Optional tests

Marco Kuhlmann

01 Text classification (3 points)

a) Use Maximum Likelihood estimation with add-one smoothing to estimate the class probabilities and word probabilities of a Naive Bayes text classifier from the following document collection. Assume that the vocabulary consists of the set of all words occurring in the documents. Answer with fractions.

document	class
ant	A
ant bear	B
bear camel	B
camel	C

b) Based on the probabilities just estimated, compute the class-specific scores that the Naive Bayes classifier uses to predict the class for the following document:

ant bear camel

Answer with fractions.

c) Here are some class frequencies in a document collection:

	class X	class Y	class Z
training data	2,460	2,952	1,968
test data	738	492	615

What is the precision for class X of the most frequent class baseline on the test data? Answer with a fraction.

Sample answers:

a) Estimated probabilities:

$$
\begin{array}{llll}
P(\mathtt{A}) = 1/4 & P(\text{ant} \mid \mathtt{A}) = 2/4 & P(\text{bear} \mid \mathtt{A}) = 1/4 & P(\text{camel} \mid \mathtt{A}) = 1/4 \\
P(\mathtt{B}) = 2/4 & P(\text{ant} \mid \mathtt{B}) = 2/7 & P(\text{bear} \mid \mathtt{B}) = 3/7 & P(\text{camel} \mid \mathtt{B}) = 2/7 \\
P(\mathtt{C}) = 1/4 & P(\text{ant} \mid \mathtt{C}) = 1/4 & P(\text{bear} \mid \mathtt{C}) = 1/4 & P(\text{camel} \mid \mathtt{C}) = 2/4
\end{array}
$$

b) Class-specific scores:

$$
\begin{aligned}
\text{score}(\mathtt{A}) &= P(\mathtt{A}) \cdot P(\text{ant} \mid \mathtt{A}) \cdot P(\text{bear} \mid \mathtt{A}) \cdot P(\text{camel} \mid \mathtt{A}) = \frac{1}{4} \cdot \frac{2}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \\
\text{score}(\mathtt{B}) &= P(\mathtt{B}) \cdot P(\text{ant} \mid \mathtt{B}) \cdot P(\text{bear} \mid \mathtt{B}) \cdot P(\text{camel} \mid \mathtt{B}) = \frac{2}{4} \cdot \frac{2}{7} \cdot \frac{3}{7} \cdot \frac{2}{7} \\
\text{score}(\mathtt{C}) &= P(\mathtt{C}) \cdot P(\text{ant} \mid \mathtt{C}) \cdot P(\text{bear} \mid \mathtt{C}) \cdot P(\text{camel} \mid \mathtt{C}) = \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{1}{4} \cdot \frac{2}{4}
\end{aligned}
$$

c) $\frac{0}{0}$ (which is mathematically undefined)

02 Language modelling **(3 points)**

The WikiText language modelling dataset is a collection of 2 million tokens extracted, comprising a vocabulary of 33,000 unique words. We have the following selected counts of unigrams and bigrams:

the	*book*	*first*	*the book*	*book the*	*first book*	*book first*
113,161	611	3,981	200	1	8	0

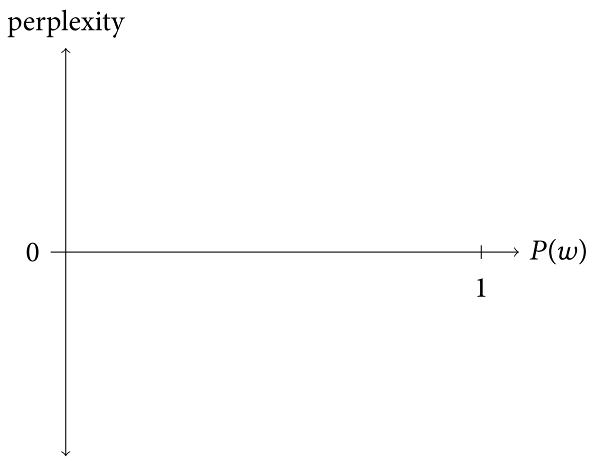
a) Estimate the following probabilities using maximum likelihood estimation without smoothing. Answer with fractions.

i. $P(\mathit{first})$ ii. $P(\mathit{book} \mid \mathit{first})$

b) Now, use additive smoothing with $k = 0.05$.

i. $P(\mathit{first})$ ii. $P(\mathit{first} \mid \mathit{book})$

c) We evaluate a unigram language model on a one-word sentence w. Sketch how the perplexity of the model varies with $P(w)$ by completing the following diagram. What is the minimal value for the perplexity measure?



Sample answers:

i. i. $\frac{3981}{2000000}$ ii. $\frac{8}{3981}$

ii. i. $\frac{3981+0.05}{2000000+0.05\times 33000}$ ii. $\frac{0+0.05}{611+0.05\times 33000}$

iii. The graph has the same shape as that for entropy in slide 35 from the slide deck for Unit 2, but the minimal value is 1 instead of 0.

03 Part-of-speech tagging **(3 points)**

a) The evaluation of a part-of-speech tagger produced the confusion matrix shown below. The marked cell gives the number of times the system tagged a word as an adjective (`ADJ`) whereas the gold standard specified it as a noun (`NOUN`).

