu profond < 5 🡪 apprentissage et prédiction plus rapide** que les FA
  + - Risque de surapprentissage
    - **Ajustement précis des paramètres**

XGBoost

LightGBM

CtBoost

# SVM à noyau (séparateur à vaste marge) :

Nous avons vu les SVM pour la classification dans les modèles linéaires. Les SVM à noyau en constituent une extension qui permet de former des modèles plus complexes qui ne sont pas simplement définis par des hyperplans dans l’espace d’entrée. Ils s’appliquent à la classification comme à la régression.

Comme vu précédemment les modèles linéaires sont assez limitatifs pour les espaces de faibles dimensions, car les lignes et les hyperplans ont une flexibilité limitée. Une façon de s’y prendre pour rendre un modèle linéaire plus souple consiste étendre l’ensemble des caractéristiques. Autrement dit, à y ajouter des caractéristiques supplémentaires/axes supplémentaires, par exemple des intersections ou des variables polynomiales.

Par rapport aux caractéristiques d’origines, le modèle linéaire n’est en fait plus linéaire. Ce n’est en effet pas une ligne, mais plutôt une ellipse (voir exemple page 99).

# **L’astuce noyau :**

Ajouter des caractéristiques non linéaires à la représentation de nos données peut rendre les modèles linéaires bien plus puissants.

Quelle caractéristique ajouter ?

Trop de possibilité…

Trop de cout de traitement...

* D’où l’astuce noyau, qui consiste à calculer directement la distance (plus précisément, les produits scalaires) des points de données pour la représentation étendue des caractéristiques, sans jamais calculer réellement cette extension.

Il existe deux manières de s’y prendre

Il considère tous les polynômes possibles à tous les degrés mais avec une décroissance de l’importance des caractéristiques pour les degrés les plus élevés.

Calcule tous les polynômes possibles jusqu'à un certain degré de caractéristiques d’origine (features1\*\*2 features2\*\*5)

SVM apprend quelle importance chaque point de donnée du jeu d’apprentissage a pour la représentation de la frontière de décision entre les deux classes. Seul un sous-ensemble de ces points compte vraiment pour définir la frontière de décision : ce sont des échantillons qui se trouve prés de la frontière entre les classes => on les appelle des vecteurs supports.

La prédiction sur un nouveau point, on mesure la distance jusqu’à chacun des vecteurs supports, et de l’importance de ces vecteurs (déterminé au cours de l’entrainement sous dual\_coef\_ de SVC).

*Ajuster les paramètres du SVM :*

|  |  |
| --- | --- |
| Détermine la portée de l’influence d’un certain échantillon de donnée. | C est un paramètre de régularisation. Il limite l’importance de chaque point (coefficient dual\_coef\_). |
| **ꝩ** Є [0.1, 10] | C Є [0.1, 1000] |
| **ꝩ <<** correspond à une grande portée.  Plus le rayon du noyau gaussien est large, et plus l’influence de chaque échantillon du jeu d’apprentissage est grande.   * Grand rayon pour le noyau gaussien. * Nombreux points considérés comme proche. * Frontière de décision lisse.   Une valeur basse, indique que la frontière va varier lentement =>   * Modèle moins complexe. | **C >>** On renforce l’influence, de ce fait :   * La frontière de décision prend une forme courbe pour classifier correctement tous les points. |
| **ꝩ >>** correspond à une portée limitée.   * Modèle plus complexe. * Une frontière qui se concentre davantage sur des points individuels. | **C<<** implique que chaque point de donnée ne peut avoir qu’une influence très limitée sur le modèle.  Les points mal classés ont une influence négligeable sur la ligne. |

Avantages/inconvénients :

+ Les SVM donnent généralement de bon résultat.

- Ils sont très sensibles au paramétrage et au réglage des données.

- En pratique**, ils exigent** que toutes les **caractéristiques** **varient selon une échelle similaire**.

**Solution**

Recalibrer chacune des caractéristiques afin qu’elles se trouvent approximativement toutes sur une même échelle. Ils existent deux méthodes courantes pour les SVM à noyau.

Prétraitement MinMaxScaler

Recalibrer les données pour qu’elles soient comprise entre 0 et 1

Forces, faiblesses et paramètres.

|  |  |
| --- | --- |
| Forces | Faiblesses |
| * Puissant. * Efficaces avec une large variété de donnée. * Permettent d’obtenir des frontières complexes. * Fonctionnent bien avec des données ayant beaucoup de dimension peu de dimension contrairement aux Modèles linéaires. * Intéressant à tester avec des pixels (même échelle). | * Ne s’adaptent pas trop bien aux grandes quantités de données (problématique en termes de temps et de consommation de mémoire si >> 100 000). * Prétraitement des données et ajustement précis des paramètres (c’est pourquoi on préfère en pratique les arbres). * Difficile d’expliquer une prédiction a des non-experts. * Corrélation forte entre les paramètres ꝩ et C. |

# **Réseau de neurones :**

1. **Fenêtre temporelle**

Ils sont vus comme une généralisation des modèles linéaires effectuant de multiples étapes de traitement pour arriver à prendre une décision.

Après calcule de la somme pondérée pour chaque unité cachée on applique au résultat une fonction non linéaire.

1. Ajuster les réseaux de neurones :

Il y’a plusieurs manières de contrôler la complexité d’un RN :

* Le nombre de couches cachées.
* Le nombre d’unités dans chaque couche cachée.
* Le paramètre de régularisation (alpha).

Attention :

Les **coefficients de pondération** sont **définis** de manière **aléatoire** **avant que l’entrainement ne débute**, et cela aléatoire **affecte** évidemment le **modèle** qui est entrainé. Ce signifie que, même **si on utilise exactement les mêmes paramètres**, nous **obtiendrons des modèles très différents** selon la manière dont le générateur de nombre aléatoires est initialisé. **Si le jeu de données est >>** et **complexité bien choisie** alors **pas** de **problème**. Si le jeu de données << alors faut garder cette notion en tête.

Les RN **s’attendent aussi que toutes les caractéristiques d’entrée** **varient de manière similaire**, et qu’elles aient dans l’idéal une **moyenne** **de 0 et une variance de 1** (StandarScalar).

Rmq : nombre de bons modèles atteignent la mémé exactitude 0.972, tous les modèles font exactement le même nombre d’erreurs, soit 4. Cela peut provenir du fait que le jeu de donnée a une taille réduite, ou que les points concernés sont réellement différents des autres.

Avec Python : MPLRegressor – MPLClassifier (ce n’est pas les plus complet).

1. Avantages et inconvénients

|  |  |
| --- | --- |
| Force | Faiblesse |
| **+** Peuvent capturer des informations contenues dans de vastes quantités de données.  **+** Bonnes données + bon réglage + temps de calcule suffisant => battent tous les autres algorithmes (Classification et régression).  **+** Construire des modèles très complexes. | * Prétraitement soigneux des données, travaillent avec des données homogène (caractéristique ont des significations similaires, comme SVM).   Si les caractéristiques sont de natures différentes, les arbres sont mieux.   * Réglage des paramètres. * Temps long pour l’entrainement. |

Estimer la complexité des réseaux de neurones

Une mesure utile pour juger de la complexité d’un RN est le nombre de poids ou de coefficients qui ont été entrainés.

Si **classification binaire** & # [**caractéristique**] = 100 & 100 **unités cachées**

=>

100 \* 100 = 10000 poids entre l’entrée et la première couche cachée + 100 \* 1 = 100 Poids entre la dernière couche cachée et la couche de sortie

…

**Une manière courante d’ajuster les paramètres** dans un RN consiste à commencer par créer un RN suffisamment **grand pour engendre un sur-apprentissage**, en s’assurer sue la tache peut être entrainé par le réseau. Puis

* Soit on **réduit** le RN,
* Soit on **augmente la valeur d’alpha** pour ajouter de la régularisation ce qui améliore les performances de généralisation.

1. Algorithme d’entrainement :

|  |  |  |
| --- | --- | --- |
| Adam | Lbfgs | Sgd |
| + Bien pour la plupart des situations.  - Sensible aux données (entre 0 et 1). | +Robuste.  -Lent avec les modèles complexes et de grande taille. | +Pour les chercheurs en depp learning.  +/-Paramètre supplémentaires à ajuster. |

1. Estimer l’incertitude pour les classifieurs binaires :

Nous ne sommes pas seulement intéressés de savoir si l’exemple appartient ou pas un une classe, mais on souhaite aussi savoir le degré de certitude du classifieur a quant à la justesse de cette prédiction. Il existe deux fonctions sous scikit-learn :

|  |  |
| --- | --- |
| Decision\_function | predict\_proba |
| Le modèle nous dit avec quelle force il croit qu’un point appartient à une classe ‘positive =1‘ ou à la classe ‘négative =0’ | Le modèle donne une probabilité pour chaque classe.  Plus facile à interpréter. |

Rmq :

La classe négative est toujours la première de l’attribut classes\_.

Un modèle est dit calibré si ses prédictions sont en accords avec la réalité.

## Grande famille de RN :

1. INSERTION IMAGE

**Points négatifs d’un RN FEEDFORWRD** :

* Nombre d’entrée/sortie fixé.
* Complexité du réseau dépend du nombre de couche et du nombre de lien entre les neurones (fixé).
* Difficile pour modéliser les séries temporelles (vidéo, musique et traduction de phrase, le réseau se souvient mal de ce qui se passe avant, ex : Je vais en France, pour apprendre le … ? français).

