Structure et Dynamique des Réseaux

Cours 14: Fouille de données et dynamique des réseaux

Clémence Magnien, Lionel Tabourier, Fabien Tarissan

LIP6 – CNRS and Université Pierre et Marie Curie

prenom.nom@lip6.fr

Outline

Le problème de la prédiction de lien

Description du problème

On suppose l'ensemble de nœuds V fixé,

- l'évolution du graphe est connue de t_0 à t'_0
- quels liens vont apparaître/disparaître entre t_1 et t'_1 ?

Pertinence?

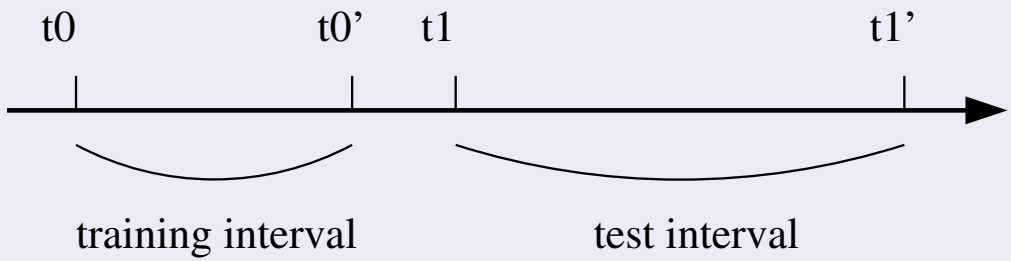
soit f la fraction de nouveaux liens faisant apparaître de nouveaux nœuds, on veut: $f \ll 1$

exemples: amitiés FaceBook? adresses IP par mesures radar?

Prédiction de lien

Principe

Liben-Nowell, Kleinberg - *JASIST, 2007*



Prédire les liens de $G[t_1, t_1']$ qui ne sont pas dans $G[t_0, t_0']$:

Trouver les propriétés des paires de nœuds qui rendent probable l'apparition d'un lien

Rq: paires non-liées bcp plus nombreuses que paires liées...
($\Rightarrow$ disparition de liens doit être formulée de manière différente)

Outline

Un problème de classification statistique

Illustrations: `scikit-learn.org` et cours d'A.Ng - *Introduction to Machine Learning*

Qu'est-ce que la classification statistique?

- classification:
  nombre fixé de familles, affecter une famille à un objet
- statistique:
  basée sur la comparaison des caractéristiques de l'objet à une population d'objets déjà classés

Exemples

- famille d'animaux selon les dimensions, l'apparence
- diagnostic médical selon les symptômes
- images: déduire ce qu'elle représente

Un problème de classification statistique

Illustrations: `scikit-learn.org` et cours d'A.Ng - *Introduction to Machine Learning*

Qu'est-ce que la classification statistique?

- classification:
  nombre fixé de familles, affecter une famille à un objet
- statistique:
  basée sur la comparaison des caractéristiques de l'objet à une population d'objets déjà classés

Et ici?

- **à classer:** paires de nœuds en **2 classes:** existe ou non
- **sources d'information:**
  structure du graphe, caractéristiques des nœuds, ...

Questions de fouille de données (*DataMining*)

But

Extraire de la connaissance de grands jeux de données

Un exemple, quelques questions

Statistiques sur populations (ex félins: chat sauvage, tigre, ...)
classer un animal à l'aide d'informations partielles?

des caractéristiques informatives:
d'autres utiles si associées à une autre:
d'autres peu ou pas:

- comment **sélectionner** les caractéristiques utiles?
- comment **pondérer** les différentes caractéristiques?
- quelle est la **limite à la qualité** d'une prédiction?

