Master on Artificial Intelligence Universitat Politècnica de Catalunya

Advanced Human Language Technologies

Final Exam

June 3^{rd}, 2021

Exercise 1. Estimation & Smoothing

We want to build a probabilistic model for sentiment analysis on tweets, and for that we have gathered the following training data:

- We have a collection of 1,000 tweets with gold standard annotations. 300 tweets are annotated as *positive* (POS), 250 as *negative* (NEG), and the remaining 450 are annotated as *neutral* (NEU).
- We collected some statistics about words appearing in these tweets:
  - 140 tweets contain the word *big*. 70 of them are annotated as NEU, 60 as POS, and 10 as NEG.
  - 90 tweets contain the word *great*. 50 of them are annotated as POS.
  - 100 tweets contain the word *awful*. 80 of them are annotated as NEG.
  - 40 tweets contain the word *terrific*. All of them are annotated as POS.

1. Compute the following probability values corresponding to a MLE model, and those smoothed using linear discount with $\alpha = 0.1$. Justify your answer and the values chosen for N and N_0 in each case where it applies.
   - Probability that a tweet is positive, $P(\text{POS})$
   - Probability that a tweet contains word *big*, $P(big)$
   - Probability that a tweet is positive and contains word *big*, $P(\text{POS} \wedge big)$
   - Probability that a negative tweet contains word *awful*, $P(awful|\text{NEG})$
2. Compute the following probability values corresponding to a MLE model, those corresponding to a model smoothed with Laplace's Law, and those smoothed using linear discount with $\alpha = 0.1$. Justify your answer and the values chosen for B, N, and N_0 in each case where it applies.
   - Probability that a tweet containing word *great* is positive, $P(\text{POS}|great)$
   - Probability that a tweet containing word *terrific* is positive, $P(\text{POS}|terrific)$
   - Probability that a tweet containing word *terrific* is negative, $P(\text{NEG}|terrific)$

   Develop your computations, do not provide just a numeric result. You may leave probabilities as fractions.

SOLUTION

1.

	P_{MLE}	P_{LD}	justification
$P(\text{POS})$	$\frac{\#\text{POS}}{N} = \frac{300}{1{,}000}$	$0.9\frac{300}{1{,}000}$	We have 1,000 tweets, i.e. 1,000 cases where a label may be observed, thus $N = 1,000$.
$P(big)$	$\frac{\#big}{N} = \frac{140}{1{,}000}$	$0.9\frac{140}{1{,}000}$	We have 1,000 tweets, i.e. 1,000 cases where a word may be observed, thus $N = 1,000$.
$P(\text{POS} \wedge big)$	$\frac{\#(\text{POS} \wedge big)}{N} = \frac{60}{1{,}000}$	$0.9\frac{60}{1{,}000}$	We have 1,000 tweets, i.e. 1,000 cases where a word-label co-occurrence may be observed, thus $N = 1,000$.
$P(awful\|\text{NEG})$	$\frac{\#(\text{NEG} \wedge awful)}{\#NEG} = \frac{80}{250}$	$0.9\frac{80}{250}$	We are counting word ocurrences conditioned to the tweet being NEG, so, $N = 250$.

2.

	P_{MLE}	P_{LAP}	P_{LD}	justification
$P(\text{POS}\|great)$	$\frac{\#(\text{POS} \wedge great)}{\#great} = \frac{50}{90}$	$\frac{50+1}{90+3}$	$0.9\frac{50}{90}$	We are counting events conditioned to the occurrence of *great* in a tweet, so $N = 90$. For P_{LAP}, Possible labels are (POS, NEG, NEU), thus $B = 3$.
$P(\text{POS}\|terrific)$	$\frac{\#(\text{POS} \wedge terrific)}{\#terrific} = \frac{40}{40}$	$\frac{40+1}{40+3}$	$0.9\frac{40}{40}$	We are counting events conditioned to the occurrence of *terrific* in a tweet, so $N = 40$. For P_{LAP}, Possible labels are (POS, NEG, NEU), thus $B = 3$.
$P(\text{NEG}\|terrific)$	$\frac{\#(\text{NEG} \wedge terrific)}{\#terrific} = \frac{0}{40}$	$\frac{0+1}{40+3}$	$\frac{\alpha}{N_0} = \frac{0.1}{2}$	We are counting events conditioned to the occurrence of *terrific* in a tweet, so $N = 40$. For P_{LAP}, , Possible labels are (POS, NEG, NEU), thus $B = 3$. For P_{LD}, N_0 is the number of outcomes unobserved with *terrific*. There are 3 possible outcomes (POS, NEG, NEU) but only one (POS) was observed, thus $N_0 = 2$

Exercise 2. Distances/Similarities

Given the sentences:

S1: *The man saw a car in the park*
S2: *I saw the man park the car*

Compute *similarity* between them using the following measures (if the measure yields a distance, convert the result to a similarity).