##### Type de RN :

* (ex : Entrée = Image, Sorties = Phrases pour décrire l’image)
* (ex : Entrées = Phrases, Sortie = Une émotion)

##### RNN :

RNN => Très profond. Une couche cachée dépend de la couche précédente, utile pour les séries temporelles.

Avec une E on une S avec une E et une S, on a une autre S

C.C

1. **Point positif/ négatif d’un RNN :**

**+** Possède une mémoire.

**-**En pratique, on a un problème d’entrainement, le gradient explose.

**- - Vanishing problem** gradient le réseau s’arrête, on pense qu’il a fini d’apprendre car la valeur de la différence du gradient est tres petite => solution **clipping** réduit les valeurs du gradient.

##### **LSTM :**

LSTM => Pour palier du problème VPGradient, avec des portes dans les neurones cachées.Rmq :

Entrées : 2 vidéos, si elles sont de tailles différentes (n’ont pas le meme nombre de pixel, Le vecteur d’Input n’ont pas la même taille) = > Problème => Solution ROGNER l’image

Si le nombre de couche caché est élevé => problème de puissance des machines. En pratique 2-3 couches caché.

Plus on a de paramètre plus on a besoin de données sinon on Over fit.

Yann Lecan, premier CNN « LENET » une combinaison entre la convolution et les RN

Une image => 16 filtres (convolution), 16 images de même taille => Relu 0 si x < 0 sinon x => Pooling, intervient sur la taille de l’image, on la réduit en gardant un max d’information.

##### Auto-encodeur :

Reproduire en sortie ce qu’il a eu en entrée en passant par un espace plus restreint (comme une PC (qui n’est pas linéaire) on peut ajouter plusieurs layer) cela le force à détecter les combinaisons de dimension les plus intéressantes et capturer le maximum d’information.

Il est composé en deux parties :

1. Encodeur : Image en entrée puis la plongé dans 🡺 🡺 🡺 🡺

🡺 Embedding : (bottleneck) dimension en espace latent réduit la dimension et capture les plus importantes caractéristiques.

1. Décodeur : Qui prend la dimension plus restreinte puis reproduit la sortie. Non supervisé.

Exemple :

Entrée : image 🡺 A l’intérieur : bruité l’image 🡺 Sortie : reconstitué l’image.

**Résumé**

# Apprentissage non supervisé

|  |  |
| --- | --- |
| Transformation du jeu de données | Partitionnement des données (Clustering) |
| 1. Réduction de dimension.    * Nouvelle représentation plus compréhensible.    * Part d’une représentation des données avec de nombreuses dimensions, pour trouver une nouvelle manière de représenter ces données afin d’en déduire ce qui est essentiel.   Souvent utilisé à des fins de visualisation en 2D.   1. Trouver les parties ou composants qui structurent les données, exemples :    * Extraire des thèmes dans des collections de documents textuels. | 1. Partitionnent les données en groupes distincts d’éléments similaires, exemples :  * Réseau pouvant regrouper les photos d’un utilisateur. |

# Pré-traitement et recalibrage

|  |  |
| --- | --- |
| From sklearn.processing import MinMaxScaler  Scaler = MinMaxScalet()  Scaler.fit(X\_train) // que sur le X\_train  X\_train\_scaled = Scaler.transform(X\_train)  X\_test\_scaled = Scaler.transform(X\_test) | From sklearn.processing import MinMaxScaler  Scaler = MinMaxScalet()  Scaler.fit\_tranform(X\_train)) |

# Blanchiment des images => utile

I make the distinction of an object recognition convolutional neural network because this process of PCA whitening can transform your data in such a way that makes an image unsuitable for certain tasks. Indeed, PCA whitening can be used for many more types of applications via a CNN. The idea of Principle Component Analysis (PCA) whitening is similar, but much less visually interesting. You are using PCA to reduce the dimensionality of your data to only the most important information. This makes it much easier for an **object recognition** convolutional neural network (CNN) to get meaningful results from your data, because you have often eliminated most of the **noise** in your image.

ZCA stretches the dataset to make it spherical, but *tries not to rotate it* (whereas PCA does rotate it quite a lot).

PCA – moins bon.

ZCA – garder l’orientation des données, composants ne sont pas orthogonales

ZCA étire l'ensemble de données pour le rendre sphérique, mais essaie de ne pas le faire tourner (alors que PCA le fait tourner beaucoup). ZCA pourrait être préférable, c'est le prétraitement pour les réseaux de neurones convolutifs.

MNIST (les pixels du bord sont blancs on peut les éliminer du jeu de d’entrainement sans perdre de l’information. Ces pixels ont une importance quasi nulle dans une tache de classification, deux pixels voisins sont souvent fortement corrélés)

# Chapitre 4 : Réduction de la dimension

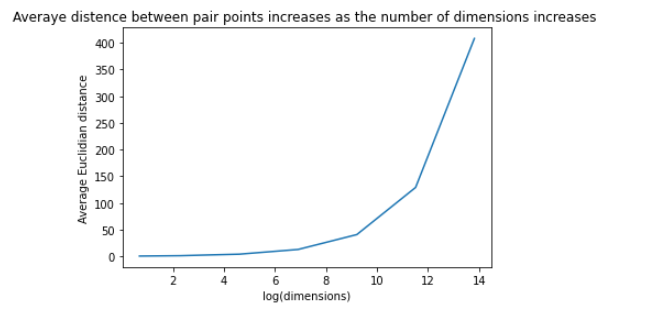
## **Fléau de la dimensionnalité**

Soit une matrice de données et un vecteur d’observations . Les observations portent sur variables, mesurées sur individus.

On parle de grande dimension dès que .

Conséquences :

* Les hypothèses dans des espaces à faibles dimensions ne sont plus valides dans des espaces à grandes dimensions
* Les distances entre toutes les paires de points deviennent identiques
* Répartition de la masse contre-intuitive, points aléatoires équidistants
* Les groupes existant dans les espaces à petites dimensions, disparaissent dans les espaces à grandes dimensions.
* Dans un carré de de taille 1X1, la distance moyenne entre deux points est de 0.52. Lorsque , les données deviennent très éparses, les observations (points) s’éloignement considérablement les uns des autres. Ce qui implique que :
  + La matrice devient très parcimonieuse et devient non inversible.
  + La densité locale des données dans le voisinage est trop faible pour correspondre à des distributions.
* 🡺 **ce qui rend les prédictions moins fiables, car elles seront basées sur des extrapolations plus larges** 🡺 **sur-apprentissage.**
* Phénomène de sur-apprentissage



* Les points s'éloignent du centre dans les hautes dimensions

|  |  |
| --- | --- |
|  | False intuition : As The points are uniformly distributed throughout the 2-ball and the radius is equal to 1. So perhaps, the average distance between the points and the origin is also uniformly distributed [0, 1]. Hence, it would be 0.5 ()  Area = π \* radius²  For the 2-ball with a radius of one :  Area = π \* 1² = π 3.14 |
|  | Let’s divide our 2-ball into 2 different non-overlapping areas :   1. Points, less than 0.5 units from the center :   Area = π \* ()² = π \* () =   1. Points, greater than 0.5 units from the center :   Area = Area – Area = π - =  **The average distance between each point and the center is actually about 0.67 and not 0.50** |
|  | How can we measure distances in spaces with more than three dimensions ?  Pythagorean theorem : a² + b² = c²   * a : distance between the center and the point’s coordinate on the x-axis * b : distance between the center and the point’s coordinate on the y-axis * c : the distance between the center and the point itself   Another name for c here is the Euclidean distance between the point and the center of the ball.  The average distance from randomly selected points to the origin increases rapidly as we increase the dimension . |

Motivation :

1. La compression des données.
2. La visualisation.
3. La recherche d’une présentation qui soit plus informative pour des traitements ultérieurs.
4. Etape préparatoire à une tâche d’apprentissage supervisé.

## Analyse en composante principale (PCA) :

Il identifie l’hyperplan le plus proche des données, et le projette sur celui-ci.

Il **choisit l’axe qui conserve la plus grande part possible** **de variance**, car il est très vraisemblable qu’il perdra moins d’informations que les autres projections. Pour justifier ce choix, on peut également dire qu’il s’agit de l’axe qui minimise la moyenne du carré des distances entre les données d’origine et leurs projections.

Axe1 : a le plus de variance.

Axe2 : axe orthogonal au premier, qui contribue le plus à la variance résiduelle.

**Le sens des flèches des composantes principales n’est pas stable,** **une petite perturbation du jeu de donnée peut les inversées.**

Attention **: PCA s’occupe de centrer les données pour nous**, si on utilise autre que celle de Sklearn, ne pas oublier de centrer les données avant d’appliquer une PCA.

**Application 1 : visualisation des données, Dim = 2 || 3.**

La méthode transforme des caractéristiques dites corrélées en caractéristiques statistiquement non corrélées. Puis sélectionne un sous-ensemble des nouvelles caractéristiques, selon l’importance qu’elles présentent pour la compréhension des données. L’algorithme commence par trouver la direction du maximum de la variance (component 1). Il s’agit de la direction (ou du vecteur) dans les données qui contient la plus grande part d’informations ou, en d’autres termes e la direction dans laquelle les caractéristiques sont le plus corrélées les unes avec les autres. L’algorithme trouve ensuite la direction qui contient le plus d’informations tout en restant orthogonale à la première. L’orientation n’a aucune importance. Les directions déterminées par ce processus sont appelées **composantes principales**, **car elles déterminent les principales directions de la variance dans les données**. Il existe autant de composantes principales que de caractéristiques originales. **La PCA se charge de la rotation et du décalage des données**, mais en conservant toutes les composantes principales.