Questions de fouille de données (*DataMining*)

But

Extraire de la connaissance de grands jeux de données

Un exemple, quelques questions

Statistiques sur populations (ex félins: chat sauvage, tigre, ...)
classer un animal à l'aide d'informations partielles?

des caractéristiques informatives: taille, poids, couleur
d'autres utiles si associées à une autre: sexe (+ poids)
d'autres peu ou pas: âge

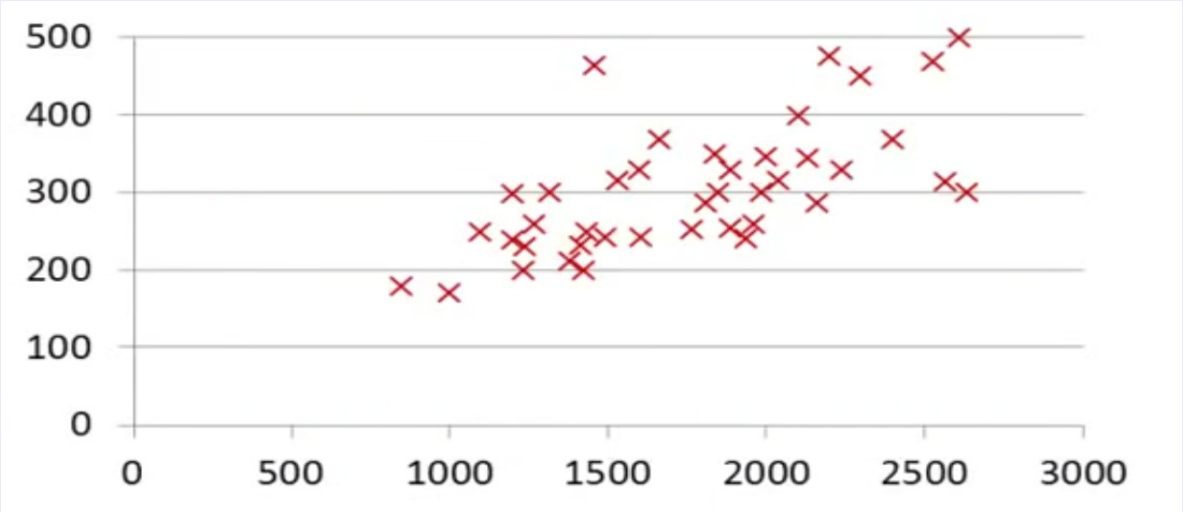
- comment **sélectionner** les caractéristiques utiles?
- comment **pondérer** les différentes caractéristiques?
- quelle est la **limite à la qualité** d'une prédiction?

Problème de régression

apprentissage supervisé:
on a des exemples pour lesquels le résultat est connu

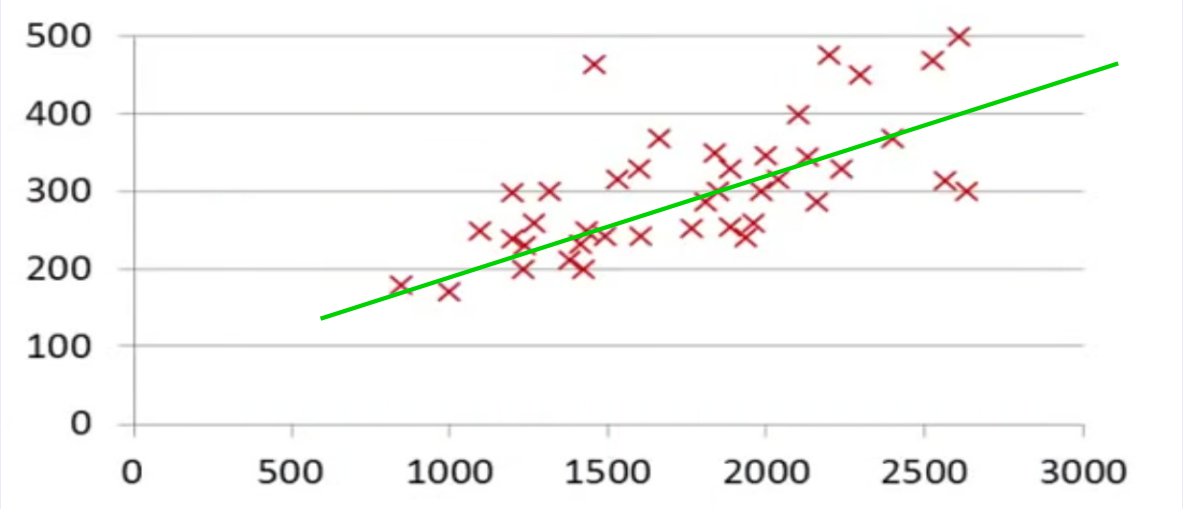
Illustration

Prix immobilier à Portland:
rechercher le prix (k$) en fonction de la surface (feet^2)



