

Classification

Quentin Fournier <quentin.fournier@polymtl.ca>

Les diapositives ont été créées par Daniel Aloise
<daniel.aloise@polymtl.ca>

19 mai 2020

Classification

- Comment peut-on résoudre le problème de classification avec ce que l'on a vu en classe ?

Classification

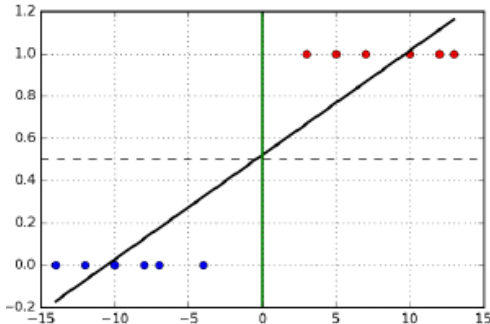
- Comment peut-on résoudre le problème de classification avec ce que l'on a vu en classe ?
- On pourrait utiliser la régression linéaire en convertissant les classes en chiffres :

- homme = 0 / femme = 1
- spam = 1 / non-spam = 0
- cancer = 1 / bénin = 0

Trjs la classe d'intérêt avec un 1

- 0/1 fonctionne pour les classifieurs binaires.
- Par convention, la classe d'intérêt est dite "positive" et prend la valeur 1 (la classe "négative" 0).

Classification



Skiena, 2017

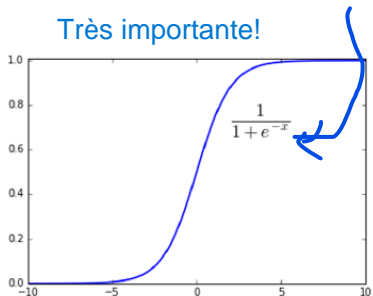
- La droite de la régression coupe ces classes, même s'il existe un séparateur.
- Ceci est dû à la minimisation de l'erreur carrée.
- La **régression logistique** est une première méthode pour trouver une fonction de séparation de deux classes.

La fonction *sigmoïde*

- La régression logistique vise à convertir une valeur continue en une valeur de probabilité $\in [0, 1]$.
quand on a juste 1 attribut

- Fonction sigmoïde :

- $f(0) = \frac{1}{2}$
- $f(\infty) = 1$
- $f(-\infty) = 0$



Skiena, 2017

- La fonction sigmoïde donne la probabilité de X_i d'appartenir à une classe particulière. proba d'appartenir à la classe 1

La fonction *sigmoïde*

quand on a plusieurs attributs

- Pour étendre la fonction sigmoïde à d attributs, on fait :

$$f(X_i, w) = \sum_{j=1}^d w^j X_i^j,$$

et ensuite :


$$h(X_i, w) = \frac{1}{1 + e^{-f(X_i, w)}}$$

Fonction d'erreur - *entropie croisée*

- La régression logistique utilise une fonction d'erreur différente de celle de la régression linéaire : l'**entropie croisée**.

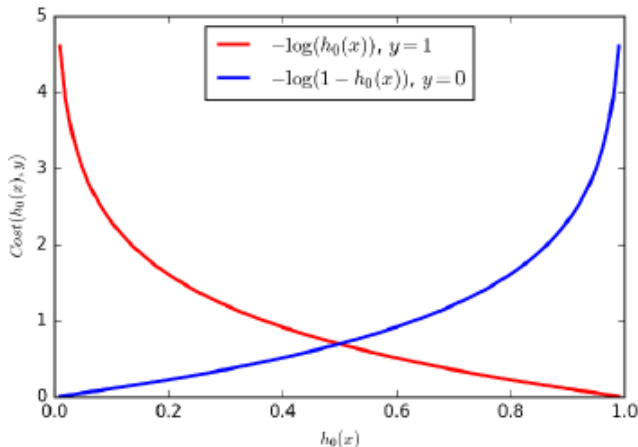
$$J(w) = \frac{1}{n} \sum_{i=1}^n -y_i \log \underbrace{h(X_i, w)}_{\text{notre prédiction}} - (1 - y_i) \log(1 - h(X_i, w))$$

ex: 1

rappelez-vous que y est binaire.  entre 0 et 1
si on a 1 on est 100% sûrs

- Alors, un seul des termes de la somme est actif pour chaque enregistrement.

