Déjeuner Scientifique
La place de l'apprentissage automatique à l'heure des données massives : que faire en présence des collections partiellement étiquetées et/ou d'une connaissance pré-acquise ?

Groupe des Jeunes Statisticiens

Juin 2018

1 Introduction

Depuis les 49èmes JdS, le groupe des Jeunes Statisticiens propose des déjeuners scientifiques. L'objectif est de créer un espace d'échanges autour de sujets scientifiques d'actualité. Un débat est mené par deux orateurs reconnus dans le domaine thématique du déjeuner et un membre du groupe jeune modère les discussions. Ce moment privilégié est l'occasion d'un tour de table, d'une introduction du thème par les orateurs, d'une session de questions-réponses ouvertes et d'une pause café conviviale pour des discussions informelles. Cette année, nous avons organisé deux déjeuners scientifiques le 31 mai 2018, un orienté méthodes modernes pour l'enseignement et un orienté apprentissage automatique. Ce rapport ne fait référence qu'au deuxième.

2 Présentation du déjeuner du 31 mai 2018

Les invités sont Massih-Reza Amini (professeur au Laboratoire d'Informatique de Grenoble) et Florence d'Alché-Buc (professeure au Laboratoire Traitement et Communication de l'Information, Télécom ParisTech). Le modérateur est Erwan Scornet (maître de conférence au Centre de Mathématiques Appliquées, Ecole Polytechnique). Il s'agit de discuter des problématiques de l'apprentissage automatique (aussi communément appelé *machine learning*) telles que la présence de données partiellement étiquetées ou la prise en compte d'une connaissance pré-acquise. Etaient présents (par ordre alphabétique) : Faïcel Chamroukhi, Antonin Della Noce, Emilie Devijver, Charlotte Dion, Damien Drubay, Mélina Gallopin, José Gregorio Gomez Garcia, Efoevi Angelo Koudou, Charlotte Laclau, Adrien Saumard, Myriam Tami.

Chacun s'est présenté en introduction de la discussion : les personnes sont majoritairement des universitaires, plus ou moins proches de l'apprentissage machine, avec des profils plus informatique ou plus statistique, ou même à l'interface, mais tous très curieux du sujet.

3 Introduction aux problématiques par les invités

Plusieurs problématiques ont été mentionnées par Massih-Reza Amini et Florence d'Alché-Buc. Les deux problématiques principales sont les suivantes : les données peuvent présenter un petit nombre d'observations étiquetées, ou alors peuvent être complétées par des connaissances physiques. Ces connaissandes par les experts du domaine peuvent être converties en loi a priori des variables. Ces deux situations sont détaillées dans les sections suivantes.

3.1 En pratique, les données sont non étiquetées ou peu étiquetées

Définition du problème Dans la vie réelle, on observe beaucoup de données, mais elles sont brutes. On a besoin d'un avis de spécialiste pour les étiqueter. Par exemple, on peut penser aux données patients, où l'avis du médecin permet de savoir si le patient est sain ou malade.

Si on a accès à un jeu de données $(X_1, \ldots, X_n) \in (\mathbb{R}^p)^n$, on peut demander à un spécialiste de l'étiqueter via une variable $Y \in \{0, 1\}$ (ou $\{1, \ldots, K\}$ dans le cas multi-classes). En pratique, il est trop coûteux d'étiqueter tout l'échantillon, nous avons donc seulement accès à des données partiellement étiquetées (aussi communément appelé apprentissage semi-supervisé) : un ensemble étiqueté ou supervisé (de taille l) et un ensemble non étiqueté ou non supervisé (de taille u) avec $l + u = n$ et $l << u$.

Méthodes principales À ce jour, trois classes de méthodes existent pour résoudre les problèmes d'apprentissage semi-supervisé : les méthodes génératives via l'estimation des densités conditionnelles, les méthodes discriminantes via le pseudo-étiquetage et les méthodes graphiques via la propagation des étiquettes sur un graphe empirique.

