Classification supervisée

Rémy Degenne

12 janvier 2022

Ces notes de cours contiennent uniquement un bref résumé de ce qui a été dit en cours. Reportez vous à vos notes pour plus de détails.

1 Cours 1: introduction à la classification

1.1 Introduction

Exemples de tâches d'apprentissage Données : photos. But : reconnaître les photos contenant des personnes.

Données : réservations de chambres d'un hôtel pour une certaine date future. But : trouver le bon prix auquel vendre la chambre.

Données : formules chimiques de molécules. But : savoir comment les molécules vont plier (pour prédire leur interaction avec d'autres molécules).

Le cours porte sur des tâches du type du premier exemple, dites tâches de classification.

Données et étiquettes On notera $\mathcal{X}$ un ensemble de données. Chaque donnée est décrite par un ensemble $\mathcal{A}$ d'attributs. Chaque attribut $a \in \mathcal{A}$ prend sa valeur dans un certain ensemble de valeurs $\mathcal{V}_a$. Ainsi, on peut considérer l'ensemble des données x dont les coordonnées balayent toutes les valeurs possibles des attributs : c'est l'espace des données que nous noterons $\mathcal{D}$. Si l'on note $a_1, \dots, a_P$ les P attributs, $D = V_{a_1} \times \dots \times V_{a_P}$. Toute donnée appartient à cet ensemble et on a $\mathcal{X} \subseteq \mathcal{D}$.

Il est souvent utile d'avoir une représentation géométrique de l'espace des données, où chaque attribut correspondant à un axe de coordonnées. C'est possible si tous les attributs ont des valeurs numériques. S'il y a P attributs, l'espace des données est un espace euclidien à P dimensions.

Types d'attributs: qualitatif, quantitatif. nominal/ordinal. valeur absolue ou non. discret ou continu.

Propriété voulue : le résultat de notre analyse ne dépend pas des unités des attributs.

Tâches: classification, régression Classification : chaque donnée $x \in \mathcal{D}$ a une classe (ou étiquette) dans un ensemble $\mathcal{Y}$. Le but est de trouver l'étiquette

de x à partir de ses attributs. Géométriquement : trouver une partition de l'ensemble $\mathcal{D}$ correspondant aux classes.

On peut par exemple chercher un hyperplan pour séparer l'espace en deux (c'est un modèle). Mais on ne peut souvent pas classer les données parfaitement de cette manière, à cause du bruit ou de la complexité des données. Il y aura des erreurs.

Régression : prédire l'étiquette de x, qui est une valeur continue (dans $\mathbb{R}$ par exemple). On dispose d'une notion de se tromper plus ou moins selon l'écart entre la prédiction et la réalité. Exemple: si $\mathcal{Y}$ est un intervalle de $\mathbb{R}$, on peut définir une perte $l(\hat{y}, y) = |\hat{y} - y|$. Plus la perte est grande, plus on se trompe en prédisant $\hat{y}$ plutôt que y.

Segmentation : grouper les points en fonction de leur ressemblance. Aucune étiquette n'est disponible.

Prédiction : l'ensemble de points est une suite temporelle et on veut prédire son évolution future. Prédire les prix en bourse par exemple.

Bruit et données manquantes Peu exploré dans ce cours mais très important: les données ne sont jamais "propres".

Certaines valeurs sont des erreurs.

Certaines valeurs sont manquantes.

Propriété voulue : obtenir un résultat fiable même si une proportion des données est mauvaise.

Préparation des données Les données brutes sont rarement dans une forme qui convient bien à la tâche ou à l'algorithme qu'on veut utiliser. On peut vouloir effectuer des opérations sur les données avant de les classer. Quelques exemples:

- diminution du nombre d'attributs
- projection dans un sous-espace
- création de nouveaux attributs (*features*)
- changement d'échelle
- interpolation des attributs manquants

1.2 Classification

On étudie dans ce cours la tâche dite de classification supervisée.

