MODEL SELECTION FOR CONTEXTUAL BANDITS

Julien Aubert [1] & Luc Lehéricy [2] & Patricia Reynaud-Bouret [3]

[1] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - jaubert@unice.fr*
[2] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - luc.lehericy@univ-cotedazur.fr*
[3] *Laboratoire J.A. Dieudonné, Université Côte d'Azur, France - patricia.reynaud-bouret@univ-cotedazur.fr*

Résumé. Ce travail aborde le problème de la sélection de modèles dans les algorithmes de bandits contextuels lorsqu'ils sont utilisés pour modéliser une tâche d'apprentissage. Plus précisément, chaque modèle représente une partition de l'espace des contextes sur chaque ensemble de laquelle un algorithme de bandit est appliqué. Notre objectif est de trouver le modèle qui correspond le mieux aux données d'apprentissage. En étendant les outils traditionnels d'estimation et de sélection de modèle aux données non i.i.d et non stationnaires, nous montrons dans un premier temps qu'une procédure de hold-out sur les données satisfait un taux de convergence classique. Ensuite, sous diverses hypothèses, nous formulons des inégalités oracles avec différents taux de convergence. Nous fournissons également des exemples pour lesquels les hypothèses sont satisfaites. Enfin, nous testons nos résultats sur des données d'apprentissage synthétiques et réelles.

Mots-clés. Estimation statistique, Sélection de modèle, Bandits contextuels, Cognition, Vraisemblance pénalisée, Hold-out.

Abstract. This work tackles the problem of model selection in contextual bandit algorithms when used to model a learning task. More specifically, each model represents a partition of the contexts space on each set of which a bandit algorithm is applied. Our goal is to find the model that best fits the learning data. By extending the traditional tools of estimation and model selection to non i.i.d and non stationary data, we show first that a hold out procedure on the data satisfies a classical rate of convergence. Then, under various assumptions, we formulate oracle inequalities with various rates of convergence. We also provide examples for which the assumptions are satisfied. Finally, we test our results on both synthetic and real learning data.

Keywords. Satistical estimation, Model selection, Contextual bandits, Cognition, Penalized likelihood, Hold-out.

1 Introduction

1.1 Cognitive Models

Imagine an individual learning a categorisation task. Without initially knowing the rule, the individual is presented with objects in a sequential way, each object having a certain amount of features, and has to classify the objects in two categories. The reward he gets is whether the object was classified in the right category. His goal is to identify what the rule for belonging to one category or the other is. Our goal is to estimate its learning strategy, meaning to understand which features of the objects the learner saw as important for belonging to each of the categories.

A model describing the previous experiment is called a categorization model. It belongs to the wider category of cognitive models. They help understanding the mechanisms that occur in the brain while learning, remembering, and predicting tasks. They have been widely studied in the cognition literature [1] and have a major impact on education for instance. In this article, we consider cognitive models that describe learning experiments. Back to our example, the choice made by the learner to classify an object in a category depends on the choices he made before.

To assess the relevance of a model, one should first evaluate its goodness of fit through maximum likelihood estimation (MLE) or least-square regression. Then, one should compare the given model to other models and choose the best model given a certain criterion. This empirical methodology is traditionally used for cognitive modelling [2]. However, no theoretical result exists to prove the relevance of either the model or the methodology. Our goal is to establish a theoretical framework for model selection in learning.

1.2 Contextual Bandits

To do so, we use machine learning (ML) algorithms. We try to infer the learning strategy of a ML model. More specifically, given learning data, our goal is to select the learning algorithm that best fits the learning data. The algorithms we work with are contextual bandits algorithms (see 1). For $d \in \mathbb{N}$, let $\mathcal{P}_d$ be the probability simplex $\{q \in [0,1]^{d+1}, \|q\|_1 = 1\}$. Let $K \geqslant 2$ be the number of actions, and let $\mathcal{X}$ be a set representing the context space. We are interested in estimating the parameters of adversarial contextual bandits algorithms which satisfy, by definition, the following interaction protocol with the environment

Algorithm 1 Interaction protocol for contextual bandits [3]

Adversary secretly chooses a sequence of losses (or rewards) $(\pi_t)_{t=1}^n$ with $\pi_t \in [0,1]^K$.
Adversary secretly chooses a sequence of contexts $(x_t)_{t=1}^n$ with $x_t \in \mathcal{X}$.
for $t = 1, 2, \ldots, n$ **do**
    Learner observes context $x_t \in \mathcal{X}$
    Learner selects a distribution $p_t \in \mathcal{P}_{K-1}$ and samples A_t from p_t.
    Learner observe loss $\pi_{A_t,t}$.

