Apprentissage multitâche en mélange gaussien: les bornes théoriques

Minh-Toan NGUYEN[1], Romain COUILLET[2]

[1]GIPSA-lab, Université Grenoble-Alpes

[2]LIG-lab, Université Grenoble-Alpes

Minh-Toan.Nguyen@gipsa-lab.grenoble-inp.fr
Romain.Couillet@univ-grenoble-alpes.fr

Résumé – Nous étudions un modèle de mélange gaussien d'apprentissage multitâche et calculons la performance optimale asymptotique de chaque tâche dans le régime des données de grande dimension. Dans le cas supervisé, nous dérivons un algorithme simple qui atteint la performance optimale. Notre analyse utilise la méthode des répliques issue de la physique statistique.

Abstract – We study a Gaussian mixture model of multitask learning and compute the asymptotic optimal performance of each task in the regime of high dimensional data. In the supervised case, we derive a simple algorithm that attains the optimal performance. Our analysis uses the replica method from statistical physics.

1 Introduction

L'apprentissage multitâche (AMT) est une approche qui résout plusieurs problèmes d'apprentissage simultanément. Comparé à l'apprentissage à tâche unique, AMT peut utiliser les données plus efficacement et mieux généraliser. La relation entre tâches joue un rôle important dans l'AMT. Les méthodes d'AMT sont généralement basées sur l'hypothèse que les tâches sont étroitement liées. Lorsqu'il est appliqué à des tâches non liées, AMT peut entraîner une perte de performances, un phénomène connu sous le nom de *transfert négatif* [1].

Dans [2], les auteurs ont étudié un modèle simple d'apprentissage semi-supervisé (SSL) et ont examiné le rôle des données étiquetées et non étiquetées via le risque asymptotique de Bayes. Nous étendons ce travail au cas de l'apprentissage multitâche, en mettant l'accent sur le rôle de la similarité des tâches. Nous travaillons sous un modèle de mélange gaussien, dans le cadre pratiquement pertinent où la dimension et la quantité de données sont à la fois grandes et comparables. Dans chaque tâche, les données sont partiellement étiquetées et proviennent de deux classes. Grâce à la simplicité du modèle, la similarité entre deux tâches données peut être quantifiée par un nombre réel dans $[0, 1]$. La performance d'un algorithme d'apprentissage est mesurée par la probabilité de son erreur de classification sur un nouveau point de données. Bien que les données soient aléatoires, dans la limite des très grandes dimensions, la performance optimale converge vers une valeur déterministe, le *risque bayésien asymptotique*, qui dépend de la similarité entre tâches ainsi que, pour chaque tâche : du rapport entre nombre d'échantillons et dimension (le taux de suréchantillonnage), de la proportion de données non étiquetées et de la séparation entre classes. Cette performance optimale peut donc être utilisée pour analyser le rôle des paramètres du modèle, notamment celui des similarités entre tâches, ainsi que pour fournir un parangonnage d'évaluation de méthodes AMT sur la base d'un mélange gaussien.

Le rôle des données étiquetées ou non dans le modèle nous conduit à la même conclusion que [2], pour tout niveau de similarité de tâche, à savoir que les données étiquetées ont un avantage décroissant à mesure qu'elles s'additionnent et que les données non étiquetées ne sont utiles que lorsque les classes sont suffisamment distinctes. Quant au rôle de la similarité des tâches, nous avons constaté que plus les tâches sont similaires, plus le gain de performance de l'apprentissage multitâche est élevé. En apprentissage non supervisé, la classification n'est possible que si le niveau de séparabilité entre deux classes dépasse un seuil fixe, un phénomène connu sous le nom de *transition de phase BBP* [3]. Ce phénomène se produit encore dans l'apprentissage multitâche lorsque toutes les tâches sont non supervisées: plus les tâches sont similaires, moins la séparation est nécessaire pour rendre une classification possible. Pour une tâche non supervisée, les données étiquetées d'une tâche connexe peuvent fournir des indications utiles et améliorer considérablement la classification d'une tâche cible (Figure 3).