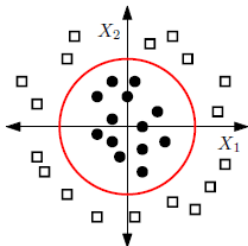
Fonction d'erreur - *entropie croisée*



Skiena, 2017

Fonction d'erreur - *entropie croisée*

- L'entropie croisée est aussi une **fonction convexe** !
- Ainsi, nous pouvons trouver le meilleur séparateur entre deux classes avec **la méthode du gradient**. **trouver le minimum global**
- L'erreur est nulle seulement si les classes sont linéairement séparables **si les données peuvent être séparées par un plan alors ce plan sera trouvé par la régression logistique.**
- Afin d'utiliser la régression logistique pour séparer des classes non linéairement séparables, il faut ajouter des colonnes non linéaires (ex. x^2 , \sqrt{x}) dans X .
pour séparer les données non linéairement séparables



Classification - Problèmes

① Ensemble d'entraînement déséquilibré

- ex. identification d'un terroriste.
- Considérons la droite de séparation optimale pour les classes grossièrement déséquilibrées, disons 1 exemple positif contre 1,000,000 exemples négatifs. *être terroriste* *ne pas être terroriste*
- La meilleure droite trouvée par la régression logistique essaiera d'être très loin du grand groupe des gens non terroristes au lieu de se placer entre les classes.
- Présence des faux négatives ! *TT temps prédire des négatifs*
- Même classer tout le monde comme non terroriste aura peu d'impact pour l'entropie croisée.
- **Solution (si possible)** utilisez le même nombre d'exemples positifs et négatifs.

Classification - Problèmes

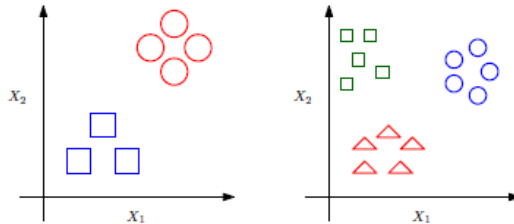
- Travaillez plus fort pour trouver des membres de la classe minoritaire.
- Supprimer des éléments de la classe majoritaire. problème on perd des données, qui peuvent être utiles pour notre prédiction
- Pesez plus lourdement les données de la classe minoritaire, mais méfiez-vous du surapprentissage ou **overfitting**.
- Répliquer les membres de la classe minoritaire, idéalement avec une perturbation aléatoire.
ex on a 10 terroristes pr 10 millions de non terroristes. Ex: on va

perturbation aleatoire: on échantillonne avec les valeurs gaussienne.

Classification - Problèmes

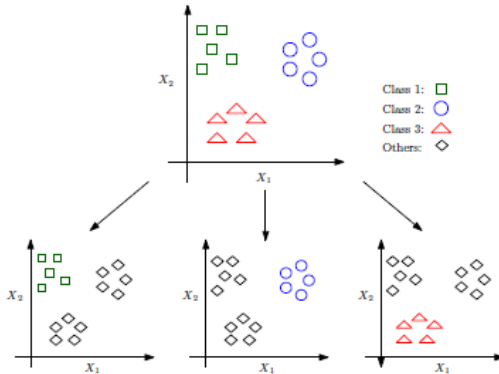
② Classification multi-classe

- Les tâches de classification ne sont pas toujours binaires
- Un film donné est-il une comédie, un drame, une sci-fi, un documentaire, etc. ?



Skiena, 2017

Classification multi-classe



Skiena, 2017

- Sélectionnez la classe de probabilité la plus élevée comme étiquette prédite
- La classification multi-classe devient beaucoup plus difficile à mesure que le nombre de classes augmente.

ex: prédire si chat chien plante
prédire leurs sous-classe: race..