Exemple :

Scaler = StandardScaler ()

Scaler.fit(X\_train)

X\_train\_Scaled = Scaler.tranform(X\_train)

From sklearn.decomposition import PCA

Pca = PCA(n\_components= 2)

Pca.fit(X\_train\_Scaled)

Pca.transform(X\_train\_Scaled)

La PCA est une méthode non supervisée, elle n’utilise aucune information provenant des classes pour trouver la rotation. Elle cherche simplement la corrélation dans les données.

*En page 149 axe-x = component principale 1, axes-y = component principale 2. On constate depuis la figure que les deux classes sont assez bien séparées, cela nous conduit à croire que même un classifier linéaire pourrait faire un travail raisonnable.*

**Application 2 : extraction de caractéristiques.**

Utile pour le classement d’image, la sécurité …

Exemple : on a une base avec des photos de personne.

Idée 1 : On construit un classifier par personne (classes distinctes). Cependant, il y’a peu d’image par classe (de la même personne) donc très peu d’échantillon d’apprentissage par classe, l’ajout d’une nouvelle personne (classe) nous oblige a ré-entrainer tout le modèle (qui devient de plus en plus volumineux).

Idée 2 : on construit un classifier KNN, mais on obtient un score 0.23

La PCA rentre en scène, car calculer la distance entre les pixels n’est pas une bonne idée pour mesurer la similitude entre les visages.

Lorsqu’on fait appel à une représentation sous forme de pixel pour comparer des visages, nous comparons en fait la valeur de niveau de gris pour chaque pixel à la valeur correspondante pour le pixel qui occupe la même position dans l’autre image. Si un pixel se voit décaler se traduit par une représentation différente. Ce qu’on espère, c’est qu’utiliser des distances en même temps que les composantes principales puissent améliorer notre exactitude.

La PCA fait une transformation dans laquelle les données subissent une rotation, les composants ayant une faible variance étant ensuite abandonnés. Il est aussi utile de voir la reconstruction des données en se servant de certaines composantes. Si component principale = nombre de pixel de l’image => on ne perd aucune information à la suite de la rotation.

**Réduire la dimension perd toujours un peu d’information et par conséquent, meme si cela accélère l’entrainement, cela peut aussi dégrader légèrement les performances de votre système et rend les pipelines plus difficiles à maintenir.**

* + Conseil **: On doit d’abord entrainer le système avec les données d’origine, puis envisager une réduction de dimension des données d’entrainement peut filtrer certains aléas et détails inutiles, occasionnant ainsi une amélioration des performances ; mais en général, cela permet uniquement de réduire la durée d’entrainement.**

De visualiser les données **; représenter les données en 2D ou 3D permet parfois d’en tirer des renseignements importants en détectant visuellement une structuration telle que des partitions.**

En pratique, on **peut combattre le fléau de la dimension en accroissant la taille des données de façon à atteindre une densité suffisante des observations**. Il faudrait plus d’observation d’entrainement que d’atomes dans l’univers observables pour que ces observations soient à une distance moyenne de 0.1 l’une de l’autre, en supposant qu’elles soient réparties uniformément sur toutes les dimensions.

**Avantage et inconvénients d’une PCA Ordinaire :**

* Entrainement plus rapide.
* Elimine le bruit et les variables redondantes.
* Visualiser les données pour avoi une idée des variables les plus importantes.
* Compression (économiser l’espace).
* Perte de l’information => dégrade les performances.
* Ne garantit pas qu’on aura une meilleure ou plus simple solution « tout dépend du jeu de donnée ».
* Ajoute de la complexité au pipeline.
* Variable transformer sont souvent difficile à interpréter.

Apres utilisation d’une réduction de dimensionnalité, il est impossible de faire marche arrière, car une partie de l’information est perdue lors de la réduction. Certain algorithme comme **PCA peuvent possèdent une transformation inverse qui peut reconstruire un jeu assez sembles**, d’autres ne peuvent pas comme t-SNE.

Evaluer une Réduction de Dimensionnalité :

* Appliquer une transformation inverse, et mesurer l’erreur de de reconstruction.
* Vérifier le score d’un algorithme d’apprentissage automatique pour les données d’origine et sur les données transformées. Si, peut écart (à la suite de la perte d’informations) alors bon, mauvais sinon.

SVD méthode de composition en valeur singulière qui permet de décomposer la matrice X en un produit de 3 matrices : U sigma V.T, V.T contient toutes les composantes principales.

U, sigma, v.t = Np.linalg.svd(X) % X = X – X.means(axis=0)

v.t[ :,0], v.t[ :,1]

PCA s’occupe que les données soient centrées autour de l’origine

pca.explain\_variance\_ratio

Choisir le nombre de dimensions représentant une continuation suffisamment importante à la variance (au moins 95%).

**Application 3 : Compression**

On peut réduire la taille du jeu de donnée en prenant l’essentiel de la variance. Ce qui permet d’accélérer considérablement un algorithme de classification (tel que SVM).

Il est aussi possible de décompresser le jeu de données réduit pour le ramener à sa taille d’origine en appliquant la transformation inverse. Il faut savoir qu’on ne trouvera pas exactement les données d’origine puisque la projection à perdu un peu d’information, mais les nouvelles données seront vraisemblablement très proches.

Erreur de reconstruction = distance quadratique moyenne entre les données d’origine et les données reconstruites.

« inverse\_transform()

### PCA incrémentale :

Si on utilise la PCA ordinaire le jeu de données doit tenir en mémoire pour pouvoir effectuer la recherche des valeurs singulières.

🡺PCAI effectue une analyse en composante principale en ligne (à la volé, au fure et à mesure de l’arrivée de nouvelle observation).

### PCA randomisée : PCA (n\_components=n, svd\_solver= ‘randomized’)

Algorithme stochastique trouvant rapidement une approximation des d premières composantes principales. Elle donc considérablement plus rapide que les algorithmes précédents.

### Projection à noyau :

Une technique mathématique qui projette implicitement les observations dans un espace de tès grande dimension appelé espace de rescription

Projection plus complexe, elle donne de bon résultat lorsqu’il s’agit de préserver des regroupements d’observation après projection. Ou parfois meme pour dérouler des jeux de données se situant à proximité d’une variété enroulée.

Est un algorithme non supervisé, il n’y a pas de mesure de performance pour nous aider à sélection le meilleur noyau et les valeurs des paramètres. Cependant, la réduction de dimensionnalité est souvent une étape préparatoire à une tache d’apprentissage supervisé.

Pipe = Pipeline ([

(‘pca’, KernelPCA(n\_component =2)), (‘log\_reg’ , LogisticRegression())

])

Params = [{‘pca\_\_gamma’ : np.linspace(0.03, 0.05,10) ,

‘pca\_\_kernel’ : [’rbf’, s=’sigmoid’}]

Grid = GridSearchCV(pipe, param\_grid = params, cv =3)

Grid.fit(X\_train,y\_train)

## Factorisation en matrice non négative (NMF) :

**Pré condition : données >= 0**

Est un algorithme **non supervisé visant à extraire des caractéristiques utiles**. Il fonctionne d’une manière semblable à la PCA, et peut aussi servir à la réduction de dimension.

On essaye d’écrire chaque point de donnée sous **la forme d’une somme non négative pondérée de certaines composantes** (les valeurs >= 0) contrairement la PCA qui elle cherche des composantes orthogonales et une variance maximale. La NMF n’est pas appliquée aux données pour lesquelles les caractéristiques sont négatives.

La NMF est utile pour des données qui ont été créés sous forme d’addition ou de recouvrement de plusieurs sources indépendantes (exemple : enregistrement audio sur lequel plusieurs personnes parlent, ou orchestre avec plusieurs instrument). Dans de telles situations NMF est capable d’identifier les composantes originales constituant les données combinées. Elle produit des composantes et des coefficients mieux interprétables que la PCA.

LDA

Réduire le nombre de composant ne supprime pas certaines directions, mais crée en fait un ensemble totalement différent de composantes. Dans NMF les composantes jouent toutes à égalité, car il n’existe pas d’ordre particulier. NMF utilise une initialisation aléatoire, ce qui pourrait conduire à des résultats différents.

Le nombre de composantes qu’on souhaite extraire << le nombre de caractéristiques en entrée (sinon on considère que chaque pixel est une composante distincte.

Faible pour la reconstruction ou l’encodage des données.

Fort pour trouver des motifs intéressants dans ces données.

## LLE

Une technique non linéaire très efficace de réduction de dimension. C’est une technique d’apprentissage de variété ne reposant pas sur des projections comme les algorithmes précédents. En bref, LLE constitue à mesurer tout d’abord les relations linéaires entre chacune des observations du jeu d’entrainement et ses plus proches voisines, puis à rechercher une représentation du jeu d’entrainement dans un espace de faible dimension qui préserve au mieux les relations locales. Ceci rend cette technique particulièrement efficace pour aplanir des variétés enroulées, en particulier lorsqu’il n’y pas trop de bruit.

-les distances ne sont pas préservées à plus grande échelle (une partie devient rétrécie/élargie alors qu’elle ne l’été pas)

+effectue en général une bonne modélisation de la variété.