Malgré la simplicité du modèle de mélange gaussien, les statistiques bayésiennes utilisant ce modèle comme loi a priori peuvent conduire à des méthodes importantes telles que le K-moyennes ou le partitionnement spectral [4]. Il est donc important de savoir s'il existe un algorithme efficace atteignant ces performances optimales, problème abordé par de nombreux travaux dans le cas d'une tâche unique sur des mélanges gaussiens. Notamment, [5] calcule les performances de plusieurs algorithmes sur des modèles de données simples, y compris sur un mélange gaussien, et montre que différentes

méthodes atteignent des performances optimales sur différents modèles de données. Pour l'apprentissage multiclasse non supervisé, [4] montre que l'algorithme AMP peut atteindre les performances optimales lorsque le nombre de classes est inférieur à un seuil ; au-dessus de ce seuil, il est conjecturé qu'aucun algorithme ne peut atteindre les performances optimales en temps polynomial. Pour l'apprentissage semisupervisé, [6] propose un algorithme dont les performances sont remarquablement proches de l'optimal. En particulier, dans le cas d'un mélange gaussien équilibré, en utilisant seulement 17% de données non étiquetées supplémentaires au pire des cas, la méthode peut dépasser les performances optimales sur le jeu de données d'origine. Dans le cas d'AMT, l'algorithme basé sur l'analyse en composantes principales de [7], avec l'idée de relaxer les contraintes discrètes sur les étiquettes, est identique à l'algorithme optimal dérivé dans cet article basé sur l'inférence bayésienne.

Il est important de noter qu'en pratique, plus de données peuvent nuire aux performances, comme dans le cas d'un transfert négatif ou dans l'apprentissage semisupervisé lorsque des données non étiquetées sont ajoutées. Cet effet dépend évidemment de l'algorithme utilisé. Dans ce travail, nous considérons le meilleur algorithme, donc ajouter plus de données ne peut pas faire de mal.

Notre résultat principal est obtenu à l'aide du calcul des répliques symétriques, une méthode non rigoureuse issue de la physique statistique qui donne des résultats corrects pour un large éventail de problèmes d'inférence [8].

2 Modèle et résultats principaux

Nous considérons T tâches d'apprentissage dans lesquelles la tâche t consiste en N_t échantillons de $\mathbb{R}^D$. Le i-ième échantillon de la tâche t est donné par

$$\boldsymbol{Y}_{ti} = V_{ti}\boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{ti} \tag{1}$$

où $\sigma_t > 0$, les variables aléatoires $\boldsymbol{V}, \boldsymbol{U}, \boldsymbol{Z}$ sont indépendantes, $\{V_{ti}\} \overset{iid}{\sim} \mathcal{U}(\{-1,1\})$, $Z_{ti} \overset{iid}{\sim} \mathcal{N}(0, I_D)$ et $\boldsymbol{U}_1, \dots, \boldsymbol{U}_T$ suivent la distribution uniforme sur

$$\left\{(\boldsymbol{U}_1, \dots, \boldsymbol{U}_T) \in \mathbb{R}^{D\times T} : \langle \boldsymbol{U}_t, \boldsymbol{U}_{t'}\rangle = C_{tt'}, 1 \le t, t' \le T\right\}$$

