Polycopié : Apprentissage Supervisé

Une Vue Statistique

Florence d'Alché-Buc & Pavlo Mozharovskyi
Télécom Paris, Institut Polytechnique de Paris

Table des matières

1 Introduction

Ce polycopié présente une vue détaillée et complète de l'apprentissage supervisé dans une perspective statistique, tel que développé dans le cours TSIA-SD210. L'objectif est d'exposer les principaux concepts théoriques (définitions, propositions, lemmes et démonstrations) qui sous-tendent la modélisation statistique des problèmes d'apprentissage supervisé.

2 Définitions et Concepts Fondamentaux

2.1 Machine Learning et Apprentissage Supervisé

Définition 2.1 (Machine Learning)**.** *Le* machine learning *est un domaine de l'intelligence artificielle qui permet à un ordinateur d'apprendre à réaliser des tâches (reconnaissance, diagnostic, planification, etc.) à partir de données, sans être explicitement programmé pour chacune de ces tâches.*

Une autre définition, donnée par Tom Mitchell, est la suivante :

Définition 2.2 (Définition de Tom Mitchell)**.** *Un programme informatique apprend de l'expérience E concernant une classe de tâches T et une mesure de performance P, si sa performance sur les tâches de T, mesurée par P, s'améliore avec l'expérience E.*

Définition 2.3 (Apprentissage Supervisé)**.** *L'*apprentissage supervisé *consiste à estimer une fonction (un classificateur ou un régressseur) à partir d'un ensemble de données étiquetées*

$$S_n = \{(x_i, y_i),\, i = 1, \ldots, n\},$$

où chaque x_i représente les caractéristiques (features) et y_i la sortie (étiquette ou valeur cible).

2.2 Composantes d'un Problème d'Apprentissage

Les éléments essentiels d'un problème d'apprentissage supervisé sont :

- **Représentation des données :** choix des variables pertinentes (features) pour modéliser l'objet.
- **Espace d'hypothèses ($\mathcal{G}$) :** ensemble des fonctions possibles parmi lesquelles est recherchée la solution.
- **Fonction de perte (ℓ) :** mesure de l'erreur entre la prédiction et la valeur réelle.
- **Algorithme d'apprentissage :** procédure permettant de sélectionner la meilleure fonction dans $\mathcal{G}$ à partir des données.
- **Évaluation :** métriques (telles que l'exactitude, l'erreur quadratique) pour mesurer la performance du modèle.

3 Cadre Probabiliste et Statistique de l'Apprentissage Supervisé

3.1 Modélisation Statistique

Soient :

- $X \in \mathbb{R}^p$, un vecteur aléatoire représentant les caractéristiques d'un objet.
- Y, une variable aléatoire représentant la cible (classe ou valeur réelle).
- $p(x, y)$, la distribution conjointe de (X, Y).

Définition 3.1 (Fonction de Risque)**.** *Pour une fonction $g : \mathbb{R}^p \to \mathbb{R}$ et une fonction de perte ℓ, le risque (ou perte vraie) est défini par :*

$$R(g) = \mathbb{E}_{(X,Y)}[\ell(g(X), Y)].$$

3.2 Cas de la Classification Binaire

Pour une classification binaire, on utilise souvent la perte 0-1 :

$$\ell(g(x), y) = \mathbf{1}\{g(x) \neq y\},$$

ce qui donne :

$$R(g) = P(g(X) \neq Y).$$

4 Empirical Risk Minimization (ERM)

4.1 Principe de l'ERM

Comme la distribution $p(x, y)$ est inconnue, le risque $R(g)$ ne peut être calculé directement. On utilise alors le *risque empirique* :

$$R_n(g) = \frac{1}{n} \sum_{i=1}^{n} \ell(g(x_i), y_i).$$

Le but est de trouver :

$$\hat{g} \in \arg\min_{g \in \mathcal{G}} R_n(g).$$

4.2 Erreur d'Estimation et Erreur d'Approximation

On distingue deux sources d'erreur :
— **Erreur d'estimation :** différence entre $R(\hat{g})$ et $\inf_{g \in \mathcal{G}} R(g)$.
— **Erreur d'approximation :** différence entre $\inf_{g \in \mathcal{G}} R(g)$ et $R(g^*)$, où g^* représente la fonction cible idéale.

5 Fonctions de Perte

Différentes fonctions de perte sont utilisées selon le problème étudié :
— **Perte 0-1 (misclassification loss) :** $\ell(g(x), y) = \mathbf{1}\{g(x) \neq y\}$.
— **Perte exponentielle :** $\ell(g(x), y) = e^{-y\, g(x)}$.
— **Log-vraisemblance binaire :** $\ell(g(x), y) = -\log(1 + e^{-2y\, g(x)})$.
— **Erreur quadratique :** $\ell(g(x), y) = (1 - y\, g(x))^2$.

