École Centrale de Lille

Science des données et Intelligence artificielle

Pierre Chainais

Semestre d’automne

Décision et apprentissage

Une image contenant texte, Police, Graphique, capture d’écran

Description générée automatiquement

# 1. Introduction

## 1.1. Objectifs de l’apprentissage

L’apprentissage comprend :

* La classification supervisée, semi-supervisée et non supervisée : la prédiction est discrète
* La régression : la prédiction est continue

## 1.2. Position du problème

Une image contenant texte, diagramme, Police, capture d’écran

Description générée automatiquement

Schéma d’un modèle

Les bonnes propriétés d’un modèle sont la capacité de généralisation et la possibilité de réutilisation du modèle dans un contexte différent.

## 1.3. Familles d’approches

Dans tous les cas l’objectif est la minimisation des erreurs de prédiction.

Cela peut être la **minimisation d’une fonction de coût** (risque) (par exemple, ). On parle **d’optimisation**.

Cela peut également être la **maximisation d’une probabilité** (par exemple, ). On parle dans ce cas de **raisonnement Bayésien**.

# 2. Modèles linéaires pour la régression

## 2.1. Position du problème

Les données sont de la forme et les cibles .

On peut faire l’hypothèse d’une dépendance linéaire en et en : .

On peut enrichir le modèle en supprimant la linéarité en : ( traduit le changement de représentation du problème).

## 2.2. Modèle linéaire en et en

Les données sont et le modèle de prédiction est où

On pose :

L’objectif est de minimiser l’erreur quadratique :

L’erreur quadratique est convexe et différentiable donc il existe une solution unique. On cherche un optimum en calculant un point on le gradient s’annule.

La matrice est symétrique **définie** positive, i.e. ses valeurs propres sont **strictement** positives si est de rang plein. Dans ce cas, on a bien la convexité.

Sous cette hypothèse, est solution unique de :

On note qu’on appelle le pseudo-inverse de Moore-Penrose de .

Si est inversible, alors , mais est rectangulaire donc cette hypothèse est rarement vérifiée.

Au lieu de résoudre , on cherche un minimum de . De plus, la résolution de n’est pas toujours évidente, donc on peut avoir recours à la descente de gradient. Cela a aussi l’avantage de ne pas avoir à inverser de matrice, et donc d’économiser du coût calculatoire.

Les prédictions sur l’ensemble d’entrainement sont : . On appelle la matrice « hat matrix ».

## 2.3. Modèle linéaire en et non en

On pose et les pour (pas forcément linéaires en ).

La fonction et permet d’obtenir la matrice de dimension .

Par exemple, si et , alors et .

Donc de manière similaire :

Par exemple, pour une régression polynomiale à l’ordre 3 d’un vecteur de dimension 2 :

Pour des sorties multiples, c’est-à-dire , avec des données sont et le modèle , on a :

On minimise le coût quadratique :

## 2.4. Moindres carrés régularisés et Ridge Regression

La dimension peut être grande, le nombre d’échantillons est fini donc limité, le modèle implique paramètres ( dans le cas où on reste dans la dimension initiale).

L’objectif est d’avoir un modèle précis donc assez riche (pas d’underfitting) mais qui ne surapprend pas (overfitting).

L’idée est de partir d’un modèle potentiellement trop riche et le régulariser, c’est-à-dire, sélectionner un sous-modèle.

### La méthode de Ridge Regression

On ne veut pas changer la moyenne mais étudier uniquement les variations autour d’elle. Donc on traite à part le terme d’ordre 0 qu’on fixe comme la moyenne des sorties : ; et on ne considère plus les vecteurs mais simplement).

On cherche qui minimise :

Le principe est de minimiser si et si de minimiser . On cherche donc un hyperparamètre intermédiaire appelé aussi **paramètre de régularisation**.

Le optimal correspond au modèle qui a la meilleure généralisation. Comme donne le même poids à tous les , cela nécessite de **normaliser** les données auparavant, c’est-à-dire de le rendre centrés-réduits :

On résout le problème pour sachant les données centrées-réduites :

On a :

Donc :

On cherche à résoudre par rapport à :

La solution est

Donc si et seulement si . Et est toujours bien définie car est toujours inversible. Numériquement, il arrive que soit mal conditionnée :

Plus généralement, on peut utiliser d’autres régularisations, par exemple en .