dans laquelle la matrice $\boldsymbol{C} = (C_{tt'})_{t,t'=1}^T$ est définie positive avec $C_{tt} = 1$ pour tout t. En d'autres termes, dans la tâche t, les échantillons sont issus de deux distributions gaussiennes isotropes centrées sur $\pm\boldsymbol{U}_t$, où $\boldsymbol{U}_t$ est un vecteur unitaire inconnu. Dans chaque tâche, les proportions de données de chaque classe sont égales. [1] Nous définissons la similarité entre la tâche t et t' par $|C_{tt'}|$ qui mesure à quel point les vecteurs $\boldsymbol{U}_t$ et $\boldsymbol{U}_{t'}$ s'alignent. La classe de $\boldsymbol{Y}_{ti}$ est indiquée par V_{ti}. Indépendamment de toute autre variable aléatoire, chaque échantillon de la tâche t est étiqueté avec probabilité η_t. On considère le régime $D \to \infty$ et $\lim_{D\to\infty} N/D = \alpha$, $\lim_{D\to\infty} N_t/D = \alpha_t$. Les paramètres α_t sont appelés *taux de suréchantillonnage*. Les paramètres $\mathrm{SNR}_t = 1/\sigma_t^2$ sont les *rapports signal sur bruit*. À mesure que SNR_t augmente, les deux classes de la tâche t se séparent et la classification est plus facile. Nous avons accès aux échantillons $\boldsymbol{Y}$ et aux étiquettes ainsi qu'aux paramètres des modèles $\sigma, \eta, \alpha, \boldsymbol{C}$.[2] Nous souhaitons étudier en quoi les paramètres du modèle affectent l'erreur de classification, en supposant que le meilleur algorithme est utilisé, ce qui nous revient à calculer le risque de Bayes asymptotique du modèle pour des paramètres donnés. Lorsque chaque tâche est entièrement étiquetée, nous décrivons un algorithme simple qui atteint les performances optimales.

Soient $\hat{\boldsymbol{U}}_t = \mathbb{E}[\boldsymbol{U}_t|\mathcal{D}]$, $\hat{\boldsymbol{V}}_t = \mathbb{E}[\boldsymbol{V}_t|\mathcal{D}]$ où $\mathcal{D}$ inclut échantillons et étiquettes. Si les limites

$$q^*_{vt} := \lim_{D\to\infty} \frac{1}{N_t}\|\hat{\boldsymbol{V}}_t\|^2 \tag{2a}$$

$$q^*_{ut} := \lim_{D\to\infty} \frac{1}{\sigma_t^2}\|\hat{\boldsymbol{U}}_t\|^2, \tag{2b}$$

existent, nous obtenons notre résultat principal:

Theorem 2.1. *Les limites q^*_{ut} (2a) et q^*_{vt} (2b) satisfont*

$$q^*_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{3a}$$

$$q^*_{vt} = \eta_t + (1-\eta_t)\psi(q^*_{ut}) \tag{3b}$$

où, pour $Z \sim \mathcal{N}(0,1)$,

$$\boldsymbol{M} = \{C_{tt'}/\sigma_t\sigma_{t'}\}_{t,t'=1}^T$$
$$\boldsymbol{D} = \mathrm{diag}\{\alpha_t q^*_{vt}\}_{t=1}^T$$
$$\psi(\gamma) = -1 + 2\partial_\gamma \mathbb{E}[\log\cosh(\sqrt{\gamma}Z + \gamma)].$$

De plus,

$$\lim_{D\to\infty} \langle \hat{\boldsymbol{U}}_t, \boldsymbol{U}_t\rangle = \sigma_t^2 q^*_{ut}$$
$$\lim_{D\to\infty} \frac{1}{N_t}\langle \hat{\boldsymbol{V}}_t, \boldsymbol{V}_t\rangle = q^*_{vt}.$$

Enfin, l'estimateur de la classe d'un nouvel échantillon $\boldsymbol{Y}_{new}$,

$$\hat{V}_{new} = \mathrm{sgn}(\langle \boldsymbol{Y}_{new}, \hat{\boldsymbol{U}}_t\rangle),$$

atteint le risque bayésien asymptotique donné par

$$\lim_{D\to\infty} \mathbb{P}\left(\hat{V}_{new} \neq V_{new}\right) = 1 - \Phi\left(\sqrt{q^*_{ut}}\right),$$

où $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_t^\infty e^{-x^2}dx$.