Fonctionnement :

Problème d’optimisation sous contrainte

* + - 1. Identifie les k plus proches voisines de chacune des observations d’entrainement
      2. Tente de reconstruire comme fonction linéaire de ses voisins. Il recherche les poids tels que le carré de la distance entre et soit la plus petite que possible
      3. Projeter les observations d’entrainement dans un espace à d dimension tout en préservant autant que possible ces relations locales.

Complexité algorithmique :

O(m log(m) n log(k)),pour la recherche des plus proches voisins

O(m n k^^3) pour l’optimisation des poids

O(d m^^2) pour la construction des représentations en faible dimension.

LLE n’est pas adapté aux données de très grande taille.

# Apprentissage de variétés avec t-SNE

Ce sont des algorithmes destinés **à la visualisation et non à la génération de nouvelles caractéristiques.**

t-SNE calcule une nouvelle représentation des données d’apprentissage, mais ils ne permettent pas d’effectuer des transformations de nouvelles données. Cela implique que ces algorithmes ne peuvent pas être appliqués aux données de test.

+Utile pour l’analyse exploratoire.

+/-Rarement utilisé lorsque le but final est un apprentissage supervisé.

Le but : **Trouver une représentation bidimensionnelle des données qui préserve le mieux possibles les distances entre les points**. Il débute par une représentation aléatoire en 2D pour chaque point de donnée, puis il essaie de rapprocher les points qui sont éloignés l’un de l’autre dans l’espace des caractéristiques d’origine, et les points qui sont éloignés. t-SNE met l’accent sur les points qui sont proches, plutôt que de préserver les distances entre des points éloignés.

**Effectue une réduction de dimension tout en essayant de conserver les observations semblables proches et les observations dissemblances éloignées ? cette méthode est utilisée surtout pour la visualisation, et en particulier pour visualiser des groupes d’observations dans un espace de grande dimension.**

MDS : réduit la dimension tout en s’efforçant de préserver les distances entre observations

Isomap : créer un graphe en connectant chaque observation à ses plus proches voisins, puis effectue une réduction de dimension en s’efforçant de préserver les distances géodésiques entre observations.

LDA : algorithme de classification détermine durant **l’entrainement les axes les plus discriminants** entre les classes, axes qui peuvent alors etre utilisés pour définir un hyperplan sur lequel projeter les données. L’intérêt est que la projection conservera les classes aussi éloignées que possibles, c’est pourquoi la LDA est une bonne technique de réduction de dimension avant d’exécuter un algorithme de classification tel qu’un classifieur SVM.

# Clustering

Le clustering consiste à partitionner le jeu de donnée en groupe (cluster) très différents.

### Partitionnement en k-moyennes :

Un algorithme de clustering très simple et utilisés. Il essaie de trouver des centres de clusters qui sont représentatifs de certaines régions des données.

Etape 1 : affecte chaque point de donnée au centre du cluster le plus proche.

Etape 2 : redéfinit le centre du cluster comme étant la moyenne des points de données qui lui ont été affectés.

L’algorithme se termine quand l’affectation des instances aux clusters ne varie plus.

* Echoue à identifier des clusters ayant une forme complexe.

|  |  |  |
| --- | --- | --- |
| PCA | NMF | K-moyenne |
| Trouver les directions ayant la variance maximale dans les données. | Trouver les composantes additives, qui correspondent souvent à des extrémités ou à un axe dans les données | Représente chaque point de donnée en utilisant un centre de cluster. |
| Algorithme de décomposition  Exprimer les données sous a forme d’une somme pondérée d’un certain nombre de composantes | | Algorithme de partitionnement  (Chaque point était une seule composante)  Quantification vectorielle |

Quantification vectorielle :

+ On peut utiliser plus de cluster que de dimensions de l’entrée pour encoder les données.

Si cluster >> Représentation plus expressive.

Si modèle complexe, utiliser plus de cluster pour couvrir les variations => On aura de nouvelles caractéristiques donc plus de dimension ce qui peut nous amener à utiliser un modèle linéaire.

|  |  |
| --- | --- |
| Force | Faiblesse |
| * Facile à comprendre. * Facile à implémenter. * Rapide. * S’adapte à de gros jeux de données. | * Initialisation aléatoire (exécuter l’algorithme 10 fois et choisir le meilleur dans le sens ou la somme des variances des clusters est minimal). * Relativement restreint quant aux suppositions de la forme des clusters. * Spécifier le nombre de clusters (nombre qu’on ne connait pas dans le monde réel). |

# Clustering agglomératif

L’algorithme commence par déclarer chaque point étant son propre cluster, puis il fusionne les deux clusters les plus similaires ou les plus proches jusqu’à satisfaire un critère d’arrêt (nombre d’itération/clusters). Il y’a plusieurs critères de lien (linkage) qui spécifient la façon exacte de mesurer ce que représente les *clusters les plus similaires*, on cite sous scikit-learn :

|  |  |  |
| --- | --- | --- |
| ward | Average | Complete - Maximim |
| * Le choix par défaut. * Il sélectionne les deux clusters à fusionner de manière à ce que la variance augmente le moins à l’intérieur de tous les clusters –l’inertie intra classe. Cela conduit à des clusters qui sont de taille relativement égale. | * Fusionne les deux clusters qui ont la plus petite distance moyenne entre tous les points. | * Fusionne les deux clusters qui ont la plus petite distance maximale entre leurs points. |
| * Fonctionne bien avec la plupart des jeux de données. | * Si les clusters ont des nombres très dissemblables de membres, les liens Average ou complète pourraient être meilleurs | |

**Clustering hiérarchique ascendant**

Procède de manière itérative

Init : chaque point est un cluster lui-même …

* Toutes ces étapes produisent un nouveau partitionnement des données (donne toutes les combinaisons de clusters possibles.

+ Donne une vue très détaillées.

-Mais s’appuie sur des données bidimensionnelles.

Correction de la lacune 2D

## DBSCAN

Pré condition : Données recalibrées StandardScaler ou MinmaxScaler.

+ L’utilisation ne définit pas a priori le nombre de clusters.

+ Capture des clusters de forme complexe.

+ Identifie des points qui n’appartiennent à aucun cluster.

+ S’adapte aux gros jeux de données.

-Lent comparé aux autres méthodes.

-Pas adapté aux petits jeux de données (changer le paramétrage ).

Fonctionnement :

Idée :

* Les clusters forment les régions les plus denses séparés par des régions relativement vides.
* Identifie les points qui se trouvent dans les régions encombrées - denses dans l’espace de caractéristiques (la ou les points sont proches les uns les autres). Ces points sont dits core points (points intérieurs).
* L’algorithme possède deux paramètres :

|  |  |
| --- | --- |
| Min\_samples  **Détermine** le statut des points (aberrants/outliers/anormaux/atypiques) qui se trouvent dans les régions moins denses => **taille du cluster**.  Si min\_samples >> on diminue le nombre de core point et on augmente la quantité de bruit. | Esp (epsilon)  Important car il **détermine le sens** de **proche**.  Si Esp >> on a plus d’exemple dans un cluster.  Si Esp >>> fusionne les clusters (fini avec 1).  Si Eps << il n’y a que du bruit. |

* S’il y’a au moins min\_samples points se trouvat à une distance < esp d’un certain point de donnée, alors celui-ci est classifié comme étant un core point. Les points intérieurs qui sont plus proches les uns des autres que la distance eps sont placés dans le même cluster.

1. L’algorithme choisit un point aléatoirement pour commencer.
2. Identifie les points qui sont < eps, si le nombre de point trouvés < n\_samples alors le point de départ est caractérisé comme un bruit (il n’appartient à aucun cluster) sinon il est un core point.
3. Tous les points proches < esp du cluster sont visités.
4. On boucle.

A la fin on aura 3 types de points :

|  |  |  |
| --- | --- | --- |
| Core point | Boundary point/points frontières (Au voisinage de core point) | Nose/bruit (-1) |

-Si on exécute l’algorithme plusieurs fois les nose et core points restent inchangés. Par contre, l’appartenance à tel ou tel cluster dépendent de l’ordre dans lequel ils ont été visités (mais pas trop important).

Représentation des données et ingénierie des caractéristiques :

|  |  |
| --- | --- |
| Caractéristique catégorielle/discrète | Caractéristique continue/numérique |

Dans le monde réel, les données ne sont pas collectées sous forme d’un tableau. On les retrouve sous deux types de données :

La manière de **représenter les données** peut avoir une **influence plus grande** sur les performances des **modèles d’apprentissage automatique** qu’un *choix approprié* des **paramètres.**

Représenter le mieux possible les données **l’ingénierie des caractéristique / feature engineering**

## Du catégorielle au numérique :

Il existe plusieurs manières de représenter des variables catégorielles en numériques, on citera :

Encodage one-hot (variables indicatrices) :

La manière de loin la plus courante consiste à utiliser l’encodage one out of N ou encore dummy variables.

