12 leçons d'apprentissage automatique : 2è résumé du cours de modélisation des problèmes complexes DIC-9251, présenté le 26 avril 2025 au Professeur Hakim Lounis

Dominique S. Loyer
Université du Québec à Montréal (UQAM)
Département d'informatique
Montréal, QC, Canada
loyer.dominique@courrier.uqam.ca

Abstract

L'article de Pedro Domingos expose, à travers le prisme de 12 cas d'utilisation (Domingos :2012), de classification de l'apprentissage automatique. Cet apprentissage permet au système d'avoir la capacité d'effectuer des tâches importantes en généralisant à partir d'exemples lui étant fournis.

Mots clés : Apprentissage automatique, Apprentissage machine, ML

1 Apprentissage = représentation + évaluation + optimisation

1.1 La représentation

Le classifieur (modèle utilisé pour la classification) doit d'abord être représenté par un langage formel que la machine comprend. Il est ainsi impératif de choisir un ensemble de classifieurs (classifier) possibles. Cet ensemble est nommé l'espace hypothétique de l'apprenant «hypothesis space learner»(1).

1.2 L'évaluation

L'autre composante est l'évaluation qui est constitué par la fonction objectif afin de distinguer les bons modèles des mauvais. (1).

1.3 L'optimisation

Finalement, l'auteur rappelle le besoin de choisir une technique d'optimisation pour rechercher dans les classifieurs, les meilleurs résultats.(1).

Aussi, Domingos rappelle par les 12 leçons, les thèmes centraux du ML tels que la généralisation, l'importance des données (qui ne sont pas en elles-mêmes suivantes cependant), les connaissances préalables, le surapprentissage des données «overfiting» (le modèle ayant de la difficulté à généraliser à partir des exemples). Cependant, le plus gros problème du ML demeure selon l'auteur, le problème de la dimensionnalité (1).

2 La généralisation

L'objectif ultime ne réside pas dans la performance exemplaire sur les données d'entraînement (facilement réalisable par la mémorisation), mais dans la capacité de généralisation à de nouvelles données non observées, au-delà des données d'entraînement. Évaluer sur les données d'entraînement constitue une erreur fondamentale. La validation croisée est une méthode employée pour estimer cette capacité de généralisation.(1).

3 Les données ne sont pas suffisantes en elles-mêmes

La généralisation étant le but principal, les données ne suffisent pas peu importe combine nous en avons. (1). Aussi, l'auteur illustre par un exemple : 100 variables d'une fonction booléenne sur 1 000 0000 d'exemples donne un chiffre astronomique de $2^100 - 10^6$. Il rappelle, par ailleurs, que ce problème n'est pas nouveau puisqu'abordé déjà par le philosophe , David Humes, il y a déjà deux décennies. (1).

4 Le surapprentissage

4.1 Difficulté à généraliser

Le risque du surapprentissage étant que le système n'arrive pas à généraliser et peut, même «halluciner»(1) un classifieur ou une partie de ce dernier. L'auteur illustre par l'exemple d'une machine qui aurait un résultat de 100% sur les données d'entrainement mais que 50% sur les données de validation, alors qu'il aurait pu obtenir un résultat de 75% sur les deux classifieurs, il surentraine les données. (1).

4.2 Le compromis biais/variance

Il illustre aussi la matrice de «biais/variance» qui consiste à analyser les résultats sous 4 quadrants

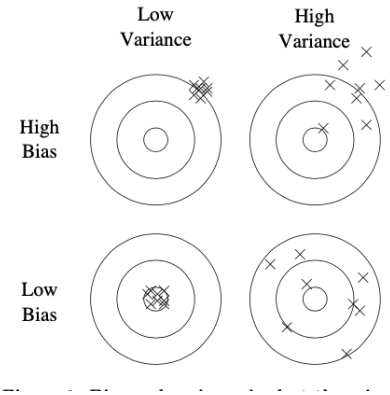

FIGURE 1: Illustration du compromis biais-variance en apprentissage automatique. Source : Domingos :2012

Le biais consiste à continuellement apprendre la même mauvaise chose alors que la variance est la tendance à apprendre de façon aléatoire qui ne tient pas compte du signal réel.(1).