6 Le Classificateur de Bayes

6.1 Définition

Définition 6.1 (Classificateur de Bayes)**.** *Pour la classification binaire, le classificateur de Bayes est défini par :*

$$g_{Bayes}(x) = \mathbf{1}\{P(Y = 1|x) > 0.5\}.$$

Ce classificateur minimise le risque pour la perte 0-1.

6.2 Propriété Fondamentale

Proposition 6.1. *Le classificateur de Bayes réalise le risque minimal parmi tous les classificateurs, c'est-à-dire :*

$$R(g_{Bayes}) = \inf_{g} R(g).$$

Démonstration : La démonstration consiste à montrer que pour chaque x, la prédiction qui minimise l'erreur de classification est celle qui choisit la classe ayant la plus grande probabilité conditionnelle. (La démonstration complète est laissée en exercice.)

7 Généralisation et Bornes de Généralisation

7.1 Dimension VC et Shattering

Définition 7.1 (Shattering)**.** *Un ensemble de points $\{x_1, \ldots, x_n\}$ est dit* shattered *par un ensemble d'hypothèses $\mathcal{H}$ si, pour toute affectation binaire des points, il existe une fonction $h \in \mathcal{H}$ qui classe ces points sans erreur.*

Définition 7.2 (Dimension VC)**.** *La* dimension VC *de $\mathcal{H}$, notée d_{VC}, est la taille maximale d'un ensemble de points qui peut être shatteré par $\mathcal{H}$.*

7.2 Exemple : Hyperplans

Pour l'ensemble des hyperplans dans $\mathbb{R}^d$, il est démontré que :

$$d_{VC}(\mathcal{H}_d) = d + 1.$$

7.3 Borne de Généralisation

Théorème 7.1 (Borne de Généralisation de Vapnik-Chervonenkis)**.** *Pour tout classificateur $h \in \mathcal{H}$ et pour tout $\delta > 0$, avec une probabilité supérieure à $1 - \delta$, il existe :*

$$R(h) \leq R_n(h) + \sqrt{\frac{8\, d_{VC} \left(\ln \frac{2n}{d_{VC}} + 1 \right) + 8 \ln \frac{4}{\delta}}{n}}.$$

8 Régularisation et Minimisation du Risque Empirique Régularisé

8.1 Principe de Régularisation

Afin de limiter le sur-apprentissage (overfitting), on introduit une pénalité sur la complexité du modèle dans l'optimisation :

$$\hat{h} = \arg\min_{h \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) + \lambda\, \Omega(h) \right\},$$

où $\Omega(h)$ est une mesure de la complexité du modèle et λ est le paramètre de régularisation.

8.2 Interprétation

Cette approche permet de trouver un compromis entre l'adaptation aux données (minimisation du risque empirique) et la généralisation (contrôle de la complexité).

9 Autres Concepts et Applications en Machine Learning

9.1 Biais, Variance et Overfitting

Remarque 9.1. *Un modèle de faible complexité peut induire un fort biais (sous-apprentissage), tandis qu'un modèle trop complexe conduit à une forte variance (sur-apprentissage). Le compromis biais-variance est ainsi crucial pour une bonne généralisation.*

9.2 Support Vector Machines (SVM)

Les SVM sont un exemple de méthode de régularisation utilisant la *hinge loss* :

$$\text{Hinge loss} : \max(0,\, 1 - y\,(h(x) + b)).$$

Le classificateur final s'exprime alors par :

$$g(x) = \text{sign}(h(x) + b).$$

10 Conclusion

Ce polycopié offre une synthèse complète des concepts théoriques de l'apprentissage supervisé dans une approche statistique. Il couvre la modélisation probabiliste, l'optimisation par minimisation du risque empirique, les bornes de généralisation ainsi que les techniques de régularisation permettant de lutter contre le sur-apprentissage.

11 Bibliographie

— Vapnik (1998) : *Statistical Learning Theory*, John Wiley & Sons.
— Bishop (1999) : *Pattern Recognition and Neural Networks*, Springer.
— Hastie, Tibshirani, Friedman (2001) : *The Elements of Statistical Learning*, Springer.
— Haykin (2009) : *Neural Networks and Learning Machines*, Pearson.
— Abu-Mostafa, Magdon-Ismail, Lin (2012) : *Learning from Data : A Short Course.*
— James, Witten, Hastie, Tibshirani (2013) : *An Introduction to Statistical Learning*, Springer.
— Bertsekas (2016) : *Nonlinear Programming*, Athena Scientific.
— Goodfellow, Bengio, Courville (2016) : *Deep Learning*, MIT Press.
— Mohri, Rostamizadeh, Talwalkar (2018) : *Foundations of Machine Learning*, MIT Press.
— Bach (2024) : *Learning Theory from First Principles*, MIT Press.