In this setting a policy is a function mapping history sequences to distributions over actions. The regret measures the performance of a policy by comparing the rewards collected by the learner and the rewards collected by the best policy. Many algorithms are able to cope with problem 1. Contextual bandits refer to the class of bandit problems in which the learner has access to additional information (context) at each time step. There are many applications of such algorithms such as ad recommendation, patient follow-up in healthcare, etc... (see [4] for more details)

When the context space $\mathcal{X}$ is small enough, a simple idea to solve the contextual bandit problem is to apply one bandit algorithm per context. By doing so, one can show that the regret is lower bounded by $2\sqrt{nK|\mathcal{X}|\log(K)}$. If the context space is too large or if the amount of data is too small, then applying one bandit per context will always be insufficient. If instead of considering the set of all policies from $\mathcal{X}$ to $[K]$, one considers the set of all policies that are constant on each part of a given partition $\mathcal{P}$ of $\mathcal{X}$, then one can apply a Bandit algorithm on each part of the partition and therefore the factor $|\mathcal{X}|$ becomes $|\mathcal{P}|$ in the regret. This is the framework we will be using.

1.3 Link between Contextual Bandits and Cognitive Models

If we go back to our categorization example, at each time step, the learner has access to the features of each object. We make the assumption that he is learning by rule [1]: an object is in one category or the other because it obeys some criteria. The learner is thus partitioning the set of objects into parts each representing a different rule. On each part, we assume that the learner is applying a bandit algorithm. We therefore are in the setting of contextual bandits, only considering the set of policies that are constant on each part of a partition of the set of contexts $\mathcal{X}$. For the categorization task, the contexts are the objects the learner is seeing at each time step.

Our goal is to estimate the way the learner partitions the set of objects he has to classify. In other words, given that he is learning by making a partition on the set of contexts and applying a bandit algorithm in each set of the partition, can we estimate the partition he actually used to learn ? Each model we study is thus a partition of the context space on each set of which a bandit algorithm is applied.

From a statistical point of view, the underlying problem is similar to selecting the best model (i.e. closest to reality) when constructing a histogram [5]. Unlike the traditional framework in which the data are i.i.d with a common density to estimate, here the learning data are not stationary and not independent. During a learning task, present choices should be affected by past choices and rewards. Thus, traditional results and tools [5] such as cross validation [6] would not apply here. In a previous work, [7] focus on estimating the parameters of Exp3 when fitting it to learning data. In this work, we tackle the problem of choosing the best model to fit the learning data.

We extend the work of [5] for model selection to non i.i.d and non stationary data. To do so, we use classical concentration inequalities for martingales [8].

1.4 Contributions

- We show that a hold out procedure in this framework satisfies a classical rate of convergence, regardless of how the initial parameters are estimated.
- Let $\mathcal{M}$ be a collection of models. Let $\hat{\theta}^m = \arg\max_{\theta^m} \ell_T(\theta^m)$ be the estimated parameter for model $m \in \mathcal{M}$, where $\ell_T(\theta^m)$ is the log-likelihood function stopped at time T. Let $\text{pen} : \mathcal{M} \to \mathbb{R}_+$ be a penalty function. Let $\hat{m}$ be the model that minimize the penalized criterion:

$$\widehat{m} := \arg\min_{m \in \mathcal{M}} \left(-\frac{\ell_T(\hat{\theta}^m)}{T} + \text{pen}(m) \right)$$

  Then under some assumptions, for a well chosen T, we show that if the penalty function scales as $1/T$, then the estimated distribution satisfies an oracle inequality with a $O(1/T)$ additive factor.
- We provide examples with Stochastic Gradient Bandits [9] and EXP3-IX [3] for which the assumptions are satisfied.
- Finally, we test our results on both synthetic and real learning data on a categorization task [1].

References

[1] G. Mezzadri. *Statistical inference for categorization models and presentation order*. Phd thesis, Université Côte d'Azur, LJAD, France, 2020.

[2] R. C. Wilson and A. G.E. Collins. Ten simple rules for the computational modeling of behavioral data. *Elife*, 8:e49547, 2019.

[3] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[4] Djallel Bouneffouf and Irina Rish. A survey on practical applications of multi-armed and contextual bandits, 2019.

[5] Pascal Massart. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

[6] Stephen Bates, Trevor Hastie, and Robert Tibshirani. Cross-validation: what does it estimate and how well does it do it?, 2022.

[7] Julien Aubert, Luc Lehéricy, and Patricia Reynaud-Bouret. On the convergence of the MLE as an estimator of the learning rate in the exp3 algorithm. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 1244–1275. PMLR, 7 2023.

[8] Y. Baraud. A Bernstein-type inequality for suprema of random processes with applications to model selection in non-Gaussian regression. *Bernoulli*, 16(4):1064 – 1085, 2010.

[9] Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvari, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *International Conference on Machine Learning*, pages 24325–24360. PMLR, 2023.