Nous pouvons vérifier l'accord entre le théorème 2.1 et notre intuition en le testant par rapport aux cas particuliers suivants: si la similarité entre deux tâches quelconques est nulle, alors l'apprentissage multitâche n'apporte aucun gain de performance par rapport à l'apprentissage à tâche unique, tandis

[1] Sans perdre aucune idée importante, nous considérons ce cas spécifique de mélanges équilibrés pour sa simplicité. Notre résultat peut être facilement généralisé au cas où la proportion de données de classe 1 dans la tâche t est égale à $\rho_t \in [0,1]$.

[2] Dans la limite $D \to \infty$, si les fractions de données étiquetées de chaque tâche sont positives, les entrées de $\boldsymbol{C}$ peuvent être estimées avec une erreur d'ordre $O(D^{-1/2})$.

que si $\sigma_t = \sigma$ et $\boldsymbol{U}_t = \boldsymbol{U}$ pour tout t, c'est-à-dire que les tâches ont la même distribution de données, alors les performances optimales de toutes les tâches sont identiques et égales à celle d'un tâche unique avec $\alpha = \sum_t \alpha_t$ et $\alpha\eta = \sum_t \alpha_t \eta_t$.

Nous pouvons aussi explorer les conséquences du théorème 2.1 lorsque les tâches sont supervisées, non supervisées ou semisupervisées.

Apprentissage supervisé. Dans le cas supervisé, l'algorithme suivant atteint les performances optimales:

1. Pour $t = 1, \dots, T$

$$\overline{\boldsymbol{Y}}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} V_{ti} \boldsymbol{Y}_{ti};$$

2. Pour $t = 1, \dots, T$

$$\tilde{\boldsymbol{Y}}_t = \sum_{s=1}^{T} a_{ts} \overline{\boldsymbol{Y}}_s$$

où

$$\boldsymbol{A} = (a_{ts})_{t,s=1}^{T} = \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}}(\boldsymbol{I} + \boldsymbol{M}\boldsymbol{D}_{\boldsymbol{\alpha}})^{-1}$$
$$\boldsymbol{D}_{\boldsymbol{\alpha}} = \mathrm{diag}\{\alpha_t\}_{t=1}^{T}$$

3. Si $\boldsymbol{Y}_{new}$ est un nouvel échantillon dans la tâche t, alors

$$\hat{V}_{new} = \mathrm{sgn}(\langle \boldsymbol{Y}_{new}, \tilde{\boldsymbol{Y}}_t \rangle).$$

En d'autres termes, nous classons un nouvel échantillon de la tâche t selon sa projection sur le "vecteur de caractéristiques" $\tilde{\boldsymbol{Y}}_t$. Si les tâches sont apprises séparément, les vecteurs de caractéristiques sont simplement $\overline{\boldsymbol{Y}}_t$ [2], alors que si les tâches sont apprises ensemble, nous prenons en compte les interactions entre les tâches encodées dans la matrice $\boldsymbol{A}$ et utilisons $\tilde{\boldsymbol{Y}}_t$ au lieu de $\overline{\boldsymbol{Y}}_t$ pour la classification.

Apprentissage non supervisé. La figure 1 se concentre ensuite sur l'apprentissage non supervisé, dans le cadre de deux tâches. Dans la tâche t, le nombre de données N_t est égal au nombre de dimensions D. Il est connu pour le cas d'une tâche qu'une transition de phase se produit alors à $\mathrm{SNR}_c = 1/\sigma_c^2 = 1$: lorsque $\mathrm{SNR} < \mathrm{SNR}_c$, il est (asymptotiquement) impossible d'obtenir une performance non triviale (l'estimateur trivial attribue la même classe à tous les échantillons et résulte en 50% d'erreur de classification asymptotique). Dans le cadre de deux tâches corrélées avec paramètre de corrélation c, le phénomène de transition de phase est toujours présent mais maintenant à $\mathrm{SNR}_c = (1 + c^2)^{-1/2}$. La figure 1 représente les régions du plan (c, SNR) pour lesquelles la classification est possible ou non. Comme anticipé, des corrélations plus élevées diminuent le seuil de transition de phase. La vitesse de croissance de la région de classification possible augmente avec des corrélations plus grandes: des données supplémentaires faiblement corrélées améliorent marginalement la performance.