Problème de régression

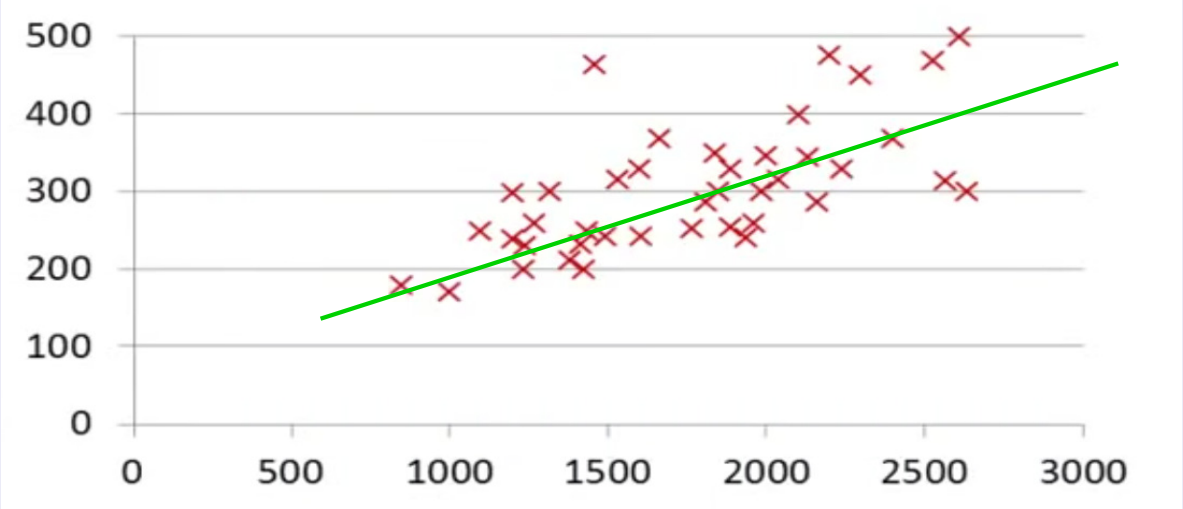
Modèle de régression linéaire $y = \theta_0 + \theta_1 . x$



Qu'est-ce qu'un bon modèle?

Problème de régression

Modèle de régression linéaire $y = \theta_0 + \theta_1.x$



Qu'est-ce qu'un bon modèle?

définir une fonction de coût à minimiser

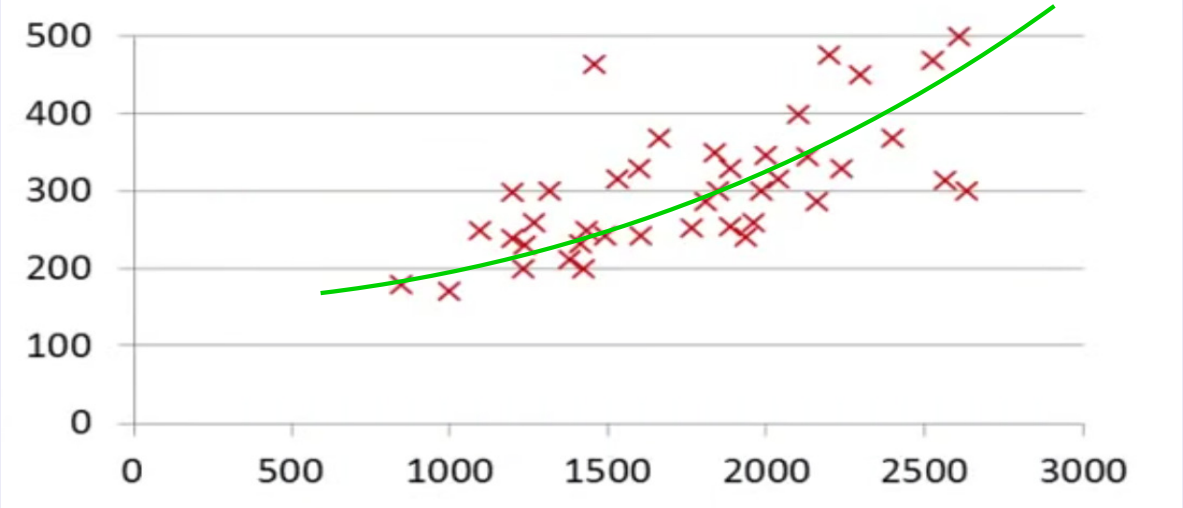
ex: erreur quadratique moyenne $\frac{1}{N}\Sigma_{i=1}^{N}(y_i - y'_i)^2$

