Before using cosine similarity between label representations consider:

a) Which classifier your representations are coming from and

b) Whether cosine similarity can answer your question

Funded by the European Union

European Research Council
Established by the European Commission

What Cosine Similarity of Label Representations Can and Cannot Tell us

Sigmoid classifier

$$p(y_v \mid \mathbf{x}) = \frac{1}{1 + e^{-\mathbf{g}(v)^\top \mathbf{f}(\mathbf{x})}}$$

E.g. Word2Vec, multi-label classifier, labels can be present or not present.

Possible label combinations are captured by the **Sign Pattern**, $\mathbf{W} \in \mathbb{R}^{k \times d}$ has $\mathbf{g}(v)$ as rows

$$\mathcal{S}(\mathbf{W}) = \left\{ \operatorname{sign}(\mathbf{W}\mathbf{z}) : \mathbf{z} \in \mathbb{R}^d,\ \mathbf{w}_i^\top \mathbf{z} \neq 0,\ 1 \leq i \leq k \right\}$$

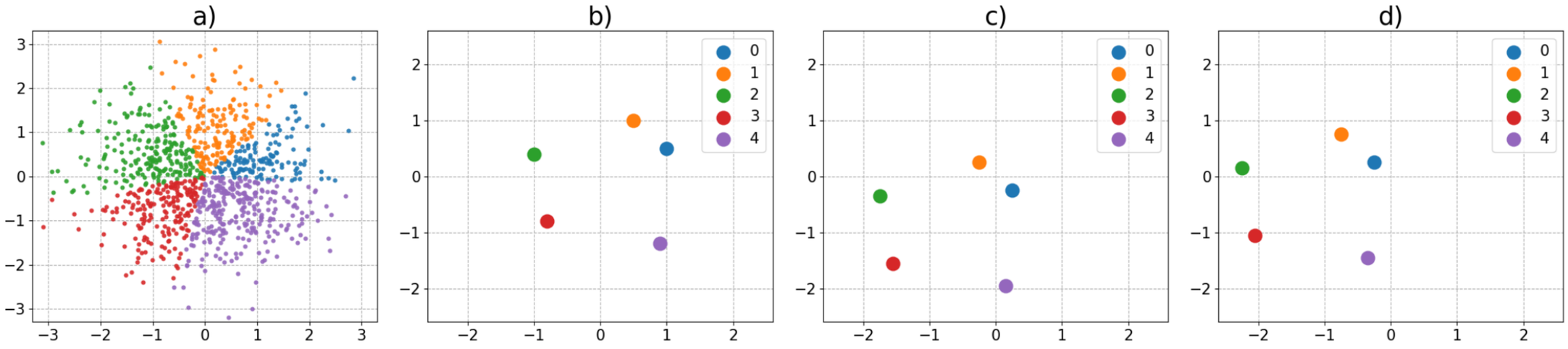
Softmax classifier

$$p(y \mid \mathbf{x}) = \frac{\exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{f}(\mathbf{x})^\top \mathbf{g}(y'))}$$

E.g. Language model, image classifier, gives a ranking of labels.

Possible rankings are captured by the **Ranking Pattern**

$$\mathcal{R}(\mathbf{W}) = \left\{ \begin{matrix} \operatorname{sign}(\mathbf{B}\mathbf{W}\mathbf{z}) : & \mathbf{z} \in \mathbb{R}^d, \\ & (\mathbf{w}_i - \mathbf{w}_j)^\top \mathbf{z} \neq 0, \\ & 1 \leq i < j \leq k \end{matrix} \right\}$$



Above, 4 softmax models with the same probabilities, but different cosine similarities between unembeddings.

Pairwise cosine similarity of label representations (unembeddings)

cannot tell us the probabilities (Thm 3.4)

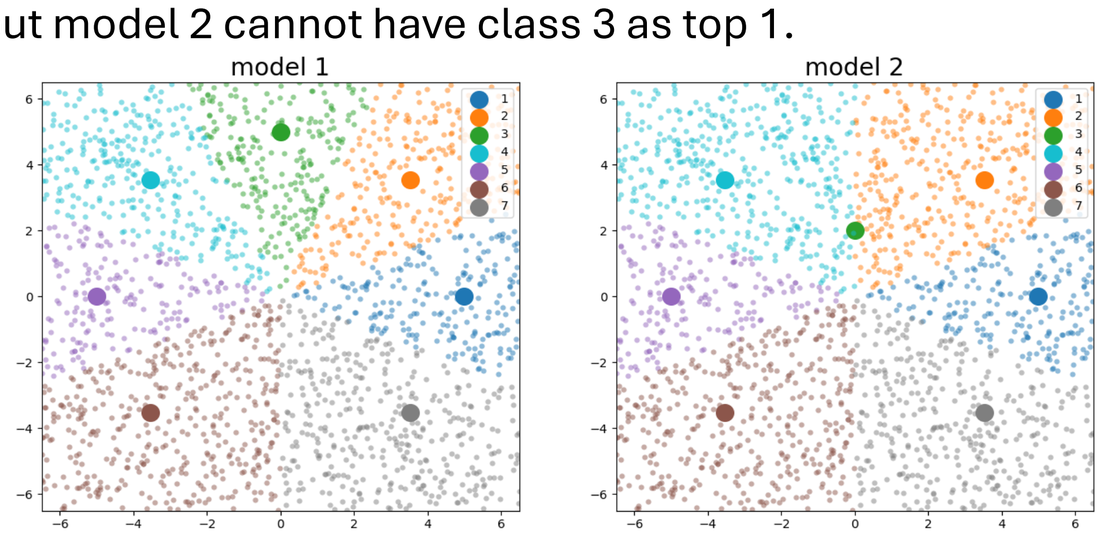
can tell us the sign pattern (Lem 5.3.)

cannot tell us the ranking pattern (Thm 5.5)

Cosine similarities of differences of unembeddings

can tell us the ranking pattern (Thm 5.6)

Below, Cosine similarities of unembeddings are the same, but model 2 cannot have class 3 as top 1.



Arxiv

Code

Beatrix M. G. Nielsen, Andreas Grivas
Mechanistic Interpretability Workshop, ICML 2026

IT UNIVERSITY OF CPH

THE UNIVERSITY of EDINBURGH
School of Mathematics