Apprentissage semi-supervisé. La figure 2 illustre le théorème 2.1 dans le cadre de deux tâches. La première tâche est composée d'un petit ensemble de données ($\alpha_1 = 0,1$) sans étiquettes ($\eta_1 = 0$), tandis que la deuxième tâche consiste en un ensemble de données entièrement étiqueté ($\eta_2 = 1$) avec deux fois plus de données ($\alpha_2 = 0.2$). On prend $C = \left[\begin{smallmatrix} 1 & c \\ c & 1 \end{smallmatrix}\right]$. Lorsque les deux tâches sont apprises ensemble et que l'estimateur optimal est utilisé, la première tâche bénéficie grandement de la seconde tâche : son risque bayésien diminue considérablement lorsque la similarité de la tâche c augmente de zéro à un.

3 Conclusion

Cet article étudie un modèle simple d'apprentissage multitâche, dans lequel chaque tâche est un problème de classification semisupervisé de données de mélange gaussien en grande dimension. Une expression analytique du gain de performance est obtenue lorsque les tâches sont apprises ensemble par rapport au cas où elles sont apprises séparément, en supposant que les meilleurs algorithmes (non nécessairement polynomiaux) sont utilisés. L'effet de la similarité des tâches sur le seuil de transition de phase lorsque chaque tâche est non supervisée a également été étudié. Malgré la simplicité du modèle de mélange gaussien, les statistiques bayésiennes utilisant ce modèle comme loi a priori conduisent à des conclusions pratiques importantes en fournissant un parangonnage absolu de performances optimales accessibles sous ce modèle. En particulier, nous avons montré que lorsque chaque tâche est entièrement étiquetée, un algorithme optimal simple existe (qu'il n'est donc pas nécessaire de chercher à améliorer). Pour le cas général, nous nous attendons à ce que les performances optimales puissent être atteintes par un algorithme du type "échanges de messages approximatifs" (AMP).

Ce travail entre ainsi dans une lignée de nouvelles considérations en apprentissage machine permettant de rompre avec la "malédiction" du fonctionnement en boite noire de nombreux algorithmes qui voient aujourd'hui le jour dans le domaine, et de réinstaurer des éléments indispensables de théorie de l'information (bornes de performance, maîtrise des algorithmes, limites éthiques et biais, etc.) au cœur du développement des outils de l'intelligence artificielle.

References

[1] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv:1706.05098*, 2017.

[2] Marc Lelarge and Leo Miolane. Asymptotic bayes risk for gaussian mixture in a semi-supervised setting. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2019.

[3] Jinho Baik, Gérard Ben Arous, and Sandrine Péché. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 2005.

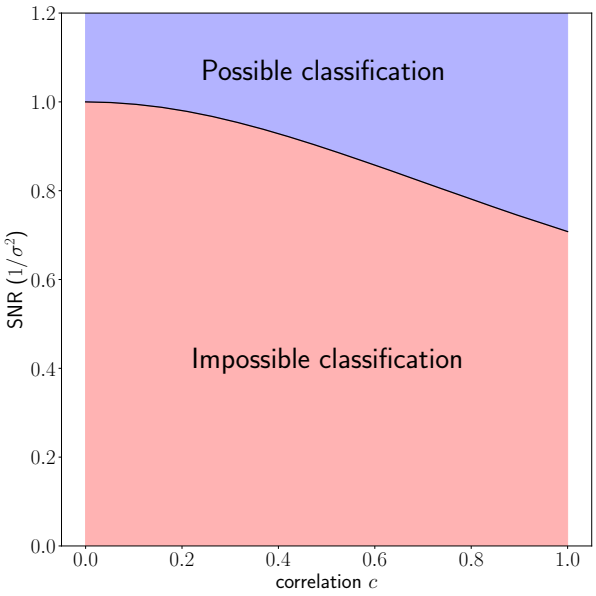

Figure 1: Transition de phase dans l'apprentissage non supervisé à deux tâches : $\eta_1 = \eta_2 = 0$, avec des taux de suréchantillonnage $\alpha_1 = \alpha_2 = 1$. Pour chaque tâche, la proportion de données de chaque classe est égale à $1/2$. ***Plus les tâches sont similaires, moins la séparation est nécessaire pour rendre une classification possible***.