si on a N enregistrements de données (x_i, y_i)

y'_i *: valeur du modèle en* x_i

Problème de régression

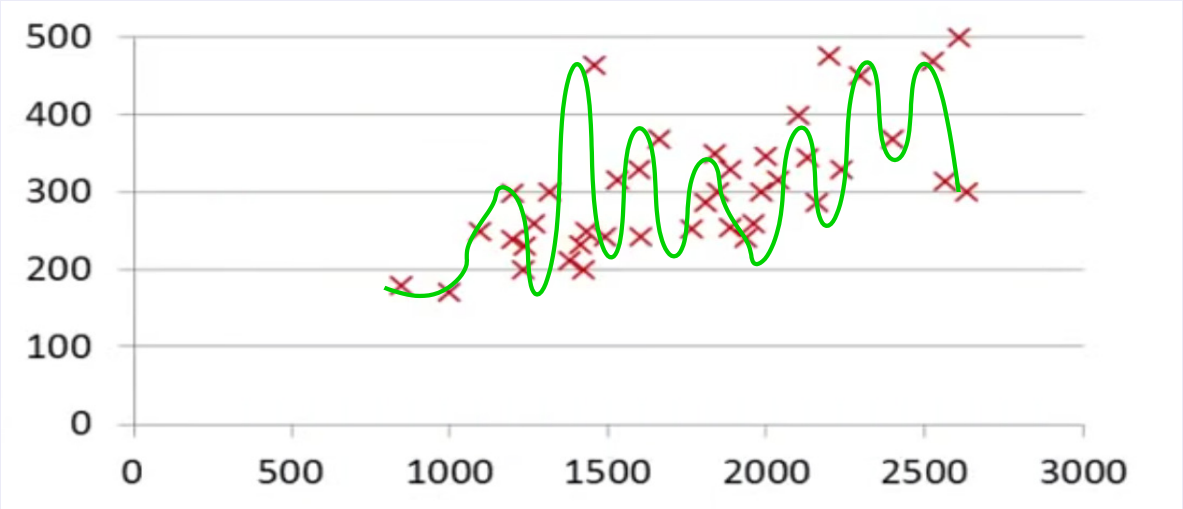
Polynome d'ordre supérieur $y = \theta_0 + \theta_1.x + \theta_2.x^2 + ...$



Minimisation du coût meilleure que linéaire (pourquoi?)

Problème de régression

Polynome d'ordre supérieur $y = \theta_0 + \theta_1.x + \theta_2.x^2 + ...$



Minimisation du coût meilleure que linéaire (pourquoi?)
mais modèle avec plus de paramètres...

problème du sur-apprentissage:
modèle peu extrapolable à d'autres données

Problème de régression

Sélection de modèle

Comment choisir entre les différents modèles?

Une partie des données utilisée en **ensemble de validation**:

- valeurs des y_i connues pour ces items
- items non-utilisés pour fixer les paramètres du modèle
- comparer y_i aux prédictions du modèle

Alors, si l'erreur désigne $\frac{1}{N}\Sigma_{i=1}^{N}(y_i - y_i')^2$:

Problème de régression

Sélection de modèle

Comment choisir entre les différents modèles?

