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

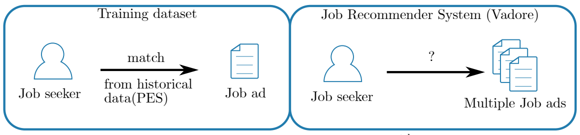
Solal Nathan[1], Guillaume Bied[1,2]
[1] LISN, Université Paris-Saclay
[2] CREST, ENSAE, Institut Polytechnique de Paris

1. Objectives

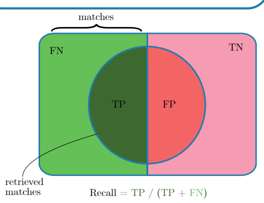
In learning to rank and recommender systems, it is typically untractable to directly optimize on metrics of interest such as recall or precision. Surrogate losses are instead used for learning: an important case are margin-based loss functions, which seek to separate relevant samples from irrelevant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

2. Surrogate loss

Case study: Vadore, a real world Job Recommender Systems (JRS)



- Historical dataset from Pôle Emploi describing job ads and job seekers
- We create a score s_{ij}, by which we rank outputs
- Final goal is to maximise the recall



- Let us define a score function:

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

with x_i the job seeker embedding and y_j the job ad embedding

- The outputs are a bilinear form of neural nets

$$s(x_i,\ y_j) = \sum_k \varphi_k(x_{ik})^T A_k \psi_k(y_{jk})$$

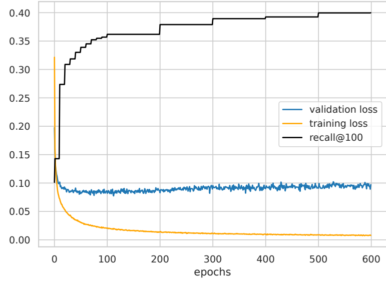
- The metric of interest is the recall@k (share of jobseekers s.t. their future job is among the algorithm's top-k recommendations). However it is too computationally expensive, thus we use the

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

with i an anchor, j a positive example and $j' \neq j$ a negative example

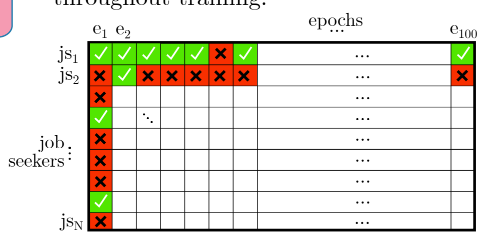
3. Overfitting

Vadore and its submodules were usually trained for 100 epochs. But does it overfit when trained for longer?

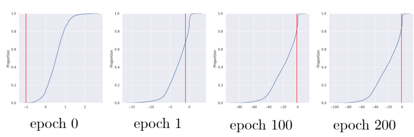


4. Participation and Cycling

- **participation:** the share of training samples with non-zero contribution to the loss.
- **cycling**: stability of the population of samples with non-zero contribution throughout training.



Empirical Cumulative Distribution Function (ECDF) at different epochs, displaying the amount of participation.



Margin in red. It represents the percentage of the population participating to the loss (everything on the right of the margin).

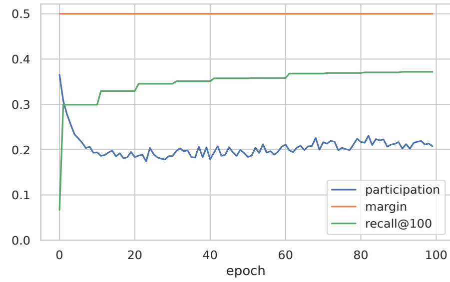
5. Adaptive Margin

- The adaptive margin algorithm enables us to control the participation.
- We set a goal g and control the margin to aim for it.

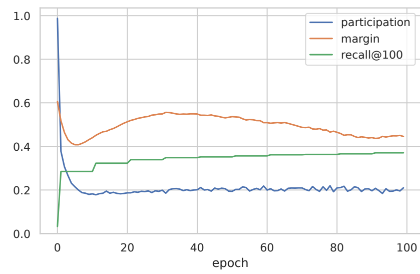
Algorithm 1 Proportional controller for margin-participation

input $\eta_0 \in [0,1]$
input $g \in [0,1]$ ▷ Goal for participation
input K ▷ Proportional constant K_p
$\eta \leftarrow \eta_0$
for e in epochs **do**
  data ← from_dataset()
  $s_{ij}, s_{ij'}$ ← model.forward(data) ▷ Scores
  loss ← $max(s_{ij} - s_{ij'} + \eta, 0)$ ▷ Triplet Margin Loss
  loss.backward()
  p ← compute_participation()
  error ← $g - p$
  $\eta \leftarrow K \times$ error
end for

Participation on the Vadore's General submodule trained at constant margin $\eta = 0.5$ for 100 epochs



Same experiment using the Proportional Controller Algorithm with a participation goal $g = 0.2$



6. Conclusion

- We investigate the relationship between **recall** and the **surrogate loss** in a real world job recommandation setting.
- Surprisingly, **overfitting** on the validation loss coincides with steady improvement on the recall.
- To understand this behavior we introduced two new concepts: **participation** and **cycling**
- Thus creating the **adaptive margin algorithm**, leading to hyperparameter optimisation and curriculum learning in the future.

References

[1] Alfonso Naya, V., Bied, G., Caillou, P., Crépon, B., Gaillac, C., Pérennes, E.: Designing Labor Market Recommender Systems: The Importance of Job Seeker Preferences and Competition, IZA Workshops on Environment and Labor Markets (2021)
[2] Volkovs, M., Yu, G. W., and Poutanen, T. Content-based Neighbor Models for Cold Start in Recommender Systems. In Proceedings of the Recommender Systems Challenge 2017, RecSys Challenge '17