[4] Lesieur et al. Phase transitions and optimal algorithms in high-dimensional gaussian mixture clustering. In *2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*. IEEE, 2016.

[5] Christos Thrampoulidis, Samet Oymak, and Mahdi Soltanolkotabi. Theoretical insights into multiclass classification: A high-dimensional asymptotic view. *arXiv:2011.07729*, 2020.

[6] Xiaoyi Mai and Romain Couillet. Consistent semi-supervised graph regularization for high dimensional data. *Journal of Machine Learning Research*, 22(94):1–48, 2021.

[7] Malik Tiomoko, Romain Couillet, and Frédéric Pascal. Pca-based multi task learning: a random matrix approach. In *25th International Conference on Artificial Intelligence and Statistics*, 2021.

[8] Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

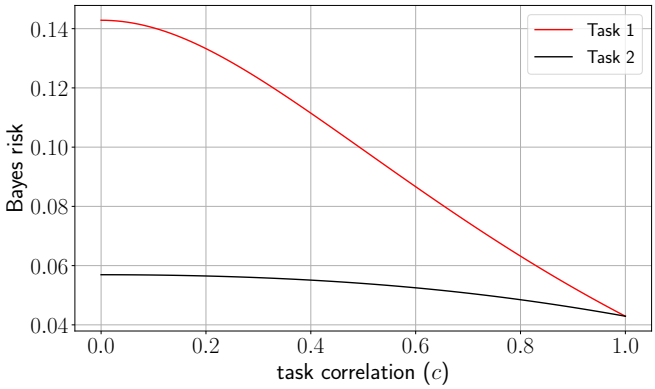

Figure 2: Risques de Bayes en fonction de la corrélation c entre 2 tâches, pour des proportions de données étiquetées $\eta_1 = 0$, $\eta_2 = 1$, taux de suréchantillonnage $\alpha_1 = 0,1$, $\alpha_2 = 0,2$ et niveau de bruit $= 0,2$. Pour chaque tâche, la proportion de données de chaque classe est $1/2$. ***Un gain significatif de classification peut être obtenu lorsque les deux tâches sont liées.***

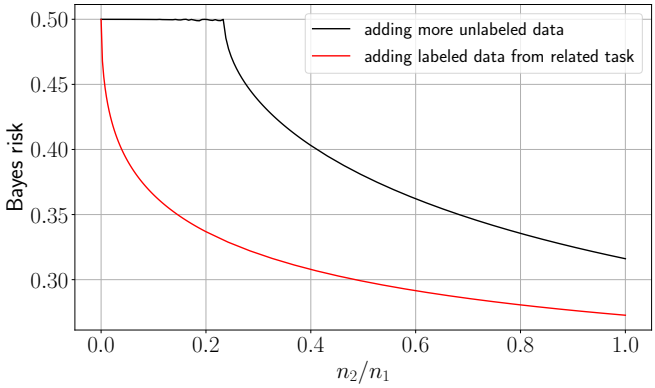

Figure 3: Risque bayésien asymptotique de la tâche 1 (tâche cible) non supervisée avec les paramètres $\alpha_1 = 1, \text{SNR}_1 = 0.9$. En régime de faible rapport signal sur bruit, il est impossible d'obtenir une performance non triviale pour la tâche cible. ***Si plus de données non étiquetées sont recueillies, une quantité de données comparable à n_1 est nécessaire avant que le risque bayésien ne commence à diminuer. En revanche, l'ajout de données étiquetées à partir d'une tâche connexe (avec corrélation $c = 0.8$), peut immédiatement et fortement réduire l'erreur de classification***.