Definition 1. On dispose d'un ensemble $\mathcal{X}$ de N données étiquetées. Chaque donnée x_i est caractérisée par P attributs et par sa classe $y_i \in \mathcal{Y}$. Dans un problème de classification, la classe prend sa valeur parmi un ensemble fini ($\mathcal{Y}$ est fini). Le problème consiste alors, en s'appuyant sur l'ensemble d'exemples $X = \{(x_i, y_i) \mid i \in \{1, \ldots, N\}\}$, à prédire la classe de toute nouvelle donnée $x \in \mathcal{D}$.

Si $|\mathcal{Y}| = 2$, on parle de classification binaire.

Si une donnée peut appartenir à plusieurs classes, on parle de problème multi-classes (exemple: photo de chat, chien? Les deux si la photo contient

chat et chien). Dans ce cours, on se concentre sur le cas où chaque donnée appartient à une seule classe.

Definition 2. Un exemple est une donnée pour laquelle on dispose de la classe.

On parle d'apprentissage à partir d'exemples ou d'apprentissage supervisé. A l'opposé, pour une tâche comme la segmentation, pour laquelle on ne dispose que de données mais pas d'étiquettes, on parle d'apprentissage non supervisé.

Hypothèse générale: les exemples à partir desquels on travail sont représentatifs des données. S'en assurer est un problème amont de la classification, qui peut être difficile à résoudre.

Definition 3. Un classeur est un algorithme (une fonction) qui à partir d'un ensemble d'exemples construit une fonction associant une classe à chaque point de $\mathcal{D}$.

A partir d'exemples, un classeur produit une fonction plus générale: c'est ce qu'on appelle généraliser. Le but n'est pas d'obtenir un classeur qui soit efficace pour classer les données de $\mathcal{X}$, mais un classeur qui soit performant sur tout $\mathcal{D}$.

La capacité d'un algorithme à généraliser dépend à la fois de l'algorithme, mais aussi des exemples qui ont servi à sa construction. Quelques sources d'erreur:

- Pas assez d'exemples: si on a très peu d'exemples, ils ne contiendront pas assez d'information pour apprendre à classer correctement sur tout $\mathcal{D}$.
- Un classeur trop rigide: le classeur qui "apprend par coeur" (voir plus bas) ne pourra jamais généraliser.
- Des étiquettes erronées: si les étiquettes des exemples ne sont pas fiables, elles peuvent introduire de l'erreur dans le classeur.

Implicitement, un classeur détermine l'importance des différents attributs pour prédire la classe. Certains classeurs font cette opération explicitement: on dit qu'ils construisent un modèle. D'autres comme k-NN (voir plus bas) utilisent les données directement.

1.3 Exemples de classeurs

Sans généralisation : tableau *look-up* table. Apprentissage par coeur.

Ce classeur fonctionne de la manière suivante:

- construction: stocker tous les exemples dans un tableau.
- évaluation d'une nouvelle donnée: si la donnée est dans le tableau, retourner la classe (ou étiquette) correspondante. Sinon retourner une étiquette au hasard.

On a une erreur de 0% sur l'ensemble de données $\mathcal{X}$, mais une erreur de 50% sur $x \notin \mathcal{D}$. Aucune généralisation.

Plus proches voisins (k-NN)

Le principe du classeur k-NN est le suivant:

- Construction: stocker tous les exemples dans un tableau.
- Evaluation: étant donné une nouvelle donnée x,
  - Calculer la distance d_i de x à x_i pour tout i de 1 à N (nombre de données)
  - Trouver les k données x_i les plus proches de x
  - Retourner l'étiquette majoritaire parmi les k plus proches voisins

La dernière étape de l'évaluation peut être remplacée par le calcul d'un maximum pondéré, où chaque donnée a un poids qui décroit avec la distance à la nouvelle donnée x.

Ce classeur dépend d'une notion de distance, ou dissimilarité.

Comment choisir k ? Pas de méthode. Il faut essayer en fonction du problème.

Problèmes:

- temps de calcul important à l'évaluation si le nombre d'exemples est grand
- problème du choix de la distance
- problème de l'importance relative des attributs: si un attribut a de plus grandes variations qu'un autre, il a une influence plus grande sur la distance.