agaetis
Big Data & Data Science

Introduction au Machine Learning
Régression linéaire

Léo Beaucourt pour Clermont'ech APIHour #42

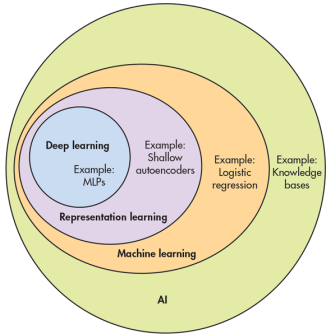
Pourquoi la régression linéaire?

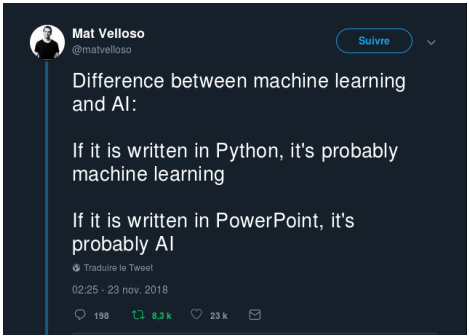
- La régression linéaire: le "*Hello world!*" du *ML*
- Résolution d'un problème de Data science: *Prédiction d'un prix*
- En pratique: *Python*, *Jupyter*. Packages *numpy pandas* et *matplotlib*.
- Pas de (trop) de math ...

Allez, on démarre en douceur ...


DON'T
PANIC

Machine learning, *qu'est ce que c'est?*





- **AI**: Domaine d'étude $\Rightarrow$ abus de langage (NN, RL)
- **ML**: Algorithmes/outils développés dans le cadre de la recherche sur l'IA

Machine Learning: Définitions

- **Arthur Samuel:**
  - *The field of study that gives computers the ability to learn without being explicitly programmed.*
- **Tom Mitchell:**
  - *A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.*

- **L'idée:** Une machine apprend *seule* à réaliser une tache complexe à l'aide de processus itératifs simple.

ML: Les principaux types d'apprentissage

Supervisé

- Utilise des données *labélisées*
- La machine apprend par l'exemple
- *Prédis* le résultat pour de nouveaux événements
- Problèmes de prédictions et de classification
- Regression linéaire et logistique
- Réseaux de Neurones
- Arbres de décisions

Non-supervisé

- Données non *labélisées*
- La machine apprend par elle même à identifier une structure
- Évaluation des performances compliqué.
- Problèmes de classification, réduction de dimensions
- K-means
- Analyse en Composante Principale

Par renforcement

- Un agent A, effectue une action Ac, l'environnement E lui renvoie une récompense.
- Récompenses à court et long terme
- Utilisé par Deepmind (alphaGo)

À quelles problèmatiques répond le Machine Learning?

- **Prédictions** Prédire une valeur continue à partir de caractéristique données
- **Projections** Prédictions spécifique de séries temporelles: $y = f(y(t-1), y(t-2), \dots)$
- **Classifications** Prédire la classe (discret) d'un objet en fonction de ses caractéristiques
- **Segmentations** Regrouper des objets par similarité dans l'espace des variables utilisé
- **Compréhensions** Comprendre l'importance de variables d'intérêt dans un contexte donné

La regression linéaire

- Déterminer une relation *linéaire* entre *input(s)* (features) et *output*:

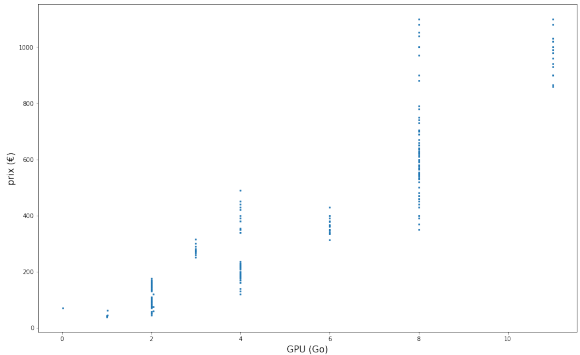
$$\Rightarrow \textbf{Apprentissage Supervisé}$$

- Prédiction d'une valeur **continue** (e.g. non discrète, non catégorielle)
- Applications:
  - Recherche de corrélations
  - En science, modélisation de phénomènes (physiques, biologiques, ...) après mesures
  - Dans le domaine médical: les études épidémiologique
  - Dans la finance/économie: prédictions des tendances, *Capital Asset Pricing Model*
  - ...