1. Euclidean
2. Vector cosine
3. Jaccard
4. Overlap

Provide the vector or set representation for each sentence used in each case. Develop your computations.

SOLUTION

Vector representations:

	the	man	saw	a	car	in	park	I
S1:	**2**	1	1	1	1	1	1	0
S2:	**2**	1	1	0	1	0	1	1

1. Euclidean: It is a distance, requires conversion.

   $$d = \sqrt{(2-2)^2 + (1-1)^2 + (1-1)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (1-1)^2 + (0-1)^2} = \sqrt{3} \approx 1.732$$

   Conversion to similarity:

   $$s = \frac{1}{1+d} = \frac{1}{1+1.732} = 0.366$$

2. Vector cosine: Already a similarity

   $$s = \frac{2 \cdot 2 + 1 \cdot 1 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 1 \cdot 0 + 1 \cdot 1 + 0 \cdot 1}{\sqrt{2^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2}\sqrt{2^2 + 1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2}} = \frac{8}{\sqrt{10}\sqrt{9}} \approx 0.843$$

Set representations:

	the	man	saw	a	car	in	park	I
S1:	1	1	1	1	1	1	1	0
S2:	**1**	1	1	0	1	0	1	1

3. Jaccard: Already a similarity

   $$s = \frac{|S1 \cap S2|}{|S1 \cup S2|} = \frac{5}{8} = 0.625$$

4. Overlap: Already a similarity

   $$s = \frac{|S1 \cap S2|}{min(|S1|, |S2|)} = \frac{5}{min(7,6)} = 0.833$$

Exercise 3. Features for log linear sequence annotation models

Negation detection is a task consisting on identifying which phrases in a sentence are affected by a negation. It is a vital task e.g. in applications related to the processing of medical documents.

The task is often modeled as a B-I-O labeling task, and solved using sequence-labeling algorithms such as CRFs.

We have the following training data:

$\mathcal{X}$	The	patient	does	not	show	any	lung	symptoms	.		
$\mathcal{Y}$	O	O	O	B	I	I	I	I	O		
$\mathcal{X}$	Dark	spots	were	observed	in	lung	X-ray	imaging	.		
$\mathcal{Y}$	O	O	O	O	O	O	O	O	O		
$\mathcal{X}$	Exhoglovifin	never	caused	adverse	reactions	and	should	not	be	banned	.
$\mathcal{Y}$	O	B	I	I	I	O	O	B	I	I	O

And the following feature templates:

$$\mathbf{f}_{1,a,l}(\mathcal{X}, i, t) = \begin{cases} 1 & \text{if } w_i = a \wedge t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{2,l}(\mathcal{X}, i, t) = \begin{cases} 1 & \text{if } w_{i-1} \in \{no, not, never, any\} \wedge t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{3,l}(\mathcal{X}, i, t) = \begin{cases} 1 & \text{if } punctuation(w_i) \wedge\ t = l \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{f}_{4,l}(\mathcal{X}, i, t) = \begin{cases} 1 & \text{if } w_{i-1} = dark \wedge\ w_i = spots \wedge\ t = l \\ 0 & \text{otherwise} \end{cases}$$

1. Which is the dimension of the feature space instantiated by this dataset? Justify your answer.
2. Given the following test sentence $\mathcal{X}$ and hypothesis tag sequence $\mathcal{Y}$:

$\mathcal{X}$	X-Ray	results	do	not	show	any	dark	spots	.
$\mathcal{Y}$	O	O	O	B	I	I	I	I	O

   compute the feature vectors $\mathbf{f}(\mathcal{X}, i, t)$ for each position i, and the global feature vector $\mathbf{f}(\mathcal{X}, \mathcal{Y})$. Highlight which features in the global vector that are present in the vector space instantiated by the three training sentences above.

SOLUTION

1. Feature f_1 is instantiated for each combination word-label seen in the training data. Sentence 1 contains 9 combinations. Sentence 2 contains 8 new combinations –combination (.,O) is repeated. Sentence 3 contains 9 new combinations –combinations (not,B) and (.,O) are repeated. Total 9+8+9 = 26 feature instances for template f_1.

   Feature f_2 is instantiated for each occurrence of `not`, `never`, or `any` combined with a label. Sentence 1 contains one occurrence (with label B), sentence 2 does not contain any, and sentence 3 contains two more occurrences, also with label B, so they generate the same feature $f_{2,B}$. Total, 1 feature instances for template f_2.