L’idée : **Remplacer une variable catégorielle par une ou plusieurs nouvelles caractéristiques (1 si vrai 0 sinon).**

*Exemple :*

*Print (data.gender.value\_counts()) % mieux connaitre la valeur des caractéristiques.*

*Print («Original features», list(data.columns)*

*Data\_dummies = pd.get\_dummies(data)*

*Print («Features after get\_dummies», list (data\_dummies.columns)*

*Rmq : les variables catégorielles ont été encodées sous forme de chaine de caractères donc possibilité de rencontrer des fautes d’orthographe ou de saisie. Mais cette présentation montre clairement que ce sont des caractéristiques catégorielles.*

*Attention :*

* *Vérifier le nombre et le type des colonnes après dummy pour les données d’apprentissage et les données de test.*

*Si la variable catégorielle n’est pas encodée en numérique (1-8), il n’est pas évident de se rendre compte que c’est une variable catégorielle et la fonction get\_dummies de panda ne traite ce genre de représentation => Méthode OneHotEncoder.*

*On encode les variables catégorielles en utilisant des entiers.*

*Il ne faut pas les traiter comme étant des variables continues.*

Binning, discrétisation, modèle linaires et ARBRES :

**La meilleure façon de les représenter dépend :**

**-De leur sémantique. -Du modèle qu’on utilise.**

|  |  |
| --- | --- |
| **Modèle** | |
| Modèles arborescents (AD, Boosting ou FA) | Modèle linéaire |

**Propriétés différentes !!**

|  |  |
| --- | --- |
| * La discrétisation n’apporte pas de bénéfice aux modèles arborescents, car ils n’en ont pas besoin pour apprendre à diviser les données (capables de traiter de multiples caractéristiques en même temps). | Ils deviennent plus puissants/flexibles en utilisant :   * **La discrétisation (binning)**, car il maintient une valeur constante pour chaque bin |

Intersection et polynomes:

Une autre manière consiste à ajouter des intersections et des polynômes aux caractéristiques d’origine. Ce type d’ingénierie est souvent employé en modélisation statistique mais aussi en ML.

|  |  |
| --- | --- |
| * Les FA battent la performance de Ridge, et le fait d’ajouter des intersections et des fonctions polynomiales fait même légèrement baisser les performances. | * Etant donnée que les ML être visualisés sous forme de droites possédant une centaine pente. On peut faire une **discrétisation** + **réintroduire la caractéristique d’origine**. * **Discrétisation** + une nouvelle caractéristique (**produit de l’indicateur bin** **et de la caractéristique d’origine**). |

Exemple :

From sklearn.preprocessing import PolynomialFeatures

Poly = PolynomialFeatures(degree = 10), include\_biais =False)

Poly.fit(X)

X\_poly = poly.transform(X) % transformer les données

Print (poly.get\_feature\_names()) % x0, x0\*\*2, x0\*\*3 ...

Linear\_poly = LinearRegression ().fit(X\_poly, y)

Cependant les polynômes élevés ont tendance à se comporter de manière extrême aux limites ou dans des régions ne comportant que peu de donnée.

Les SVM à noyau donnent une prédiction aussi bonne dans le cas d’une régression polynomiale, mais sans transformation explicite des caractéristiques.

Transformation non linéaires uni-variées:

L’ajout des caractéristiques carrées ou cubiques peut aider à **améliorer les performances des modèles linéaires pour la régression**. Il existe aussi des applications mathématiques telles que log, exp, cos et sin.

Log et exp : Capture les échelles relatives des données ou pour dénombrer des entiers (jamais négatifs). Exemple : Combien de fois un utilisateur à … np.bincount(X) de zéro.

Sin et cos : Intéressants pour des données qui encodent des signaux périodiques.

|  |  |
| --- | --- |
| Modèles arborescentes | Modèle linéaire et RN |
| Se soucient que **de l’ordre des caractéristiques**. | Très liés à l’**échelle** et à la **distribution** de chaque caractéristique, et par se fait de savoir s’il **existe une relation non linéaire entre la caractéristique et la cible**. |

La plupart des modèles **fonctionnent mieux lorsque chaque caractéristique** (aussi la cible dans le cas de la régression) **est distribuée de manière à peu près gaussienne**. En d’autres termes, un **histogramme de chaque caractéristique** devrait, dans le meilleur des mondes, donner une représentation ayant à peu près la forme d’une **courbe en cloche**.

Histogramme avec beaucoup de **petites valeurs et peu de grandes valeurs** => **distribution en poisson**, assez fondamentale dans le comptage de données. Appliquer une transformation logarithmique (approximation grossière de l’utilisation de la régression de Poisson) pour avoir une courbe en cloche (Si 0 Є X alors **LOG (X+1)**).

*Trouver la transformation qui fonctionne le mieux pour chaque combinaison de jeu de données et de modèle est tout un art.*

*Il peut arriver que chaque caractéristique ait besoin d’être transformer de manière différente*.

*Bonne idée de transformer la variable cible y dans une régression*.

Ces types de **transformations** ne sont pas **adaptés aux modèles arborescents**, ils sont souvent **capables de découvrir par eux-mêmes des interactions importantes**.

Sélection automatique des caractéristiques :

Il existe plusieurs façons de créer de nouvelles caractéristiques, nous sommes tentées *d’augmenter la dimensionnalité* des données => *modèle complexe => augmenter le risque de sur-apprentissage.*

*Dans le cas de grand nombre de dimensions, on réduit le nombre de caractéristiques pour ne conserver que les plus utiles et laisser les autres de côté => modèle simple qui généralise mieux.*

A quel point chaque caractéristique est bonne ?

Trois stratégies se présentent : les *statistiques uni-variées*, *sélection basée sur le modèle* et *la sélection* *itérative,* tous sont des méthodes supervisées (besoin de la cible pour l’ajustement du modèle).

Statistique uni-variées :

Cette méthode calcule s’il existe une relation **statistiquement signifiante** entre chaque **caractéristique et la cible**. Ensuite, les **caractéristiques** qui **sont liées avec le plus haut degré de confiance** sont **sélectionnées**. Dans le cas de la classification, on appelle cela l’analyse de variance (ANOVA-ANalyst Of VAriance). Uni-variée dans le sens qu’elle **considère uniquement chaque caractéristiq**ue de manière individuelle. Par conséquence, une **caractéristique est rejetée si** elle est seulement **informative lorsqu’elle est combinée avec une autre.**

|  |  |
| --- | --- |
| f\_classif (classification) | F\_regressor (régression) |

Via scikit lean => Utiliser la méthode

Caractéristique rejetée << p-values => peu susceptible d’être liée à la cible.

Calculer le seuil from sklearn.feature\_selection import :

* SelectKBest (sélectionne un nombre fixé k de caractéristiques).
* SlectPercentile (sélectionne un pourcentage fixé k de caractéristiques).

+Calcule est rapide.

+Ne nécessite pas la construction d’un modèle.

+Indépendant du modèle.

+/- **Supprimer le bruit améliore les performances, même si certaines des caractéristiques d’origine sont perdues**.

+Utile s’il y’a trop de caractéristiques pour construire un modèle.

+Utile si on suspecte que de nombreuse caractéristique **comporte des informations sont inutiles**.

sélection de caractéristiques basées sur le modèle :

Cette méthode utilise un modèle d’apprentissage automatique supervisé pour juger de l’importance de chaque caractéristique afin de ne conserver que les plus importantes.

Le modèle supervisé servant à cette sélection n’a pas besoin d’être le même que celui qui va servir à la modélisation finale (il doit fournir une certaine mesure de l’importance de chaque caractéristique afin de pouvoir effectuer un classement en fonction de cette mesure).

Les modèles arborescents ont un attribut feature\_importances\_ qui encode l’importance de chaque caractéristique.

Les modèles linéaires disposent de coefficients qui peuvent aussi capturer l’importance des caractéristiques en considérant les valeurs absolues.

* Une forme de sélection pour le modèle lui-même.
* Un prétraitement afin de choisir des caractéristiques pour un autre modèle.

La méthode considère toutes les caractéristiques de manière globale, contrairement à la sélection univariée, peut donc capturer les interactions (si le modèle en est capable).

From sklearn.feature\_selection import SelectFromModel

Si on ne sait pas quelle méthode utilisée, mieux vaut utiliser la sélection automatique de caractéristiques.

+Adaptée pour la réduction de dimensionnalité.

+Accélère les prédictions.

+Rend les modèles plus faciles à interpréter.

Selection iterative de caractéristiques :

La méthode consiste à construire une **série de** **modèles** avec un nombre variable de caractéristiques. Il existe deux façons : Commencer sans caractéristique puis les ajouter une par une jusqu’a ce qu’un certain critère d’arrêt soit atteint ou inversement.

-Couteuse en ressource.

L’algorithme RFE (Recursive Feature Elimination) débute avec toutes les caractéristiques, construit un modèle, et rejette la caractéristique la moins importante pour ce modèle. Un nouveau modèle est ensuite construit à partir des caractéristiques restantes et ainsi de suite …

Attention : Le modèle utilisé doit pouvoir déterminer l’importance des caractéristiques.

Une fois les caractéristiques sélectionnées les MA et ML obtiennent les mêmes scores de prédiction.

Dans le monde réel, ces méthodes apportent peu de gains de performances mais reste un bon outil pour l’ingénierie des caractéristiques.

Savoir utiliser l’expertise :

Une expertise métier est importante dans l’ingénierie des caractéristiques. Bien que le but de l’apprentissage automatique soit d’éviter d’avoir à créer un ensemble de règles conçues par des experts.