Hypothèses relatives à ces méthodes On distingue 3 hypothèses classiques dans le cadre de l'apprentissage semi-supervisé :

- L'hypothèse de continuité stipule que si deux exemples sont proches dans une région à haute densité, alors leurs étiquettes de classes devraient être similaires.
- L'hypothèse de partition implique que si deux exemples sont dans le même groupe (au sens d'une partition externe, via un modèle de mélange par exemple), alors ils sont susceptibles d'appartenir à la même classe.
- Enfin, l'hypothèse de variété stipule que pour des problèmes de grande dimension, les exemples se trouvent se des espaces topologiques localement euclidiens (ou variétés géométriques de faible dimension).

Notons que suivant les modèles et les méthodes, les hypothèses faites sont différentes et plus ou moins difficiles à valider.

3.2 Connaissance a priori

Un contexte possible est celui de la maintenance prédictive. Par exemple, dans le cadre d'applications industrielles, il est possible d'avoir accès à des connaissances physiques des données. Cette connaissance vient alors compléter le jeu de données partiellement ou non étiqueté. Cette connaissance peut être traduite par exemple en loi a priori sur les variables. Il peut être pertinent d'injecter cette information dans le critère à minimiser : que doit-on privilégier entre les connaissances a priori ou les quelques étiquettes connues ?

3.3 Données massives

D'ici 2020, on s'attend à avoir 40 zetta bytes (soit 10^{21} bytes) de données non structurées. Que faire avec toutes ces données ?

Comment peut-on utiliser/adapter les méthodes précédemment citées sur des jeux de données si volumineux ?

3.4 Résultats théoriques

Quelles garanties théoriques peut-on attendre ? Quelles types d'hypothèses on peut demander / utiliser ? Comment obtenir des résultats théoriques sur l'apprentissage avec des données volumineuses ?

Ou plus simplement, pourquoi ces méthodes fonctionneraient ?

Retour sur les hypothèses. En classification, on fait l'hypothèse (en discrimination) qu'autour des quelques données étiquetées, on est dans le même cluster. L'objectif est alors d'accentuer les dissimilarités. En

régression, on fait l'hypothèse de continuité. Par exemple, il existe des algorithmes de maximisation de la marge (type SVM), ou alors sur l'entropie, mais on cherche toujours à renforcer la continuité de la fonction sous-jacente.

En semi-supervisé, l'hypothèse de clustering est importante : par exemple, voir l'article [4].

La question maintenant, ce serait de savoir si localement, c'est risqué d'utiliser le classifieur : c'est plutôt les chercheurs de la communauté statistique qui se demandent si on croit ou non en la prédiction, pas seulement via la moyenne mais pus globalement avec la distribution. On sait que dans certaines régions, le classifieur ne répondra pas bien localement. Les industriels et les praticiens sont souvent demandeurs de savoir si le classifieur appris sera bon localement (car en moyenne cela ne suffit pas toujours)?.

Autre exemple : comment mesurer le biais apporté par un nouveau modèle quand on pseudo-étiquette ?

En apprentissage semi-supervisé, beaucoup de challenges sont à dépasser (voir les dernières publications dans les conférences internationales, type NIPS).

4 Débat - questions

Le débat s'est alors ouvert.

4.1 Consistance de la méthode ?

Une fois qu'on a suffisamment étiqueté, est-ce qu'on peut jeter les observations qui ne sont pas étiquetées, comme le cadre est devenu asymptotique ? Pourrait-on alors obtenir des garanties théoriques (type bornes) ? Doit-on prendre en compte les données qu'on n'a pas étiquetées ?

4.2 Quantifier l'apport des données étiquetées ?

Comment quantifier l'apport des données étiquetées ? Par exemple, est-ce que des techniques non supervisées pourraient marcher mieux que les méthodes semi-supervisées ?