	ADJ	DET	NOUN	VERB
ADJ	1475	0	221	31
DET	5	1835	3	0
NOUN	45	5	3887	167
VERB	28	1	387	2135

Compute the following values. Answer with fractions.

i. precision on verbs

ii. recall on adjectives

b) Training a Hidden Markov Model (HMM) amounts to estimating two types of probabilities. What is the total number of probability values you need to estimate when training a model with 10 tags and a vocabulary of 29,508 unique words? Answer with a formula that evaluates to a concrete number (example: 2×3). Ignore the beginning-of-sentence and end-of-sentence markers.

c) One difference between a multi-class perceptron tagger and a tagger based on an HMM is in the feature sets. Which (zero or more) of the following features would you have to choose to provide the multi-class perceptron tagger with the same information that the HMM tagger has access to?

i. current word

ii. word to the left of the current word

iii. word to the right of the current word

iv. part-of-speech tag of the word to the left of the current word

Sample answers:

a) i. $\frac{2135}{31+0+167+2135}$ ii. $\frac{1475}{1475+0+221+31}$

b) $10 \times 10 + 10 \times 29508$

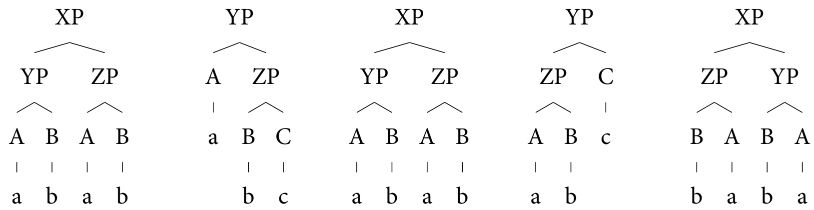
c) i. and iv.

04	**Syntactic analysis**	**(3 points)**

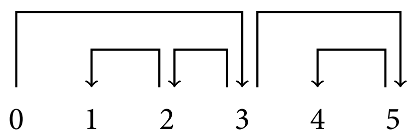
a) You sum up all rule probabilities in a certain probabilistic context-free grammar. Which (zero or more) of the following values can you *not* get as a result?

i. 0.42 ii. 1 iii. 4.2 iv. 42

b) Below is a small phrase structure treebank. Read off all rules with left-hand sides XP, YP and ZP and estimate their rule probabilities using maximum likelihood estimation (no smoothing).



c) State two different sequences of transitions that make the transition-based dependency parser produce the following dependency tree:



Sample answers:

a) i. and iii.

b) Rules and estimated probabilities:

$$\mathrm{XP} \to \mathrm{YP}\ \mathrm{ZP}\ \tfrac{2}{3} \qquad \mathrm{XP} \to \mathrm{ZP}\ \mathrm{YP}\ \tfrac{1}{3}$$

$$\mathrm{YP} \to \mathrm{A}\ \mathrm{B}\ \tfrac{2}{5} \qquad \mathrm{YP} \to \mathrm{A}\ \mathrm{ZP}\ \tfrac{1}{5} \qquad \mathrm{YP} \to \mathrm{ZP}\ \mathrm{C}\ \tfrac{1}{5} \qquad \mathrm{YP} \to \mathrm{B}\ \mathrm{A}\ \tfrac{1}{5}$$

$$\mathrm{ZP} \to \mathrm{A}\ \mathrm{B}\ \tfrac{3}{5} \qquad \mathrm{ZP} \to \mathrm{B}\ \mathrm{C}\ \tfrac{1}{5} \qquad \mathrm{ZP} \to \mathrm{B}\ \mathrm{A}\ \tfrac{1}{5}$$

c) Possible answers:

- SH SH SH LA SH LA SH SH LA RA RA
- SH SH SH LA SH SH SH LA RA LA RA

05 Semantic analysis (3 points)

a) Choose the correct semantic relation: synonym, antonym, hyponym, hypernym?

pigeon	is a/an … of	animal
big	is a/an … of	large
parent	is a/an … of	child
begin	is a/an … of	start
screwdriver	is a/an … of	tool

b) Here are three signatures (glosses and examples) from Wiktionary for different senses of the word *course*:

1 A normal or customary sequence. **2** A learning program, as in university. *I need to take a French course.* **3** The direction of movement of a vessel at any given moment. *The ship changed its course 15 degrees towards south.*

Based on these signatures, which of the three senses of the word *course* does the Lesk algorithm predict in the following sentence? Ignore the word *course*, punctuation, and stop words.

In the United States, the normal length of a course is one academic term.

c) We read off word vectors from the following co-occurrence matrix (target words correspond to rows, context words correspond to columns):

	caws	*dafad*
cheese	6	2
sheep	0	4
goat	1	6
bread	5	0

Sort the target words in decreasing degree of semantic similarity (most similar to least similar) to the word *cheese*, assuming that semantic similarity is measured in terms of cosine similarity.

Sample answers:

a) Semantic relations:

pigeon	is a hyponym of	animal
big	is a synonym of	large
parent	is an antonym of	child
begin	is a synonym of	start
screwdriver	is a hyponym of	tool

b) Sense 1 (match with *normal*)

c) *cheese*, *bread*, *goat*, *sheep*