4.3 La validation-croisée

L'auteur argumente aussi que la validation-croisée «cross-validation» peut aider à réduire le surapprentissage. (1).

4.4 La régularisation

Il y a aussi la régularisation qui peut aider à le combattre tout comme les tests statistiques comme le Chi carré avant d'introduire une nouvelle structure d'ailleurs.(1).

5 La courbe de dimensionnalité

Après le surapprentissage, l'auteur (1) mentionne qu'il s'agit du second plus grand défi en ML. Aussi. la "malédiction de la dimensionnalité" complique de manière exponentielle la généralisation à mesure que le nombre de variables augmente. Les concepts de similarité et de distance deviennent antinomiques, et les données occupent une portion extrêmement réduite de l'espace.

6 Les garanties théoriques

Les limites théoriques (par exemple, sur la quantité d'exemples requise) sont souvent très larges en pratique et servent davantage à orienter la conception des algorithmes qu'à informer des décisions concrètes. Les assurances asymptotiques (avec données infinies) ne reflètent pas toujours les performances observées avec des données finies.(1)

7 L'ingénierie des caractéristiques

La réussite d'un projet d'apprentissage automatique repose souvent davantage sur la qualité des variables exploitées que sur l'algorithme d'apprentissage en lui-même.(1) Élaborer de bonnes variables nécessite du temps, de la créativité ainsi qu'une compréhension approfondie du domaine.

8 Les données massives remportent sur le meilleur algorithme possible

Acquérir un volume accru de données est souvent la méthode la plus expéditive pour améliorer la performance, plutôt que de concevoir un algorithme d'une complexité accrue. Néanmoins, la scalabilité, en termes de temps de traitement, devient alors une problématique cruciale.(1)

9 Apprentissage sur plusieurs modèles différents

Les techniques d'ensemble (telles que le bagging, le boosting, le stacking) qui agrègent les prédictions de multiples modèles offrent quasi systématiquement des performances supérieures comparativement à un modèle unique, principalement par le biais d'une réduction de la variance.(1)

10 La simplicité n'implique pas la précision

Contrairement à une interprétation commune du principe du rasoir d'Occam, un modèle plus simple n'est pas nécessairement plus précis sur les données de validation.(1) Des modèles sophistiqués tels que les ensembles ou les machines à vecteurs de support peuvent très bien assurer la généralisation. La simplicité constitue une vertu par essence, mais ne garantit pas l'exactitude.

11 La représentation n'implique pas l'apprentissage

Le fait qu'une fonction puisse théoriquement être modélisée par une structure, telle qu'un arbre de décision, ne garantit pas qu'elle puisse réellement être apprise avec des ressources limitées en termes de données, de temps et de mémoire. Les contraintes algorithmiques et l'existence d'optima locaux influencent ce processus.(1)

12 La corrélation n'implique pas la causalité

Les modèles d'apprentissage automatique identifient des corrélations dans les données issues d'observations.(1) Il est nécessaire d'exercer une vigilance accrue avant de considérer ces corrélations comme des liens de cause à effet. Pour démontrer la causalité, il est préférable de recourir à des données issues de méthodes expérimentales, dans lesquelles les variables sont maîtrisées.

Conclusion

En somme, l'article met en lumière que la réussite en apprentissage automatique repose sur l'assimilation de ces enseignements pratiques, souvent obtenus par l'expérience, en parallèle des connaissances théoriques formelles.

Méthodologie d'utilisation d'un grand modèle de langue (LLM) : GEMINI de Google

Aucune nouvelle idée n'a été tiré de GEMINI. L'utilisation n'a été réalisée le 3 avril 2025 que pour synthétiser le document afin de m'assurer que je n'oubliais pas de détails importants.

Table des matières

Références

[1] P. Domingos. 2012. https ://doi.org/10.1145/2347736.2347755 A few useful things to know about machine learning. *Communications of the ACM*, 55(10) :78–87.

Bibliographie et médiagraphie supplémentaire

Notes du cours DIC-9251 *Modélisation des problèmes complexes*