Exemple page 246 :

La prédiction effectuée par une foret aléatoire en utilisant uniquement le temps POSIX est une ligne droite (depuis la dernière heure pour laquelle les données ont été observées), car **la valeur temps POSIX pour le jeu de test se trouve en dehors de la plage des valeurs disponibles dans le jeu d’apprentissage.**

Une régression linéaire avec un encodage du jour, de la semaine et de m’heure donne de très mauvais résultats, du fait que l’encodage de ces derniers en entiers, qui sont interprétés comme des variables continues. On peut essayer de capturer ces informations en **interprétant les entiers sous forme de variables catégorielles** (***transformation OneHotEncoder***). => améliore le score !

On peut faire appel à des intersections **pour que le modèle apprenne un coefficient pour chaque combinaison jour/heure du jour** => cela améliore encore mieux ~ modèle FA

+ Modèle FA donne de bon résultat mais ML nous permet de tracer les coefficients.

Les modèles linéaires bénéficient de ka génération de nouvelles caractéristiques via la discrétisation ou l’ajout de polynômes et d’intersections. Mais ce n’est pas le cas pour les modèles plus complexes - - **SVM et FA**.

# Evaluation et amélioration du modèle

**Echantillonnage et suites de variables aléatoires**

Un échantillon de taille est une partie de n éléments choisis aléatoirement dans une population

On dit qu’il est **non exhaustif lorsque le tirage se fait avec remise (évènements indépendants)**, exhaustif dans le cas contraire.

Si la population est très grande ou infinie, on peut considérer que le tirage est non exhaustif : le fait de remettre ou non l’élément tiré devient sans importance.

On suppose maintenant qu’on étudie un caractère statistique quantitatif C sur P. On considère l’ensemble des évènements : ”tirage avec remise d’un échantillon de taille n”. On obtient une suite de variables aléatoires ,... , où est la valeur du caractère C sur le i -ème élément de l’échantillon tiré. Ces variables sont indépendantes et suivent toutes la même loi de probabilité, d’espérance la moyenne de C et d’écart-type celui de C. A partir des n variables aléatoires Xi, on en construit de nouvelles, appelées ”estimateurs”.

On choisit un échantillon particulier en (d’où l'appellation ”**ponctuelle**”).

## Estimateurs, bais et variance

L’estimation ponctuelle consiste à fournir la meilleure prévision d’une certaine quantité d’intérêt. Elle peut être un paramètre unique, vecteur (poids d’un modèle) ou une fonction entière.

Soit

Un bon estimateur ponctuel ou statistique possède une valeur proche du véritable paramètrequi a généré les données d’apprentissage.

Biais et variance

**Le bais mesure l’écart prévu par rapport à la valeur réelle de la fonction ou du paramètre.**

**La variance, fournit une mesure de l’écart par rapport à la valeur en espérance possiblement fournie par un quelconque rééchantillonnage des données.**

Faut-il choisir un modèle avec un grand biais ou plus de variance ?

Dans un cadre applicatif, le compromis est négociable grâce à la validation croisée. On peut également comparer l’erreur quadratique moyenne (EQM) :

L’équation intègre à la fois le biais et la variance. Les estimateurs souhaitables sont ceux qui ont une petite EQM.

La relation entre biais et variance est liée à la capacité du modèle, le sur-apprentissage et le sous-apprentissage.

Capacité croissante :

Sous-apprentissage 🡪 biais >> + variance <<

Sur-apprentissage 🡪 biais << + variance >>

Généralisation : capacité d’avoir de bonne performance sur des entrées non observées précédemment

Plus on a de données, plus nos estimateurs sont fiables.

* Estimation par maximum de vraisemblance
* Log-vraisemblance conditionnelle et erreur en moyenne quadratique

## Cross Validation

### Il existe plusieurs variantes de la VC dans Sklearn

1. Cross validation croisée non stratifiée
2. Cross validation croisée standard

- Très utiliser pour la régression

1. Cross validation croisée Leave-one-out

* Pour chaque division, un seul point de donnée est choisi pour construire le jeu de test.
* Processus tres coûteaux pour de grand jeu de données.
* Bonne pour les données de petite taille.

1. Cross validation croisée Shuffle Split

* On échantillonne de manière aléatoire *train\_size* de données pour le trainset et *test\_size* pour le testset, *nsplit* fois.
* Cette méthode permet de contrôler le nombre d'itération indépendamment de la taille des jeux d'apprentissage et de test.
* Permet d'utiliser une partie des données lors de chaque itération (ne couvre pas la totalité des échantillons).
* Utile pour des jeux de données de grandes tailles.

1. Cross validation croisée stratifiée

* Très utiliser pour la classification
* Equilibre des classes déséquilibrées

1. Cross validation croisée avec groupes

* Les groupes de données partagent les mêmes traits de similarités.
* Un groupe contient les données qui ne devraient pas être séparées lors de la création des jeux de d’apprentissage et test. Exemple :
  + Toutes les expressions faciales d’une même personne
  + Echantillons du même patient

## **Métriques d’évaluation :**

Une image contenant table

Description générée automatiquement



### **Classification binaire :**

Classe positive vs négatif

#### Types d’erreurs :

|  |  |
| --- | --- |
| ***Faux positif/erreur type 1*** | ***Faux négatif/erreur type 2*** |
| *Un patient en bonne santé, classifié positivement malade => analyses supplémentaires, coût inutile, désordre mentale…* | *Un patient malade, classifié non-malade => pas de traitements, mort...*  Dans le cas d’un diagnostic cancéreux, un faux négatif est plus grave qu’un faux positif. |

#### Jeux de données déséquilibrés

Avec un mauvais équilibre des classes (exemple : les clics des utilisateurs), on peut obtenir une prédiction de 99% car la prédiction ne concerne que la classe majoritaire.

Il faut une métrique capable d’éliminer ces prédictions absurdes pour bien évaluer notre modèle.

#### *Matrice de confusion*

Une image contenant table

Description générée automatiquement

Comment résumer la matrice de confusion ?

##### Exactitude :

EXACTITUDE =

* Le rapport entre le nombre de prédictions correctes et le nombre total d’échantillons traités.
* **Efficace pour la classification de données équilibrées**
* Facile à calculer et à comprendre
* Ne prend **pas** compte la **conséquence des erreurs (type d’erreur)**

##### Précision :

PRECISION =

Mesure **la proportion des échantillons prédits comme étant positifs et qui le sont effectivement**.

* Cette une métrique de **performance lorsque le but est de limiter le nombre de faux positifs**.

Exemple :

Une entreprise veut lancer un produit, les matières premières coûtent cher. Elle veut être sûre que le produit marchera pour débuter la production.

##### Rappel :

RAPPEL =

* Identifier tous les échantillons positifs, autrement dit, **éviter les faux négatifs**.

Example :

Diagnostique cancéreux, un faux négatif est plus grave qu’un faux positif.

**Il existe un compromis entre optimiser la Précision et optimiser le Rappel.**

##### F-Mesure :

* La moyenne harmonique de la précision et du rappel
* Meilleur que les deux séparés
* Classification de données déséquilibrées
* Difficile à interpréter et à expliquer que l’exactitude

#### Prendre en compte l’incertitude :

Si les données sont trop déséquilibrées, changer le seuil de décision « decision\_function » ou « predict\_proba » peut améliorer les résultats facilement.

Un seuil à 0.5 signifie que le modèle est sûr à plus de 50% qu’un point est positif.

Augmenter le seuil signifie que le modèle est à besoin d’améliorer son taux de certitude pour prendre une décision positive et donc qu’il a besoin de moins de certitude pour prendre une décision négative.

Ne pas confondre seuil arbitraire est probabilité.

Un arbre qui atteint toute sa profondeur est sûr à 100% de ses décisions, même s’il peut se tromper.

Un modèle recalibré est un modèle qui fournit une mesure précise de son incertitude.

Changer le seuil pour prendre une décision de classification est une manière d’ajuster le compromis entre précision et rappel.

##### Courbe précision-rappel & précision moyenne :

###### Courbe précision-rappel :

Chaque point de la courbe correspond à un seuil possible de function. Le seuil 0 correspond à la valeur par défaut de decision\_function. Ce point est appliqué lors de l’appel de la méthode . Poser une exigence sur un classifier, sur la valeur du rappel ou la précision c’est définir le **point de fonctionnement (operating point).** C’est utile pour apporter des garanties de performances à des clients.

Pour mieux comprendre un problème de modélisation, il est recommandé de considérer tous les seuils / compromis possibles entre précision et rappel avec **la courbe précision-rappel.**

La fonction attend comme second argument une mesure de certitude pour la classe positive.

Le meilleur seuil qui équilibre entre la précision et le rappel, revient à optimiser la F-mesure qui résume la moyenne harmonique des deux mesures : Mesure

* La courbe précision rappel reste néanmoins un processus manuel. Pour une comparaison automatique, il est possible de résumer les informations contenues dans la courbe et ce sans nous limiter à un seuil ou à un point de fonctionnements spécifique 🡪 Calculer l’aire sous la courbe précision-rappel 🡪 Précision moyenne.

###### Précision moyenne :

Résumer les informations contenues dans la courbe précision-rappel, revient à calculer l’intégrale ou l’aire sous la courbe. Cette technique s’appelle précision moyenne.

##### Courbe ROC (Receiver Operating Characteristic) et métrique de l’AUC (l’air sous la courbe) :

###### Courbe ROC (Receiver Operating Characteristic) :

Un outil également utilisé pour analyser le comportement des classifieurs en fonction de différents seuils. La courbe ROC considère tous les seuils possibles pour un classifieur donné, mais lieu de fournir un lien entre la précision et le rappel, elle montre le **taux de faux positifs par rapport au taux de vrais positifs (rappel / sensibilité)**.