Une partie des données utilisée en **ensemble de validation**:

- valeurs des y_i connues pour ces items
- items non-utilisés pour fixer les paramètres du modèle
- comparer y_i aux prédictions du modèle

Alors, si l'erreur désigne $\frac{1}{N}\Sigma_{i=1}^{N}(y_i - y_i')^2$:

- erreur d'apprentissage $\searrow$ avec le nombre de paramètres
- erreur de validation $\searrow$ puis $\nearrow$

Problème de régression

Ajout d'autres sources d'info

modèle quadratique $\Leftrightarrow$ deux caractéristiques: x_i et x_i^2

Ajouter d'autres caractéristiques?

nombre de pièces, d'étages, âge, distance au centre ville...

$$y = \theta_0 + \theta_1.x_1 + \theta_2.x_2 + \theta_3.x_3...$$

Problème de régression

Ajout d'autres sources d'info

modèle quadratique $\Leftrightarrow$ deux caractéristiques: x_i et x_i^2

Ajouter d'autres caractéristiques?

nombre de pièces, d'étages, âge, distance au centre ville...

$$y = \theta_0 + \theta_1.x_1 + \theta_2.x_2 + \theta_3.x_3 \ldots$$

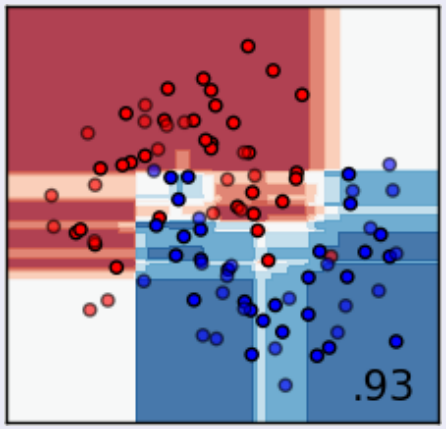
Choisir un modèle

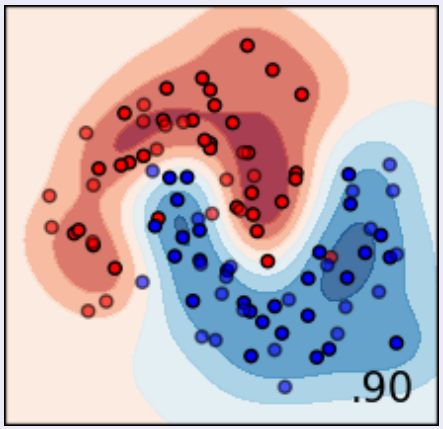
- choix du **type de modèle** et du **critère de coût**
- choix des **caractéristiques** et du **nombre de paramètres**

Problème de classification

apprentissage supervisé:
on a des exemples pour lesquels le résultat est connu

Représentation schématique





- comment tracer les frontières? **choix du modèle**
- combien de paramètres? **éviter le sur-apprentissage**
- comment définir l'erreur? **coût fausse classification**

Outline

Un exemple : arbres de classification

Ensemble des données : ensemble E_{tot} de points
$(x_1, x_2, ..., x_n, y)$

Principe

Construire un arbre depuis les données tel que:

- la **racine** correspond à l'ensemble des données E_{tot}
- chaque **nœud** i est un sous-ensemble des données E_i
- chaque **embranchement** en i réalise une partition des données de E_i selon une condition type $x_i \leq \alpha$ ou $x_i > \alpha$
- **les embranchements sont tels que chaque nœud-fils soit aussi homogène que possible**

Un exemple : arbres de classification

Comment faire...

- ... pour partager en sous-ensembles homogènes?
  critère de segmentation
  ex: minimiser diversité de Gini: $\Sigma_k f_k(1 - f_k)$
  avec f_k *fraction des éléments dans classe k*

- ... pour arrêter de la procédure?
  si la séparation n'améliore plus la prédiction
  ou **critère d'homogénéité** des feuilles de l'arbre
  ex: si une feuille à 90% des éléments dans une classe

Application à la prédiction de liens

Liben-Nowell, Kleinberg - *JASIST, 2007*

Pujari et Kanawati - *WWW 2012 companion*

Jeux de données

Réseaux de **collaborations scientifiques**:

- nœud = auteur, lien = co-publication
- publications dans *DBLP, arXiv, Medline...*
- nombre d'articles: qq milliers par an
- nombre d'auteurs: qq milliers