Sujet Data Science $\Rightarrow$ Premier algorithme à tester!

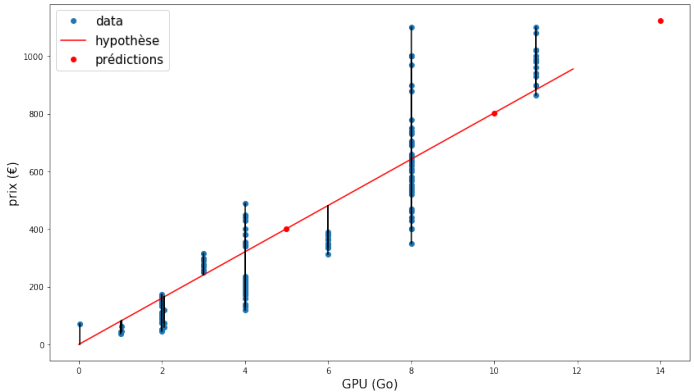
Un exemple: le prix d'une carte graphique

- La propriété principale d'une carte Graphique: valeur de **GPU**
- Jeu de données: $\{GPU; prix\}$:



On peut maintenant faire une prédiction

- Quel serait le prix de cartes avec 5, 10 et 14 Go de GPU?

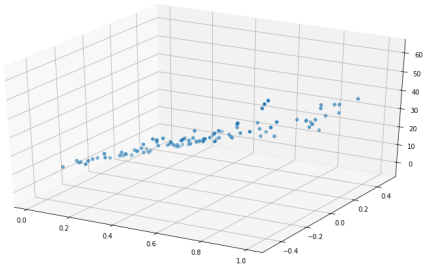


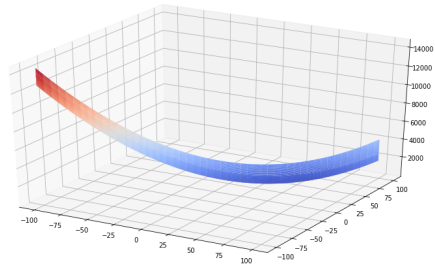
- On pourra les vendre autour de 400, 800 et 1100 euros!

La regression linéaire multivariables

- Le principe est le même, mais avec plusieurs variables x_i (donc plusieurs paramètres θ_i):

$$\hat{y} = \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i$$



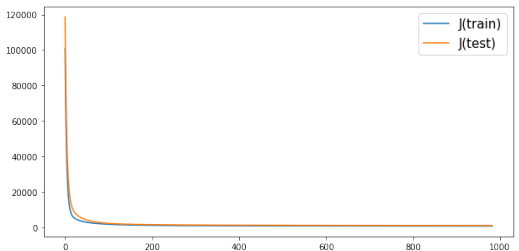


Affinons notre modèle de carte graphiques

- Plus de features: chipset, fréquence, consommation, ...
- Il va falloir explorer et nettoyer les données:
  - Gestion des données manquantes / abbérantes
  - *Features engineering*
  - Normaliser le dataset (pour accélérer la descente de gradient)

Régression linéaire multivariables: Résultats

- Modèle simple: $err \approx 100$ (biais)
- Modèle multivariable: $err \approx 40 \ / \ 50$ (variance)



Pour conclure sur la régression linéaire

- **Regression Linéaire:** $\hat{y}$ est une valeur *continue*
  - Valeur discrète: **Regression Logistique** (*classification*)
- Le résultat $\hat{y}$ dépend **linéairement** des variables x_i si:

$$\hat{y} = \theta_1 x_1 + \cdots + \theta_n x_n = \sum_{i=1}^{n} \theta_i x_i$$

- **Apprentissage supervisé**: y est connu pour chaque x_1 dans le jeu de données d'entrainement
- Facile à implémenter (encore plus avec Scikit-learn ...), rapide: bon point de départ sur un sujet

agaetis

Big Data & Data Science

Merci !

Des questions?

Léo Beaucourt

“There is a theory which states that if ever anyone discovers exactly what the Universe is for and why it is here, it will instantly disappear and be replaced by something even more bizarre and inexplicable.

There is another theory which states that this has already happened.”

Douglas Adams