Apprentissage statistique

TP7 – Méthodologie du traitement de données

Olivier Schwander <olivier.schwander@lip6.fr>

Certificat Big Data
UPMC - LIP6

2018-2019

Résumé

Différentes tâches

- Classification
- Régression
- Détection d'évènements
- Segmentation
- Recherche d'information
- Recommandation

Démarche générale

1. Données: chargement, étude, filtrage
2. Méthodes: choix, compréhension, paramètres
3. Évaluation: score, temps, mémoire, interprétation

Données

Première étape

1. Charger les données
2. Étudier les données
3. Filtrer, nettoyer, choisir les données

Méthodes

Énormément de méthodes disponibles un peu partout, de la plus simple à la plus complexe.

Rasoir d'Ockam

- Ne pas rajouter de complexité inutile

Paramètres

- Choix délicat mais indispensable

Évaluation

Dépend de la tâche à accomplir

Différentes mesures

- Précision
- Faux positifs, vrais positifs

Ne pas oublier

- Temps, mémoire
- Autres contraintes

Données: chargement

Formats faciles: directement des matrics

- Texte brut: `numpy.loadtxt`
- Format numpy: `numpy.load`
- Format Matlab: `scipy.io.loadmat`
- *Comma Separated Values*: `pandas.read_csv`

Formats standards: à transformer en matrices

- XML
- JSON: `json.load`

Formats baroques

- Format spécifique à un jeu de données
- Utiliser la documentation fournie

Données: apprentissage et test

Deux bases séparées

- **SÉPARÉES**

La séparation est parfois déjà faite

- Parfait

À faire

- Mélanger les données (avec `sklearn.utils.shuffle` par exemple)
- Découper: 80%-20%, 90%-10%
- Reproductible: sauvegarder la graine aléatoire (*random seed*), sauvegarder le découpage dans des fichiers

Données: études

Calculer des valeurs

- Moyennes, médianes, min, max, fréquences

Tracer des figures

- Histogrammes, courbes

Objectifs

- Repérer des valeurs aberrantes, des erreurs manifestes
- Voir si toutes les données sont utiles
- Tester des modèles très simples (utilisant une seule colonne par exemple)

Données: valeurs manquantes

Données manquantes

- Not a Number: `NaN`
- Codes d'erreur: -1, -9999

Supprimer ce qui gène:

- une colonne entière ? une ligne ?
- Attention à ne pas supprimer tout le contenu de la base !

Compléter les données:

- des 0
- la valeur moyenne, la médiane
- une valeur proposée par un expert

Données: *features engineering*

Travail sur les données originales

- Transformations non-linéaires: exp, log, somme, produit
- Sur une seule colonne ou entre plusieurs
- Remplacer des valeurs continues par des valeurs discrètes: histogramme, quantification, clustering
- Transformer des catégories en nombres

