Deep Learning for Natural Language Processing

Lecture 3, part 1: Representing documents for neural networks

UNIVERSITY OF GOTHENBURG

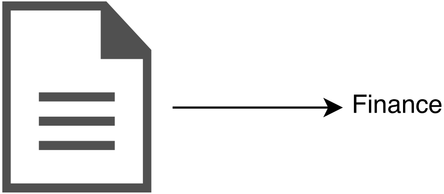
CHALMERS

WASP WALLENBERG AI, AUTONOMOUS SYSTEMS AND SOFTWARE PROGRAM

Richard Johansson

richard.johansson@gu.se

case study: categorizing documents



representation

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	target
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

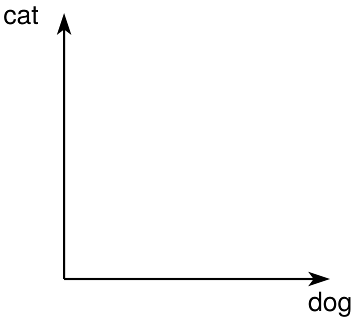
representation

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	target
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K

$$\rightarrow \begin{bmatrix} -7 & 8 & -12 & -19 & 10 & 0 & -1 & 17 \\ 5 & 5 & 15 & 13 & 3 & -13 & -11 & 16 \\ -19 & 0 & -14 & 20 & 9 & 0 & -1 & -18 \\ 7 & -19 & -18 & 1 & -17 & 13 & 13 & 9 \\ -14 & 7 & 1 & 19 & 6 & 12 & -2 & -3 \\ 13 & -17 & -15 & -13 & -4 & 14 & 12 & -18 \\ -4 & 1 & -3 & 6 & 5 & -9 & -3 & -20 \\ 20 & -14 & -16 & -5 & -12 & 0 & -7 & 19 \end{bmatrix}$$

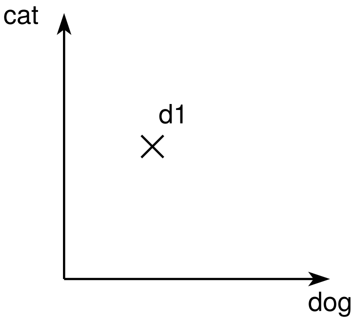
old-school solution: bag-of-words representation

d1 = "cat dog cat cat dog"



old-school solution: bag-of-words representation

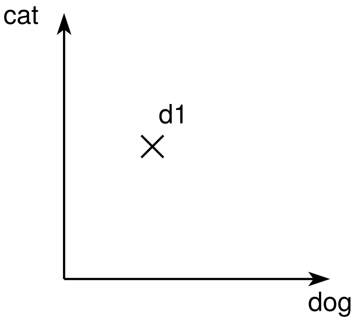
d1 = "cat dog cat cat dog"



old-school solution: bag-of-words representation

d1 = "cat dog cat cat dog"

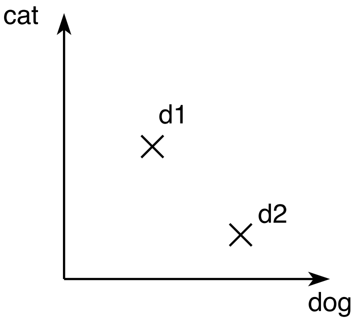
d2 = "dog cat dog dog dog"



old-school solution: bag-of-words representation

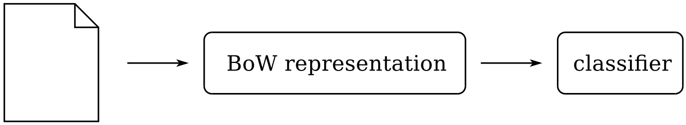
d1 = "cat dog cat cat dog"

d2 = "dog cat dog dog dog"



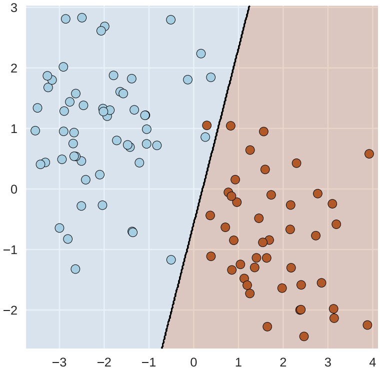
building a document categorizer with bag of words

- to build the complete document classifier, we combine the bag-of-words representation with a linear or nonlinear (neural) classifier



old-school solution, continued: linear classification

$$\text{score}(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b$$



interpretation of linear models with BoW representation

$$
\begin{aligned}
\text{score}_{\text{Finance}}(d) \quad = \quad & 0.73 \cdot \text{count}(\text{dollars}) \\
& + 0.64 \cdot \text{count}(\text{market}) \\
& + 0.89 \cdot \text{count}(\text{bonds}) \\
& \ldots \\
& - 0.45 \cdot \text{count}(\text{football}) \\
& - 0.75 \cdot \text{count}(\text{goals}) \\
& \ldots
\end{aligned}
$$

limitations of linear model with BoW representation

- words are atomic: we don't model their "meaning"
- difficult to model interactions: explicit features are needed
- "knowledge bottleneck" in feature design
- difficult to reuse knowledge from other tasks

limitations of linear model with BoW representation

- words are atomic: we don't model their "meaning"
- difficult to model interactions: explicit features are needed
- "knowledge bottleneck" in feature design
- difficult to reuse knowledge from other tasks
- these are all limitations of **representation**!