On the impact of overfitting in learning to rank using a margin loss: a case study in job recommender systems

Solal Nathan
Laboratoire Interdisciplinaire des Sciences du Numérique (LISN)
solal.nathan@lisn.fr

Guillaume Bied
Laboratoire Interdisciplinaire des Sciences du Numerique (LISN)
Centre de Recherche en Economie et Statistique (CREST)
guillaume.bied@lri.fr

1 Introduction

In learning to rank and recommender systems, it is typically untractable to directly optimize on metrics of interest such as recall or precision. Surrogate losses are instead used for learning: an important case are margin-based loss functions, which seek to separate relevant samples from irrelevant ones. This paper studies the relation between margin loss and the true metric in the real-world setting of job recommender systems.

Intriguingly, in this setting, overfitting the margin loss does not translate to overfitting on the metric of interest. To understand this phenomenon, we introduce novel concepts of participation (the share of training samples with non-zero contribution to the loss) and cycling (stability of the population of samples with non-zero contribution throughout training).

The paper is structured as follows. Section 2 introduces the job recommendation problem of interest and sets up notations. Section 3 presents preliminary results.

2 Learning to recommend jobs with a margin loss

The VADORE job recommendation system We study the margin loss' behavior in the context of VADORE, a job ad recommender system developed by researchers in computer science and economics in collaboration with French Public Employment Service (PES) Pôle emploi [3]. VADORE attempts to reduce frictional unemployment: that is, the part of unemployment due to imperfect information and the impossibility for a job seeker or a recruiter to process all possible job ads. Note that the data at hand is proprietary and highly regulated.

Problem Statement Since VADORE learns from past observed hires, few matches (if any) are observed in the past for any given job seeker. Classical algorithms like Matrix Factorization[2] are therefore not suitable. Instead, in order to recommend a job ad j to a job seeker i, a set of features x_i and y_j, described in appendix A, are leveraged. They are used to learn a matching score:

$$s_{ij} = s(x_i, y_j) \tag{1}$$

For a given jobseeker i, recommendations proceed by sorting job ads according to s_{ij}, and selecting the top k.

In practice, the score is specified as $s(x_i, y_j) = \phi(x_i)^T A \psi(y_j)$, where ϕ and ψ are feedforward neural networks, and A is a matrix. The parameters to be learned are the components of A and the weights of ϕ and ψ.

BayLearn 2022, San Francisco

The true metric of interest to measure the recommender system's performance is the recall@k; that is, the share, for jobseekers who find a job in the validation set, of times the job ad on which they are hired is among the model's top k recommendations.

Since optimizing the recall directly is untractable, a Triplet Margin Loss[1] is used as a surrogate loss. It is defined at the level of a triplet (i, j, j') where the anchor i is a job seeker, the positive example j is a job ad with which i matched, and the negative example j' is a job ad with which i did not match. The goal of this loss is to separate the a set of positive examples from a set of negative examples by a scalar margin $\eta > 0$. Formally, the Triplet Margin loss, to be minimized, takes the form:

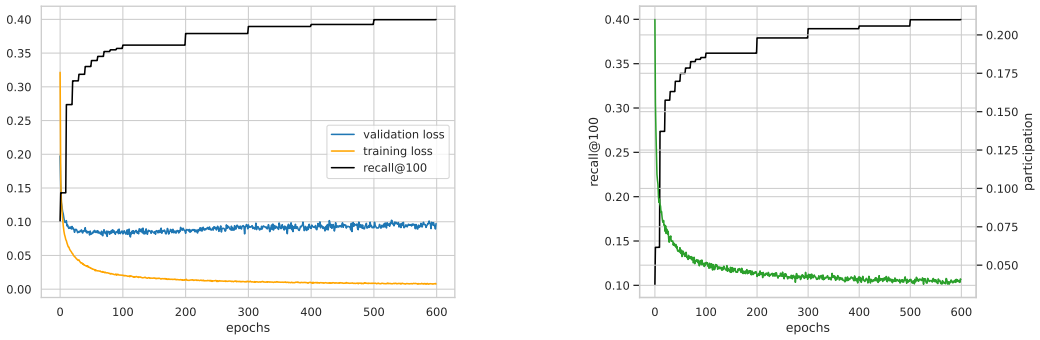
$$\mathcal{L}(i, j, j') = \max\{s_{ij} - s_{ij'} + \eta, 0\} \tag{2}$$

Optimization then proceeds using Adam.