   Feature f_3 is instantiated for each occurrence of a punctuation sign combined with a label. Each sentence has one occurrence of the combination (.,O), thus only one instance is generated for f_3.

Feature f_4 is instantiated for each occurrence of *dark spots* combined with a label. This only happens once in sentence 2 (with label O), thus only one instance is generated for f_4.

So, the total number of generated features (i.e. our feature space dimension) is $26+1+1+1 = 29$.

2. Feature vectors for each position are:

$$\mathbf{f}(\mathcal{X}, 1, \texttt{O}) = \{f_{1,XRay,O}\}$$
$$\mathbf{f}(\mathcal{X}, 2, \texttt{O}) = \{f_{1,results,O}\}$$
$$\mathbf{f}(\mathcal{X}, 3, \texttt{O}) = \{f_{1,do,O}\}$$
$$\mathbf{f}(\mathcal{X}, 4, \texttt{B}) = \{f_{1,not,B}\}$$
$$\mathbf{f}(\mathcal{X}, 5, \texttt{I}) = \{f_{1,show,I}, f_{2.I}\}$$
$$\mathbf{f}(\mathcal{X}, 6, \texttt{I}) = \{f_{1,any,I}\}$$
$$\mathbf{f}(\mathcal{X}, 7, \texttt{I}) = \{f_{1,dark,I}, f_{2,I}\}$$
$$\mathbf{f}(\mathcal{X}, 8, \texttt{I}) = \{f_{1,spots,I}, f_{4,I}\}$$
$$\mathbf{f}(\mathcal{X}, 9, \texttt{O}) = \{f_{1,.,O}, f_{3,O}\}$$

Thus, the global feature vector $\mathbf{f}(\mathcal{X}, \mathcal{Y}) = \sum_i \mathbf{f}(\mathcal{X}, i, \mathcal{Y}_i)$ is:

feature	value	in training feature space?
$f_{1,XRay,O}$	1	✓
$f_{1,results,I}$	1	× (word *results* is not in training)
$f_{1,do,O}$	1	× (word *do* is not in training)
$f_{1,not,B}$	1	✓
$f_{1,show,I}$	1	✓
$f_{2,I}$	2	✓
$f_{1,any,I}$	1	✓
$f_{1,dark,I}$	1	× ($f_{1,dark,O}$ appears in training, but $f_{1,dark,I}$ does not)
$f_{1,spots,I}$	1	× ($f_{1,spots,O}$ appears in training, but $f_{1,spots,I}$ does not)
$f_{4,I}$	1	× ($f_{4,O}$ appears in training, but $f_{4,I}$ does not)
$f_{1,.,O}$	1	✓
$f_{3,O}$	1	✓

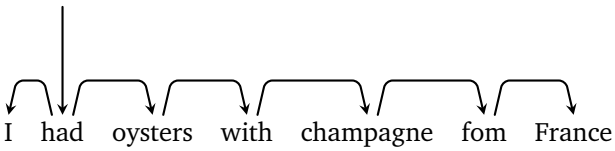
Exercise 4. Parsing

Given the sentence *I had oysters with champagne from France*.

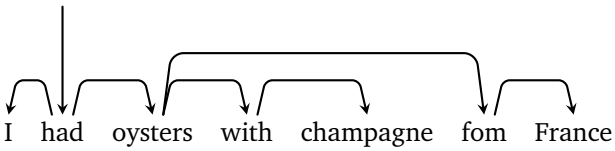
1. Draw unlabeled dependency trees for the following interpretations:
   (a) I had oysters which had champagne on them. The champagne was from France.
   (b) I had oysters which had champagne on them. The oysters were from France.
   (c) I had oysters while having also champagne. The champagne was from France.
   (d) I had oysters while having also champagne. The oysters were from France.
2. Is any of the obtained trees non-projective? Justify your answer.

SOLUTION

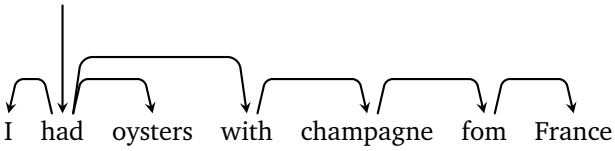
1. Draw unlabeled dependency trees for the following interpretations:
   (a) I had oysters which had champagne on them. The champagne was from France.



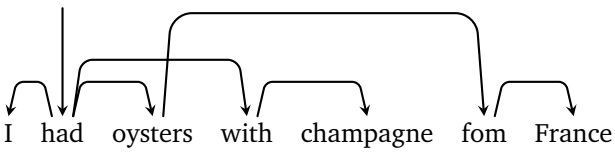
   (b) I had oysters which had champagne on them. The oysters were from France.