Classification - Problèmes

③ Fonctions de partition : normaliser les valeurs de sorties pr obtenir des probabilités

- Les classifieurs binaires indépendants ne produisent pas de vraies probabilités (la somme ne vaut pas 1).
- Une possible solution :

$$T = \sum_{\ell=1}^k h^{\ell}(X_i, w)$$

où k est le nombre de classes

$$p(\ell) = h^{\ell}(X_i, w)/T \quad \forall \ell = 1, \dots, k$$

- La **régression multinomiale (softmax)** combine les classifieurs au niveau de l'entraînement.

Apprentissage machine

regression lineaire et logistique + les algos qui apprenent à partir d'exemples
tous les algorithmes qui resoudent des problemes complexes

*Machine Learning is the study of computer algorithms
that improve automatically through experience.*

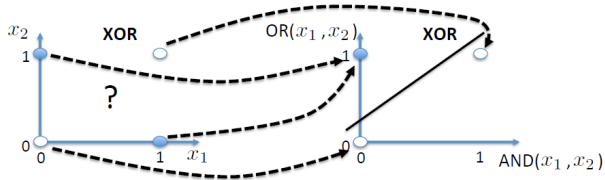
Tom M. Mitchell, 1997

*Most of what is being called “AI” today, particularly in
the public sphere, is what has been called “Machine
Learning” (ML) for the past several decades.*

Michael I. Jordan, 2018

Limitations de la régression

- On a besoin d'attributs non linéaires (ex. $x_1 x_2$) pour que la régression puisse bien séparer des données non linéairement séparables.



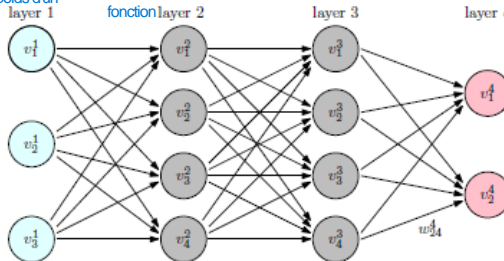
- Cela peut impliquer une énormité de combinaisons d'attributs à tester.
- Et si ces combinaisons pouvaient être extraites toutes seules ?

Apprentissage profond DEEP LEARNING

- Le domaine le plus en ^{populaire} vogue de l'apprentissage machine implique aujourd'hui de grandes architectures de réseaux neuronaux profonds.
- Les **couches cachées** créent de compositions de fonctions non linéaires \Rightarrow **un plus grand pouvoir de représentation.**

données en entrée
ex: taille, âge, poids d'un individu

régression
linéaire + une
fonction layer 2



Skiena, 2017

Apprentissage profond

- Faire **apprendre** le réseau signifie définir les valeurs des coefficients w . trouver les valeurs de w pr résoudre le problème, donc minimiser l'erreur
- Plus il y a de connexions entre les neurones, plus il y a des paramètres à apprendre.
- En principe, l'apprentissage consiste à analyser l'**ensemble de données d'entraînement étiquetées** et à ajuster les coefficients de telle sorte que les nœuds de sortie génèrent quelque chose proche de y_i lorsqu'ils le réseau est alimenté avec X_i .

Apprentissage profond

- Pourquoi les réseaux de neurones sont-ils si efficaces ?

Apprentissage profond

- Pourquoi les réseaux de neurones sont-ils si efficaces ?
- Personne ne le sait vraiment.

Apprentissage profond

- Pourquoi les réseaux de neurones sont-ils si efficaces ?
- Personne ne le sait vraiment.
- Il y a des indications que pour de nombreuses tâches, la complexité de ces réseaux n'est pas vraiment nécessaire.