If the triplet margin loss is a good surrogate, we would expect a correlation between the overfitting of the loss and the degradation of the true metric of interest, namely the recall. Since overfitting was not monitored at first, we expect to gain in performance either way (underfitting or overfitting) by measuring it and stopping at an optimal fit. Moreover, if there is no correlation between the overfitting of the surrogate loss and the real metric, it is probably necessary to evaluate the recall on the validation set at different steps during the training to do early stopping.

For computational reasons, it is undesirable to iterate through all possible triplets (i, j, j'). Instead, negative examples j' are sampled uniformly at random (attempts to implement more elaborate sampling techniques were not conclusive).

3 Preliminary results



(a) Training and validation losses compared to recall@100

(b) Participation compared to recall@100

Figure 1: Overfitting experiments on VADORE

Figure 1a documents the evolution the training loss, validation loss and validation recall@100 across 600 epochs. Overfitting between the validation loss and the training loss occurs. The validation loss stops decreasing after the first 50 epochs, and slowly increases throughout the rest of training. However, the surprise is that while the training loss is stagnating, and the validation loss increasing, the recall keeps improving throughout training. We notice diminishing returns over time, yet a steady improvement.

To better understand this behavior, we introduce a new concept: **participation**, defined as the share of triplets (i, j, j') in the training set which participate to the Triplet Margin Loss (Eq 2) at a given epoch; that is, such that $s_{ij} - s_{ij'} \geq \eta$.

Another question which arises from this result: is there **cycling** in the participating pairs at each epoch? That is, are the pairs that contribute to participation (i.e. contribute to the loss) for a given epoch a subset of the ones participating in the previous epoch, or do previously unactive pairs become often become active again during training? (see Figure 1b)

References

[1] Gal Chechik et al. “Large Scale Online Learning of Image Similarity through Ranking”. In: *Pattern Recognition and Image Analysis*. Ed. by Helder Araujo et al. Vol. 5524. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 11–14. ISBN: 978-3-642-02171-8 978-3-642-02172-5. DOI: 10.1007/978-3-642-02172-5_2. URL: http://link.springer.com/10.1007/978-3-642-02172-5_2 (visited on 07/13/2022).

[2] Yehuda Koren, Robert Bell, and Chris Volinsky. *Matrix Factorization Techniques for Recommender Systems*. 2009.

[3] Victor Alfonso Naya et al. *Designing Labor Market Recommender Systems: The Importance of Job Seeker Preferences and Competition*. 2021. URL: https://www.semanticscholar.org/paper/Designing-labor-market-recommender-systems%3A-the-of-Naya-Bied/b085fe9ae2e30e0601d8fa446258c9617473440a (visited on 07/11/2022).

Acknowledgment

This work takes place in the VADORE team. VADORE stands for "Valorisation des données pour la recherche d'emploi" (which could be translated by "VAlorizations of Data to imprOve matching in the laboR markEt"). It is a multidisciplinary team between the LISN and the CREST. The team is composed of Victor Alfonso Naya, Guillaume Bied, Philippe Caillou, Bruno Crepon, Christophe Gaillac, Solal Nathan, Elia Perennes, Michèle Sebag and Francisco Vàsquez.

A Appendix: data and features

Table 1: Brief description of the tabular data

job seeker	job ads
search criterion	job
experiences	required experience
diplomas	required diploma
skills	required skills
plain text	plain text
localization	localization
required salary	salary
driving license	required driving licenses
socio-demographic data	
administrative data	
...	...

Features x_i and y_j describing job seekers and job ads are approximately of size 500 in both cases. The embeddings $\phi(x_i)$ and $\psi(y_j)$ are respectively of size 100.

Statistics on the dataset The dataset is composed of 1.07 million jobseekers and 1.78 million job ads shared among 112 weeks. The train-validation split is done per week. The training set includes a random selection of 85% of the weeks from Jan. 2019 to Dec. 2021; the validation set includes all remaining weeks from 2019 to 2021, plus the first 8 weeks of 2022. On average, circa 400k job seekers, 64k job ads and 1.4k matches are observed per week.