Exemple: prédiction de la valeur d'une maison

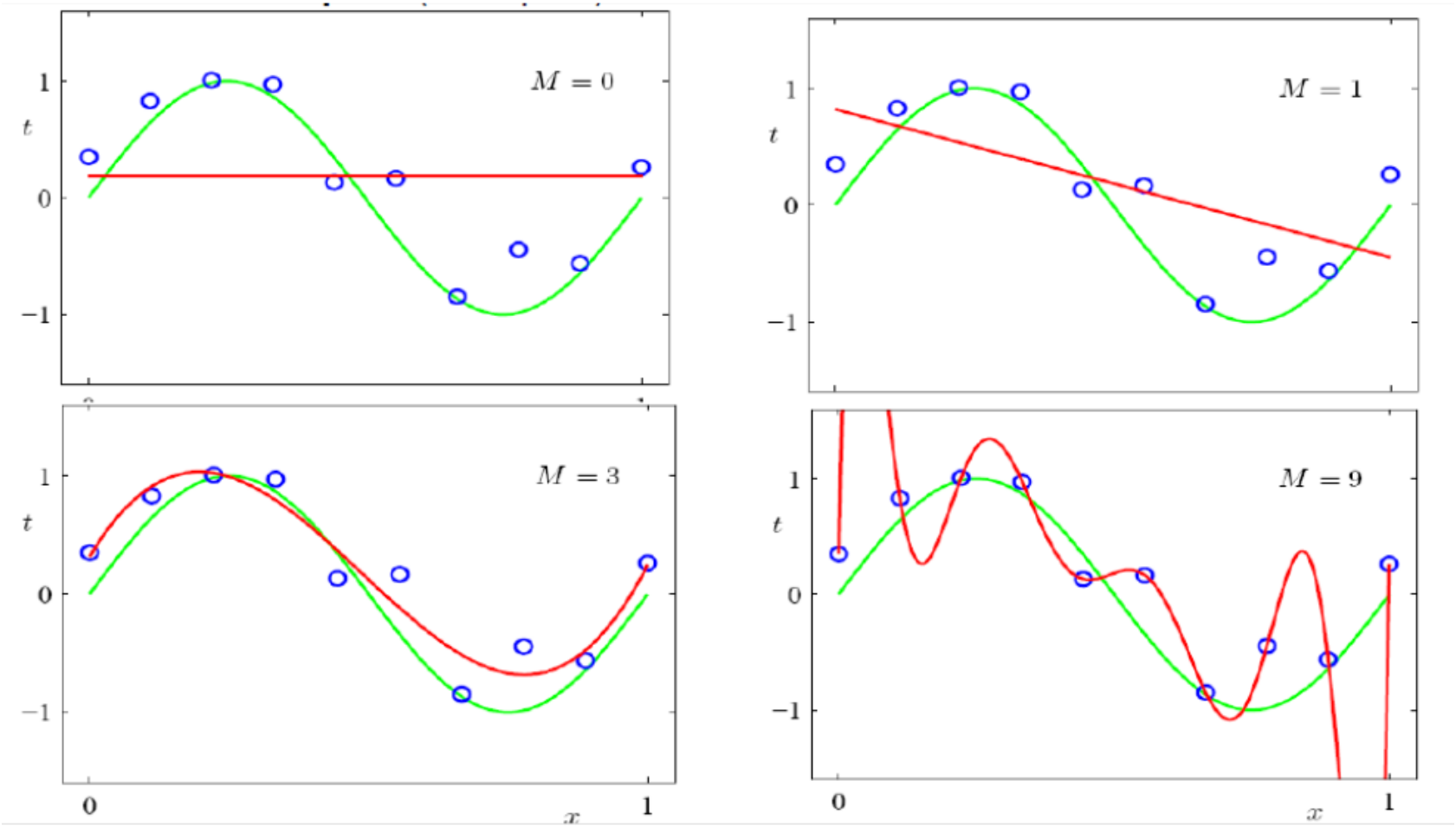
- Données: longueur et largeur de la maison
- Nouvelle *feature*: $X = \text{longueur} \cdot \text{largeur}$
- Modèles linéaire: $\text{prix} = \alpha X + \beta$

Méthodes: réflexion prélimnaire

En pratique

- Ne pas oublier les méthodes les plus simples !
- Ne pas trop subir les effets de mode
- Réfléchir aux contraintes de la méthode: mémoire, temps, quantité de données
- Comprendre les méthodes (pas forcément tous les détails, mais avoir une idée de ce qu'on calcule)

Méthodes: éviter le sur-apprentissage



Source: Ludovic Denoyer, cours FDMS

Méthodes: bibliothèques

Quelques exemples en Python

- *Machine learning*: `sklearn`
- *Modèles statistiques*: `Statsmodels`
- *Image et vision*: `skimage`

Comprendre

- Quel modèle calcule-t-on ?
- Quelle grandeur optimise-t-on ?

Méthodes: paramètres

Simplicité

- Rasoir d'Ockam
- Limiter le nombre de paramètres

Ajustement manuel

- Score
- Autres contraintes

Validation croisée

- Méthode automatique
- Découpage de l'ensemble d'apprentissage: une partie pour apprendre, une partie pour évaluer les paramètres

Évaluation: score

Précision

- Nombre de bonnes réponses

Précision, rappel

- Précision: nombre de documents pertinents parmi les documents retournés
- Rappel: nombre de documents pertinents retournés sur le nombre total de documents pertinents
- Score F1: $2\frac{\text{precision}\cdot\text{rappel}}{\text{precision}+\text{rappel}}$

Vrais positifs, faux positifs

Évaluation: autres coûts

Phase d'apprentissage et phase de prédiction

Classique

- Temps de calcul
- Mémoire utilisée

Et aussi

- Consommation électrique (smartphone, datacenter)
- Bande passante utilisée (réseau mobile)

Compromis

- *There is no free lunch*

Évaluation: interprétation

Que peut-on expliquer à un expert ?

- Qualité des données
- Utilité des différentes informations
- Explication des données