En fait, du semi-supervisé peut être vu comme du non supervisé contraint par quelques étiquettes, ou comme du supervisé avec d'autres observations non étiquetées.

4.3 Extension à un codage non discret des données ?

Le clustering, c'est un codage (drastique) des données. En allant plus loin, on peut avoir une représentation des données vectorielles, et donc repenser les problèmes d'apprentissage via le codage des sorties (voir par exemple [3]). Dans ce cas, on transforme les problèmes de classification multi-classes en régression, et on se focalise sur la réponse et comment la recoder.

4.4 Apprentissage par transfert [1]

Dans l'apprentissage par transfert, on utilise un jeu de données similaire à celui qu'on a pour apprendre une structure. Par exemple, on infère un réseau de régulation de gènes sur une population A, et on utilise ce réseau pour une population B.

4.5 Apprentissage actif / passif [2]

L'apprentissage actif, a contrario de l'apprentissage passif, prend en compte l'intervention d'un oracle (typiquement, l'être humain) qui étiquette quelques données bien sélectionnées pour améliorer le modèle.

[1] référence datacamp
[2] référence wikipédia

4.6 Zero/one/few-shot learning : d'autres paradigmes de la classification où il y a des challenges

En apprentissage, on a accès à un ensemble d'apprentissage sur lequel on apprend notre fonction de prédiction. On considère qu'on n'observe pas (zero-shot) ou très peu (one/few-shot) toutes les classes dans cet ensemble d'apprentissage.

C'est une hypothèse réaliste sur les grands jeux de données : on dit que les données suivent une loi de puissance sur les affectations des classes, car beaucoup d'exemples sont concentrés sur peu de classes, tandis beaucoup de classes sont représentées par peu d'observations. Comme on observe un échantillon fini, il y a des classes qui ne seront pas représentées.

Détecter les nouvelles classes parmi les observations non labellisées, ou parmi l'ensemble de test est un problème difficile. On peut par exemple jouer avec le codage des sorties (mentionné précédemment), mais quels résultats théoriques peut-on obtenir ?

4.7 Bruit d'étiquetage

Un autre soucis se pose si les données étiquetées que l'on observe (en apprentissage semi-supervisé ou même dans le cas d'apprentissage supervisé) sont mal étiquetées. On parle alors de bruit d'étiquetage. Existe-t-il des travaux dans ce cadre ? (nécessite des hypothèse sur le type de bruit). Quelques références : [1, 2]

5 Pour aller plus loin

On pourrait aussi parler d'apprentissage par renforcement [3].

6 Conclusion

Comme on a pu le voir pendant ce déjeuner, ce problème ouvre de nombreux questionnements et des pistes de travail, à la fois méthodologiques et théoriques. Les réflexes sont différents en apprentissage supervisé, en apprentissage non supervisé mais aussi en apprentissage semi-supervisé.

D'un point de vue purement organisationnel, quelques remarques sur le déjeuner à proprement parler. Le tour de table du début a permis à chacun de se présenter, et aux orateurs de savoir auprès de quel public ils allaient communiquer. Les deux orateurs ont joué le jeu, en donnant leur point de vue sur le thème.

References

[1] Raj S Chhikara and jim McKeon. Linear discriminant analysis with misallocation in training samples. *Journal of the American Statistical Association*, v79:p899(8), 1984-12-01. graph Boundary parameters.

[2] C. B. Chittineni. Learning with imperfectly labeled patterns. *Pattern Recognition*, 12:281–291, 1980.

[3] Moussab Djerrab, Alexandre Garcia, Maxime Sangnier, and Florence d'Alché Buc. Output fisher embedding regression. 05 2018.

[4] Yury Maximov, Massih-Reza Amini, and Zaïd Harchaoui. Rademacher complexity bounds for a penalized multi-class semi-supervised algorithm. *J. Artif. Intell. Res.*, 61:761–786, 2018.

[3] Cours Master MVA par Munos