###### Métrique de l’AUC (l’air sous la courbe) :

AUC résume l’information de la courbe de ROC. Elle calcule l’aire sous la courbe comprise entre 0 et 1 de ROC, son résultat est également compris entre 0 (le pire) et 1 (le meilleur – signifie que tous les points positifs ont un score plus élevé que tous les points négatifs). Une prédiction aléatoire renvoie une valeur d’AUC de 0.5, et ce quel que soit le déséquilibrage des classes dans le jeu de données. AUC est meilleure métrique que l’exactitude dans les problèmes de classification déséquilibré. Il peut être interprété comme une évaluation des échantillons positifs. Il est équivalent à la probabilité qu’un point de la classe positive sélectionné de manière aléatoire ait un score plus élevé pour le classifieur qu’un point sélectionné aléatoirement dans la classe négative. Utiliser AUC pour sélectionner un modèle est plus efficace que de se baser sur l’exactitude.

Une image contenant texte

Description générée automatiquement

La métrique AUC + GridSearchCV ou cross\_val\_score (argument scoring) peuvent aider à sélectionner un modèle.

## **Métrique pour la classification multi-classe :**

Les métriques pour la classification multi-classe sont dérivées des métriques pour la classification binaire mais moyennées sur toutes les classes. Elle est définie en tant que fraction des échantillons correctement classifiés. Et, là encore, lorsque les classes sont déséquilibres, l’exactitude n’est pas une bonne mesure d’évaluation.

On utilisera

* + - * La **matrice de confusion** puis **Classification\_report**
      * F-mesure

L’idée consiste à calculer une valeur f-mesure binaire par classe, celle-ci devenant la classe positive, les autres formant les classes négatives. Ces scores sont ensuite « moyennés » en utilisant les stratégies :

* Macro : effectue le calcul de la moyenne pour l’ensemble des classes.
  + Utile si on veut que toutes les classes soient considérées à l’identique.
* Pondéré : calcule la moyenne des f-mesures pour toutes les classes, pondérée par leur support (**Classification\_report)**.
* Micro : : effectue le calcul de la moyenne pour le total de faux positifs, de faux négatifs et de vrais positifs pour l’ensemble des classes, puis recalcule précision, rappel et f-mesure en utilisant ces résultats.
  + Utile si on veut que chaque échantillon soit traité à l’identique.

## **Métriques de régression :**

* Erreur quadratique moyenne.
* Erreur absolue moyenne.

# Chaînage d’algorithme et Pipeline

## **Avantage :**

* Outil généraliste servant à chaîner plusieurs étapes de traitement dans un flux d'apprentissage automatique et chaîner le tout dans un seul objet Python
* Outil très utile pour éviter la fuite d'information lors de la validation croisée et la sélection de paramètres par recherche de grille
* Pipeline permet d'écrire du code de manière plus succincte
* Réduit les erreurs

# Chapitre 6 Données textuelles

Il existe deux types de caractéristiques :

Corpus : jeu de donnée

Document : chaque point de donnée.

Donnée catégorielle

Texte

Réponse possible obtenues à partir du contenu d’un champ de texte. Pouvant être associée d’un point de vue sémantique à des données catégorielles.

Chaine libre

Attention aux fautes d’orthographe et aux doublant !!

Text\_train = [doc.replace(« … »,  « … ») for doc in text\_train]

Convertir le format en une représentation

Numérique afin d’appliquer un algorithme

1. Représenter les données textuelles sous forme de sacs de mots (bag of word)

Méthode simple et courante, servant à transformer les textes afin de leur donner un format reconnaissable pour l’apprentissage automatique. On **élimine l’essentiel de la structure du texte** présenté en entrée. On **compte le nombre d’apparitions de chaque mot dans chaque texte du corpus.**

*Etape 1 :* ***Tokenization***

Partager un document dans les mots qui le composent => un texte devient une séquence de *tokens*.

*Etape 2 :* ***Construction du vocabulaire***

**Construire un dictionnaire de tous les mots qui apparaissent dans les documents** et les numéroter (exemple dans l’ordre alphabétique).

*Etape 3 :* ***Encodage***

Dénombrer le nombre d’occurrences de chaque mot du vocabulaire.

Exemple :

« This is how you get ants »

Words = [‘The fool doth think he is wise’, ‘but the wise man knows himself to be a fool’]

From sklearn.feature\_extraction.text import CountVectorizer

Vect = CountVectorizer()

Tokenization et construction du vocabulaire

Vect.fit(words)

Vect.vocabulary\_ % {‘the’: 9, ‘fool’: 3... ‘be’: 0}

* On utilise Toarray, car words est de petite taille.
* Le vocabulaire est par ordre alphabétique.

Words.toarray() % [[0 0 1 1 1 0 1 0 0 1 1 0 1]

[1 1 0 1 0 1 0 1 1 1 0 1 1]]

1. Éliminer les mots sans importances : Pour améliorer les performances du modèle
2. Augmenter le nombre d’apparition :

Vect = CountVectorizer(min\_df /max\_df= 5).fit(text\_train)

X\_train = vect.transform(text\_train)

1. Mots vides (stop words) :

Supprimer les mots très courants qui n’apportent aucune information utile (par exemple les articles définis).

* 1. Utiliser une liste de mots vides dans une langue particulière.

(sklearn.feature\_extraction.text import ENGLISH\_STOP\_WORDS).

* 1. Supprimer les mots très fréquents.

CountVectorizer(min\_df=5, stop\_words =”English”).fit(text\_train)

1. Pondérer les données avec tf-idf ( term frequency inverse document frequency):

Donner un poids élevé aux termes qui apparaissent souvent un document, mais pas dans de nombreux documents du corpus (très représentatif dans un certain document).

* 1. TfidfTransformer

Score Tf-idf (w, d) =tf\*log+1

N le nombre de document

Le nombre de document dans le quel le mot w apparait.

Tf Le nombre d’occurrences du mot w dans le document considéré d

* 1. TfidfVectorize

Cette norme euclidienne implique que la longueur d’un document, c.-à-d. le nombre de mot qui le composent, ne change pas la représentation vectorisée.

### Des sacs avec plusieurs mots (n-grammes) :

L’ordre est totalement laissé de côté. Exemple « c’est mal, pas bien du tout » et « c’est bien, pas mal du tout ».

Il existe plusieurs manières de capturer ce contexte dans une représentation par sac de mots :

Méthode 1 : Extraction des caractéristiques countVectorizer et TfidfVectorizer + à prendre en compte non seulement la fréquence d’un certain *token*, mais aussi celle de paires (bigrammes) ou de triplets (trigrammes) ou n-grammes réglable avec le paramètre

|  |  |
| --- | --- |
| Réglable avec le paramètre ngram\_range(min\_token, max\_token) | |
| countVectorizer | TfidfVectorizer |

Utilise des séquences plus longues fournit généralement bien davantage de caractéristiques, et notamment des caractéristiques plus spécifiques => pas de bi-grammes communes entre les deux phrases (via toarray()) + Conduit une explosion du nombre de caractéristiques et entrainerait un sur-apprentissage, puisqu’il y’a plusieurs caractéristiques spécifiques.

En principe :

Nombre de bigrammes =

Nombre de trigrammes = …

+/-Méthode simple

Méthode2 : Tokenisation avancé, racinisation et lemmatisation

Mot + ing, verbe, pluriel, singulier … le mot a plusieurs formes et les distingué ne ferait qu’accroitre le su-apprentissage

Solution

|  |  |  |
| --- | --- | --- |
| Des formes de normalisation qui essaient d’extraire une certaine forme normale à partir d’un mot. | | |
| Racinisation (stemmer) | Lemmatisation | Correction orthographique |
| Représenter le mot à partir de la racine afin de repérer les éléments d’une même famille (méthode heuristique) | Utiliser un dictionnaire de formes connues (un système explicite et non automatisé) pour déterminer le rôle du mot dans la phrase |  |
| -Limiter à l’ajustement du mot à sa racine.  -Normalise mal les mots.  -Réduit les occurrences (meeting = meet) | +Trouver la forme verbale de base correcte.  +Arriver à normaliser les mots.  +Arriver à faire ma différence entre meeting et meet.  +Complexe.  +Produit de meilleurs résultats.  +Peut être perçue comme une forme de régularisation (elle associe en quelque sorte certaine caractéristique). | |

Méthode 3 : l’allocation de Dirichlet latente (LDA)

Essaie de trouver des groupes de mots (les sujets, ou topics).

Méthode 4 : RNN

# **Chapitre 7 : Pour conclure**

# Aborder un problème d’apprentissage automatique :

1. Se poser les bonnes questions :

Il faut réfléchir aux types de questions auxquelles vous voulez répondre.

* Avez-vous besoin d’une démarche exploratoire ou de trouver quelque chose d’intéressant dans les données ?
* Avez-vous déjà un objectif particulier présent à l’esprit ?

Vous devriez réfléchir en premier lieu non pas à construire un système, mais à comment définir et mesurer votre réussite, et à songe à l’impact qu’une solution efficace et réussie aurait sur votre activité ou vos objectifs de recherche.

1. définir le problème à résoudre
2. avoir les bonnes données et les mettre en forme
3. mesurer la réussite et l’impact de la solution

Mesurer les performances de l’algorithme via une métrique « business », comme l’augmentation du profit ou la réduction des pertes.