Protocole

Année A pour prédire **nouvelles** collaborations année $A + 1$

Quelques caractéristiques de prédiction

Caractéristiques structurelles locales

- nombre de voisins communs
- ressemblance des voisinages *(ex: indice de Jaccard)*

Caractéristiques structurelles globales

- nombre de chemins de longueur ν de i à j
- *hitting time* en j partant de i
- centralités des nœuds

Caractéristiques non structurelles

- mesures de similarité des nœuds i et j *(âge, spécialité...)*

Évaluer la qualité des prédictions

Métriques élémentaires (pour classification à 2 classes):

	prédiction: +	prédiction -
réalité: +	vrai positif	faux négatif
réalité: -	faux positif	vrai négatif

Métriques usuelles:

- **précision**, $\mathbf{Pr} = \frac{\#vp}{\#vp+\#fp}$
- **rappel** (*recall*), $\mathbf{Ra} = \frac{\#vp}{\#vp+\#fn}$
- **F-score**, $\mathbf{F} = \frac{2.\mathbf{Pr}.\mathbf{Ra}}{\mathbf{Pr}+\mathbf{Ra}}$ (compromis précision-rappel)
- et d'autres (*fall-out*, *AUC*,...)

Prédiction dans les grands réseaux sociaux:

Évaluer la qualité des prédictions

Métriques élémentaires (pour classification à 2 classes):

	prédiction: +	prédiction -
réalité: +	vrai positif	faux négatif
réalité: -	faux positif	vrai négatif

Métriques usuelles:

- **précision**, **Pr** $= \frac{\#vp}{\#vp+\#fp}$
- **rappel** (*recall*), **Ra** $= \frac{\#vp}{\#vp+\#fn}$
- **F-score**, **F** $= \frac{2.\mathbf{Pr}.\mathbf{Ra}}{\mathbf{Pr}+\mathbf{Ra}}$ (compromis précision-rappel)
- et d'autres (*fall-out*, *AUC*,...)

Prédiction dans les grands réseaux sociaux: **fort risque de FP**

Outline

Fixer le nombre de prédictions

Pujari et Kanawati - *WWW 2012 companion*

Supposons le nombre de liens à prédire inconnu...
méthodes usuelles inadaptées

Fixer le nombre de prédictions

Pujari et Kanawati - *WWW 2012 companion*

Supposons le nombre de liens à prédire inconnu...
méthodes usuelles inadaptées

Si une valeur d'attribut élevée correspond à un lien probable

Alors **classer les paires par score**:
pour T prédictions, choisir les T premiers éléments

Fixer le nombre de prédictions

Pujari et Kanawati - *WWW 2012 companion*

Supposons le nombre de liens à prédire inconnu...
méthodes usuelles inadaptées

Si une valeur d'attribut élevée correspond à un lien probable

Alors **classer les paires par score**:
pour T prédictions, choisir les T premiers éléments

Comment utiliser plusieurs attributs de classement?
Apprendre à ordonner

Des pistes

Méthodes non-supervisées: règles de consensus

exemple: méthode de Borda

$c_i(k)$: classement i de l'item k, N_i: nombre d'items dans i

score Borda (k) $= \Sigma_i \, (N_i - c_i(k))$

Des pistes

Méthodes non-supervisées: règles de consensus

exemple: méthode de Borda

$c_i(k)$: classement i de l'item k, N_i: nombre d'items dans i

score Borda (k) = $\Sigma_i (N_i - c_i(k))$

mais l'apprentissage supervisé est + contrôlé et + efficace

Méthodes supervisées

exemple: transformation 2-à-2

ordonner les items $\rightarrow$ est-ce que k est classé au-dessus de k' ?

retour à un problème de classification

Des pistes

Méthodes non-supervisées: règles de consensus

exemple: méthode de Borda

$c_i(k)$: classement i de l'item k, N_i: nombre d'items dans i

score Borda (k) = $\Sigma_i (N_i - c_i(k))$

mais l'apprentissage supervisé est + contrôlé et + efficace

Méthodes supervisées

exemple: transformation 2-à-2

ordonner les items $\rightarrow$ est-ce que k est classé au-dessus de k' ?

retour à un problème de classification

mais limite de taille

Comment faire sur de grands graphes?