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

Dans [1], un modèle simple de classification semisupervisée est introduit, dans lequel on considère N échantillons $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N \in \mathbb{R}^D$ tels que

$$\boldsymbol{Y}_i = V_i \boldsymbol{U} + \sigma \boldsymbol{Z}_i, \quad i = 1, \ldots, N \tag{1}$$

où $\sigma > 0$ et

$$\begin{aligned} \boldsymbol{U} &\sim \mathcal{U}(S^{D-1}), \\ V_i &\sim \mathcal{U}(\{-1, 1\}), \\ Z_i &\sim \mathcal{N}(0, I_D), \quad i = 1, \ldots, N \end{aligned}$$

sont indépendants. En d'autres termes, chaque échantillon a une probabilité $1/2$ d'appartenir à chacun des deux classes de données qui suivent les lois gaussiennes de moyennes $\pm\boldsymbol{U}$ et de la covariance $\sigma^2 I_D$. On dit qu'un échantillon $\boldsymbol{Y}_i$ est étiqueté si V_i est connu. Dans le cas semisupervisé, $\lfloor \eta N \rfloor$ échantillons choisis au hasard sont étiquetés, où $\eta \in (0, 1)$. Le cas $\eta = 1$ et $\eta = 0$ correspondent à l'apprentissage supervisé et non supervisé. Le modèle est étudié en régime de grande dimension où la dimension et la quantité de données tendent vers l'infini à un taux fixe $\alpha = \lim_{D\to\infty} N/D$, appelé *taux de suréchantillonnage*. Le paramètre SNR $= 1/\sigma^2$ est appelé le *rapport signal sur bruit*. À mesure que SNR augmente, les deux classes de la tâche t se séparent et la classification est plus facile. Les auteurs étudient le rôle des données étiquetées et non étiquetées en calculant l'erreur de classification minimale asymptotique, c'est-à-dire le *risque de Bayes* en fonction des paramètres du modèle α, σ, η, supposés connus.

Dans ce travail, nous étendons ce modèle au cas de tâches multiples. Considérons T tâches, dans lesquelles la tâche t consiste en N_t échantillons $\boldsymbol{Y}_{t1}, \ldots, \boldsymbol{Y}_{tN_t} \in \mathbb{R}^D$ tels que

$$\boldsymbol{Y}_{ti} = V_{ti} \boldsymbol{U}_t + \sigma_t \boldsymbol{Z}_{ti} \tag{2}$$

où $\sigma_t > 0$,

$$\begin{aligned} V_{ti} &\sim \mathcal{U}(\{-1, 1\}), \\ Z_{ti} &\sim \mathcal{N}(0, I_D), \end{aligned}$$

et les moyennes $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_T$ suivent la loi uniforme sur

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

dans laquelle la matrice $\boldsymbol{C} = (C_{tt'})_{t,t'=1}^T$ est définie positive avec $C_{tt} = 1$ pour tout t. La quantité $|C_{tt'}| \in [0, 1]$ mesure la similarité entre les tâches t et t'. Dans la tâche t, la proportion de données étiquetées est η_t. On considère le régime où $\lim_{D\to\infty} N_t/D = \alpha_t$. Nous avons accès aux échantillons $\boldsymbol{Y}$ et aux étiquettes ainsi qu'aux paramètres des modèles $\sigma, \eta, \alpha, \boldsymbol{C}$.[1] Nous souhaitons étudier en quoi les paramètres du modèle, en particulier la similarité entre les tâches, affectent l'erreur de classification, en supposant que le meilleur algorithme est utilisé, ce qui nous revient à calculer le risque de Bayes du modèle pour des paramètres donnés. Lorsque chaque tâche est entièrement étiquetée, nous décrivons un algorithme simple qui atteint les performances optimales.

Notation. $S^{D-1} = \left\{x \in \mathbb{R}^D, \|x\|_2 = 1\right\}$, $\langle ., . \rangle$ désigne le produit scalaire sur $\mathbb{R}^D$: $\langle x, y \rangle = \sum_i x_i y_i$.

Related works. Il est important de savoir s'il existe un algorithme efficace atteignant la performance optimale, problème abordé par de nombreux travaux dans le cas d'une tâche unique sur des mélanges gaussiens. Pour l'apprentissage multiclasse non supervisé, [2] montre que l'algorithme AMP peut atteindre les performances optimales lorsque le nombre de classes est inférieur à un seuil. Pour l'apprentissage semisupervisé,

[1]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 $O(D^{-1/2})$.

[3] propose un algorithme dont les performances sont remarquablement proches de l'optimal. Dans le cas d'AMT, l'algorithme basé sur l'analyse en composantes principales de [4], 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.

2 Résultats principaux

Soient $\hat{\boldsymbol{U}}_t = \mathbb{E}[\boldsymbol{U}_t|\mathcal{D}]$, $\hat{\boldsymbol{V}}_t = \mathbb{E}[\boldsymbol{V}_t|\mathcal{D}]$ où $\mathcal{D}$ comprend toutes les informations disponibles, c'est-à-dire les échantillons $\boldsymbol{Y}$, les étiquettes de données ainsi que les paramètres du modèle $(\alpha_t), (\sigma_t), (\eta_t)$ et $\boldsymbol{C}$.