   (c) I had oysters while having also champagne. The champagne was from France.



   (d) I had oysters while having also champagne. The oysters were from France.



2. Structure (d) is non-projective, since there are crossing arcs.

Exercise 5. Neural Language Modeling

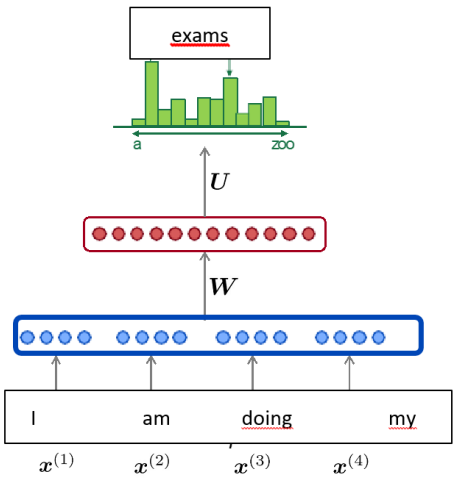
Feedforward Neural Networks (FNN) and Recurrent Neural Networks (RNN) were the two first neural network architectures that were proposed for Language Modeling. Later, the Transformer, was proposed.

1. Draw the language model schemes in both FNN and RNN at inference time for the input sentence *I am doing my* and the predicted word of *exams*. Suppose the feedforward network has a context of 3 previous words.
2. What is the limitation of feedforward neural networks with respect to recurrent neural networks when we are training a language model?
3. What are the improvements of the Transformer over the RNN?

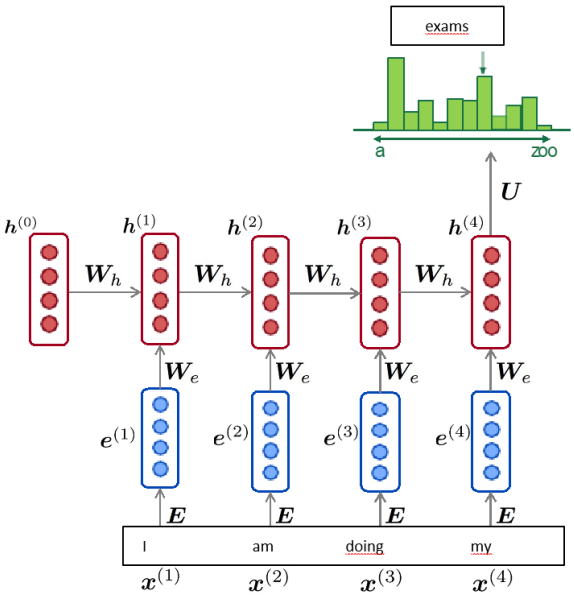
SOLUTION

1.

FNN schema



RNN schema



2. The limitation of the FNNs with respect to RNNs is the fact that they are using limited context.
3. The transformer allowed to parallelize the training procedure because the encoder is based on attention-based modules and there is no recurrence involved.

Exercise 6. Transformers and Advanced Transformers

1. Summarize key aspects of the Transformer architecture and describe at least 2 remaining challenges.
2. Explain at least 2 tasks on which Efficient Transformers based on the Transformer Encoder can be trained.

SOLUTION

1. *Key aspects:* self-attention, multi-head positional encoding, masking, enc-dec attention.

   Challenges: auto-regressive inference, large amount of data required to train.
2. Masking, next sentence prediction

Exercise 7. Ethics in AI

1. Bias is a major area of ethical concern in Artificial Intelligence. Fill in the following tables regarding the examples of bias cases in AI and resources to mitigate these cases:

Example of bias case	**Application**	**Ethical concern**
COMPAS		
Gender Shades		

Resource	**Application**	**Advantatge**
GeBioToolkit		
MT-DataSheet for Dataset		

2. Explain why word embeddings contain gender biases. Explain if contextual word embeddings still keep these biases. Mention at least two methodologies to evaluate these biases and mention what it is necessary to work with the proposed evaluation methods.

SOLUTION

1.

Example of bias case	**Application**	**Ethical concern**
COMPAS	predict recidivism risk	Accuracy varies with race: Darker skins, higher risk
Gender Shades	face recognition	Accuracy varies with race: Darker skins, lower accuracy

Resource	**Application**	**Advantatge**
GeBioToolkit	Machine Translation	Balanced in gender
MT-DataSheet for Dataset	Machine Translation	Details about the corpus

2. Word embeddings contain gender biases because certain neutral professions are associated to male or female depending on the stereotype. Contextual word embeddings still keep these biases, but in a lower amount. We can evaluate using direct bias measure, classification, clustering. For all these evaluations, we need lists of pre-defined words.