1. batir un prototype fonctionnel (choix du modèle)
   1. analyser ces modèles

Analyser les erreurs faites par un modèle fournit de bonne indication sur ce qui manque dans les données, sur quelles données supplémentaires il faudrait collecter, ou encore sur la manière dont la tache devrait être reformulée afin de rendre le processus d’apprentissage automatique plus efficace

Meilleur que d’exécuter à l’infini des recherches sur grille afin d’affiner les paramètres.

Exemple détection de fraude :

* Comment puis-je mesurer la validité de ma prédiction de fraudes ?
* Est-ce que je dispose des bonnes données pour les faire évaluer par l’algorithme ?
* Si je réussis, quel sera l’impact économique de ma solution ?

# **tester des système de production**

Le Test A/B et l’algorithme du bandit manchot.

Utiliser un modèle probabiliste puis utiliser l’apprentissage automatique ou l’inférence probabiliste pour **déterminer le niveau de confiance à accorder à chaque mesure**, et donc raisonner sur la détermination de la meilleure **estimation possible quant à la position de l’utilisateur** (exemple GPS et prédiction de la prochaine positon) Utiliser un **langage de programmation probabiliste** tel que PyMc et STAN

#### **Apprentissage automatique**

Système dans le quel nous voulons retrouver des réponses à une certaine requête en les classant selon leur pertinence.

Exemple :

* Fonctionnement des moteurs de recherche.

Système qui fournit des suggestions aux utilisateurs en fonction de leurs préférences.

Exemple :

* Facebook, Google+ (amis potentiels qui vous sont aimablement suggérés).
* Amazon et sites marchands (Les clients ayant acheté cet article ont également acheté …).
* La prédiction d’évolution dans le temps (cours de la bourse).

Erreur de généralisation

Peut s’exprimer comme la somme de trois erreurs très différentes :

* 1. BIAIS : Mauvaise hypothèses (Exemple :

- supposer que les données sont linéaires alors qu’elles sont quadratiques

Avoir plus de photos d’homme blanc que de d’homme noir (FB).

Si erreur Test/train >> au fur et à mesure + score minable => Biais >>>

* 1. Variance :

Un modèle avec beaucoup de degrés de liberté (nombre de valeurs aléatoires qui ne peuvent etre déterminées ou fixés par une équation, X + Y = 12).

* 1. Erreur irréductible : Due au bruit présent dans les données => nettoyer les données.

|  |  |
| --- | --- |
| La fonction de cout Vs la fonction de performance | |
| Elles sont différentes | |
| La fonction de cout peut etre régularisé | La fonction de performance n’est jamais régularisée |
| La fonction de cout doit avoir des dérivées permettant une bonne optimisation. | La fonction de performance doit vérifier si l’estimation est proche de l’objectif final. |

|  |  |  |
| --- | --- | --- |
| Exactitude (le taux de classification correctes). | Précision (taux d’observation classées comme positives et qui le sont effectivement). | Rappel (le taux d’observations positives qui sont bien détéctées). |

### Arrêt précoce

Quand l’erreur de validation cesse de croitre et commence à augmenter à nouveau (on commence à sur-ajuster). Avec l’arrêt précoce, on interrompt l’entrainement lorsque l’erreur de validation atteint le minimum.

# Méthodes de régularisation

L’objectif principal de la régularisation est de contrôler :

* Le surapprentissage dans l’apprentissage automatique
* L’erreur de type variance

La régularisation s’applique uniquement de l’entraînement du modèle.

## Régularisation L1 et L2 :

|  |  |
| --- | --- |
| L1 | L2 |
| La somme des poids en valeur absolue multiplié par une constante α | La somme des poids au carré multiplié par une constante α |
| Formules | |
| Une image contenant texte  Description générée automatiquementL1 essaie d'estimer la médiane des données. | Une image contenant texte  Description générée automatiquement  L2 essaie d'estimer la moyenne des données. |
| Intuition | |
| Si on a beaucoup de features, on peut penser que certaines sont inutiles.  Contraindre les coefficients à se rapprocher de zéro (qu’ils sont totalement ignorés par le modèle). Cela permet aussi de mettre en évidence les caractéristiques les plus importantes.  Permet d’effectuer une feature selection automatique (trouver des sous-ensembles de données pertinents).  Facile à interpréter du fait qu’elle sélectionne un sous ensemble de caractéristique. | La méthode pénalise le carré de la norme L2, ou encore la longueur euclidienne de W.  En pratique, elle est la plus utilisée.  L2 a tendance à réduire les coefficients de façon régulière.  Utile lorsque vous avez des caractéristiques colinéaires/co-dépendantes.  Sensible aux valeurs aberrantes. |
| Commun | |
| La régularisation impose une contrainte sur les poids du modèle. Une sorte de compromis entre la recherche d’un bon ajustement et le maintien d’une valeur basse pour certains poids lorsque les exposants sur les caractéristiques augmentent.  Les petits poids conduisent à des hypothèses plus simples, qui s’avèrent plus généralisables.  A mesure que le jeu d’entrainement augmente, l’effet de la régularisation diminue. | |

## Early stopping :

Interrompre l’entraînement lorsque ses performances sur le jeu de validation commencent à baisser.

Il est recommandé d’associer l’arrêt prématuré avec d’autre technique de régularisation.

## Dropout :

Chaque neurone a une probabilité p (taux d’extinction) d’être temporairement éteint (ignoré au cours de cette étape d’entrainement).

A chaque étape d’entrainement génère un réseau de neurones uniques, soit (où N est le nombre total de neurone), non indépendant car ils partagent leurs poids.

## Régularisation max-norm :

## Batch normalisation :

## Augmentation des données :

Utiliser des instances d’entraînement existantes pour en générer de nouvelles (réalistes), grossissant ainsi artificiellement la taille du jeu de données.

L’ajout du bruit blanc ne contribue pas à l’apprentissage (donc non).

Décaler, pivoter, redimensionner, modifier le contraste, ce qui rend le modèle plus robuste à la position, l’orientation, et la taille des instances, l’éclairage.

Si les instances sont symétriques, on peut retourner horizontalement les images.

# Fonctions d’activations

* Transformer non linéairement la combinaison formée par les entrées
* Imposer un intervalle aux données [0-1] ou [-1, 1]

# Fonctions de perte

Qualifie jusqu’à quel point un modèle est proche de l’idéal.

Une métrique qui calcule l’erreur que fait notre modèle, on peut lui ajouter une régularisation.

La recherche de cet idéal équivaut à trouver les paramètres (poids et biais) qui minimisent l’erreur.

Les paramètres peuvent être déterminés de façon analytique ou par des algorithmes d’optimisation itératifs.

### Fonctions de perte pour la régression

#### Erreur quadratique moyenne (MSE).

Fonction convexe, mais lorsqu’il s’agit de considérer multiples couches cachées, la notion de convexité ne tient plus.

Optimiser la MSE est équivalent à celle de la moyenne.

Sensible aux valeurs aberrantes.

Erreur absolue moyenne

Calcule la moyenne de l’erreur absolue sur l’ensemble du jeu.

#### Mean Squared Log Error (MSLE)

#### Erreur absolue moyenne en pourcentage (MAPE)

Remarque :

Les fonctions MSLE et MAPE méritent d’être prise en compte si l’intervalle des sorties varient fortement. Mais, en règle générale, on normalise les données d’entrées à l’intérieur d’un intervalle approprié puis on utilise MSE ou MAE.

### Fonctions de perte pour la classification

#### Perte hinge :

Classification binaire stricte, classifieur {0, 1} ou classifieur{-1, 1}.

Utilisée chez les SVMs.

Il existe des variantes pour la classification multi-classe : One-vs-All, One-vs-One.

Fonction convexe.

#### Perte logistique – Cross entropie - Log Loss :

N’est pas une classification stricte mais s’intéresse à la probabilité.

La dernière couche doit être une softmax, car contrairement à la sigmoïde, elle modélise les dépendances entre valeurs de sorties.

Optimiser le maximiser de vraisemblance

#### Divergence Kullback-Leibler :

La divergence de Kullback-Leibler est une mesure de dissimilarité entre deux distributions de probabilités

La classification multiclasse avec la fonction Softmax

## le word embedding. Le word embedding (plongement de mots) désigne un ensemble de méthode d’apprentissage visant à représenter les mots d’un texte par des vecteurs de nombres réels.

Problématique :

Un algorithme a été développé pour prédire le prix des maisons en real-time En novembre 2021, Zillow ferme le service real time car l'algorithme a suréstimé le prix de chaque propriété, engendrant de grosses pertes pour l'entreprise Si vous étiez chef de projet / advisor data scientist, qu'elles sont les recommendations que vous auriez faites pour vous assurer de la bonne mise en production du modèle.

Real time, mise à jour du modèle, continuel learning (not real time plus pour des fenetres temporelles)

Interpreter la decision du modèle : lime, shap (modification sur l’entrée pour voir comment impacter le score, plein de test) –> mini base de donnée, regression linéaire, analyser les poids w, pour connaitre les features importantes.

Confiance du modèle,

Pour un exemple donnée

* Feature seclection importante

Cas de l’image: Grad cam, bclasse activation map

Avant le soft, reponderer la feature

Phrase, a titre indicadif,

2019 – 2021: y a une pandémie 🡪 données ont pu change <!> gros drift temporal

Tester le modèle après les périodes 2008, robuste à ça

Feature selection

Se choisir un signal corrélé avec l’évolution du produit vendu