Résultat. *Les limites*

$$q^*_{vt} := \lim_{D\to\infty} \frac{1}{N_t}\|\hat{\boldsymbol{V}}_t\|^2 \tag{3a}$$

$$q^*_{ut} := \lim_{D\to\infty} \frac{1}{\sigma_t^2}\|\hat{\boldsymbol{U}}_t\|^2, \tag{3b}$$

existent et satisfont

$$q^*_{ut} = [\boldsymbol{M} - \boldsymbol{M}(\boldsymbol{I} + \boldsymbol{D}\boldsymbol{M})^{-1}]_{tt} \tag{4a}$$

$$q^*_{vt} = \eta_t + (1-\eta_t)\psi(q^*_{ut}) \tag{4b}$$

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

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

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

Nous pouvons vérifier l'accord entre ce résultat 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 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 notre résultat lorsque les tâches sont supervisées, non supervisées ou semisupervisées.

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

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

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

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

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

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 $\overline{\boldsymbol{Y}}_t$ [1], 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 semisupervisé. La figure 2 illustre notre résultat 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.

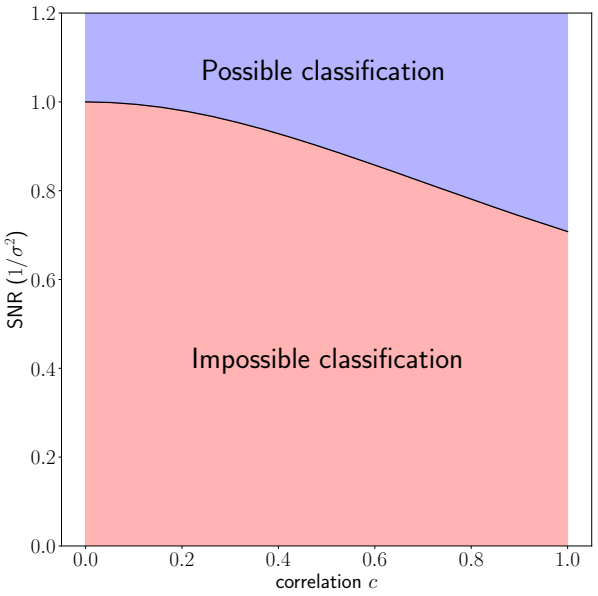

Figure 1: Transition de phase dans l'apprentissage non supervisé à deux tâches : $\eta_1 = \eta_2 = 0$, avec des taux de suréchantillonnage $\alpha_1 = \alpha_2 = 1$. Pour chaque tâche, la proportion de données de chaque classe est égale à $1/2$. ***Plus les tâches sont similaires, moins la séparation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

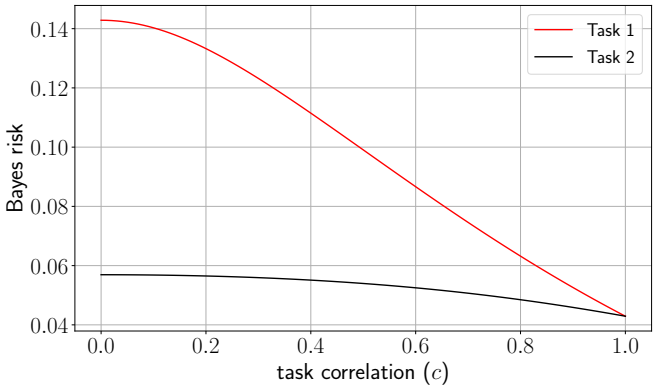

Figure 2: Risques de Bayes en fonction de la corrélation c entre 2 tâches, pour des proportions de données étiquetées $\eta_1 = 0$, $\eta_2 = 1$, taux de suréchantillonnage $\alpha_1 = 0,1$, $\alpha_2 = 0,2$ et niveau de bruit $= 0,2$. Pour chaque tâche, la proportion de données de chaque classe est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

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 des 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).

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.

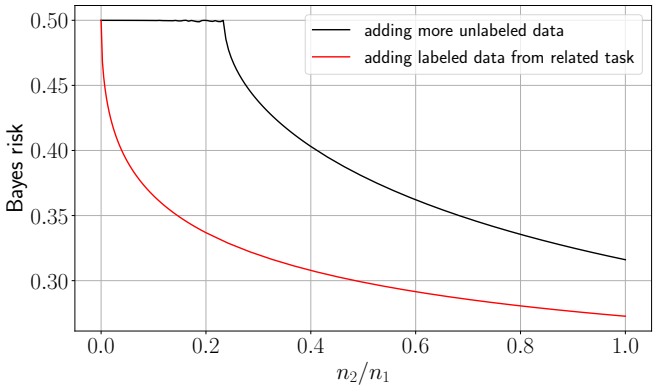

Figure 3: Risque bayésien asymptotique de la tâche 1 (tâche cible) non supervisée avec les paramètres $\alpha_1 = 1, \text{SNR}_1 = 0.9$. En régime de faible rapport signal sur bruit, il est impossible d'obtenir une performance non triviale pour la tâche cible. ***Si plus de données non étiquetées sont 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***.

References

[1] 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.

[2] 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.

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

[4] 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.