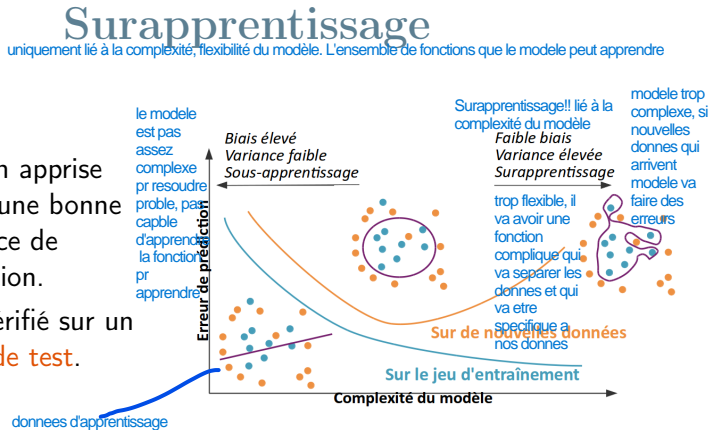
Apprentissage profond

- Pourquoi les réseaux de neurones sont-ils si efficaces ?
- Personne ne le sait vraiment.
- Il y a des indications que pour de nombreuses tâches, la complexité de ces réseaux n'est pas vraiment nécessaire.
- Les réseaux de neurones semblent fonctionner par **surapprentissage**, trouvant un moyen d'utiliser des millions d'exemples pour s'adapter à des millions de paramètres.

EXAMEN

Définition sous-apprentissage et surapprentissage

- La fonction apprise doit avoir une bonne performance de généralisation.
- Cela est vérifié sur un ensemble de test.



Source : [https://openclassrooms.com/fr/courses/](https://openclassrooms.com/fr/courses/4297211-evaluez-et-ameliorer-les-performances-dun-modele-de-machine-learning/)

4297211-evaluez-et-ameliorer-les-performances-dun-modele-de-machine-learning/

Apprentissage profond

- L'apprentissage profond généralement évite le pire comportement du surapprentissage, peut-être en utilisant des moyens moins précis d'encoder la connaissance.

arreter apprentissage un peu d'avance pr eviter le surapprentissage quand modele devient complexe

Apprentissage profond

- L'apprentissage profond généralement évite le pire comportement du surapprentissage, peut-être en utilisant des moyens moins précis d'encoder la connaissance.
- L'apprentissage profond est une technologie très excitante, bien qu'il soit mieux adapté aux domaines avec d'énormes quantités de données étiquetées.

Apprentissage profond

- L'apprentissage profond généralement évite le pire comportement du surapprentissage, peut-être en utilisant des moyens moins précis d'encoder la connaissance.
- L'apprentissage profond est une technologie très excitante, bien qu'il soit mieux adapté aux domaines avec d'énormes quantités de données étiquetées.



PyTorch

- TensorFlow et PyTorch facilitent la construction de modèles d'apprentissage profond.
- Pourtant vous aurez besoin de bons GPU pour faire tourner vos expériences.

Apprentissage profond

- Chaque **neurone** j du réseau calcule une fonction non linéaire $\phi(v_j)$ où v est donné par :

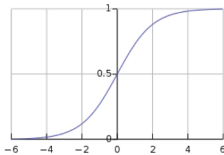
chaque neurone calcule un v_j

$$v_j = \sum_i w_{ij} t_{ij}$$

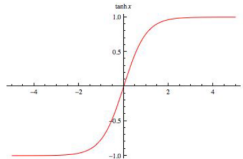
où : t_{ij} est l'input au neurone j avenant du neurone i et w_{ij} est le poids donné à cet input.

Apprentissage profond

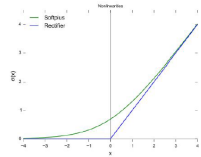
- Le neurone j est dit **activé** en fonction de la valeur $\phi(v_j)$ et de la **fonction d'activation** du neurone :



Sigmoid



Tanh



ReLU

IVADO, 2018

Backpropagation

La sortie depend de la couche d'avant, qui depend de la couche d'avant....

- Les réseaux de neurones sont entraînés par un algorithme de type descente de gradient (préf. stochastique).
- Les changements pour chaque exemple d'entraînement sont ramenés à des niveaux inférieurs.
- Les fonctions d'activation non linéaires aboutissent à une fonction d'erreur non convexe, mais son optimisation produit généralement de bons résultats.
- Beaucoup plus de détails dans le cours **INF8225 : I.A. : tech. Probabilistes et d'apprentissage** - offert à l'hiver.