Practice exam questions

> **Note**
>
> This document gives *examples* for tasks similar to those that will appear on the digital written exam. The **solutions are provided at the end** of this document, so you can first try solving the tasks for yourself.
>
> Be aware that the final exam may have more/fewer questions, other types of tasks/questions, or ask about other course topics than what is shown here. Also, to achieve the highest grades, there may be free-text essay-style questions (requiring longer, well-formulated answers) in addition to what is shown here.

Task 1

What is the main reason for using separate training and test sets when training and evaluating a machine learning classifier?

- ○ To make the training process faster by reducing the amount of data the model sees.
- ○ To estimate how well the classifier generalizes to new, unseen data.
- ○ To guarantee that the model achieves a higher accuracy.
- ○ To allow the model to learn from its mistakes during evaluation.

Task 2

Given the following counts of sentiment class labels in a document collection, what is the accuracy of the most-frequent class baseline on the test data?

	positive	neutral	negative
training data	150	66	175
test data	55	12	50

Task 3

Below is a confusion matrix from evaluating a classifier that predicts which category of a newspaper an article was published under. Rows correspond to gold-standard class labels, while columns correspond to predicted class labels; for example, the **highlighted** cell contains the number of times the classifier predicted *"lifestyle"* where the gold-standard class was *"culture"*.

	news	**sports**	**culture**	**lifestyle**
news	505	0	2	4
sports	0	428	15	8
culture	1	4	320	**45**
lifestyle	10	15	72	155

Based on the confusion matrix above, compute the following evaluation metrics:

precision with respect to *"news"*	**recall with respect to *"culture"***

Task 4

Which of the following probabilities does a bigram language model use to compute the probability of the following sentence?

My cat sleeps

Mark all that apply.

- ☐ $P(\text{cat} \mid \text{sleeps})$
- ☐ $P(\text{cat} \mid \text{My})$
- ☐ $P(\text{My} \mid \text{<BOS>})$
- ☐ $P(\text{<EOS>} \mid \text{cat sleeps})$
- ☐ $P(\text{<EOS>} \mid \text{sleeps})$
- ☐ $P(\text{sleeps} \mid \text{<EOS>})$
- ☐ $P(\text{My cat} \mid \text{sleeps})$
- ☐ $P(\text{cat} \mid \text{<BOS> My})$
- ☐ $P(\text{sleeps} \mid \text{cat})$

Task 5

Consider a dataset of newspaper articles, containing a total of 2000000 (two million) tokens with a vocabulary of 40000 unique words. The following counts of unigrams and bigrams were extracted from this dataset:

claims	*report*	*claims report*	*report claims*
1550	2672	29	190

Estimate the following probabilities using maximum likelihood estimation without smoothing.

P(*claims*)	P(*report* | *claims*)

Task 6

Using the same unigram and bigram counts as in Task 5, estimate the following probabilities using maximum likelihood estimation with Laplace smoothing (*i.e.* add-k smoothing with $k = 1$).

P(*report*)	P(*claims* | *report*)

Task 7

Language models are often evaluated using the perplexity measure. Which of the following statements are true? Mark all that apply.

- ☐ The perplexity will always be a value ≥ 1.
- ☐ The perplexity will always be lower than the entropy.
- ☐ A high probability score corresponds to a high perplexity value.
- ☐ Perplexity is an example for an intrinsic evaluation metric.

Task 8

Below is a sentence with named entity spans highlighted, both according to a gold-standard dataset and the prediction of named entity recognition (NER) model.

gold-standard: *The Bangladesh National Party was founded by Ziaur Rahman in 1978.*

prediction: *The Bangladesh National Party was founded by Ziaur Rahman in 1978.*

Based on the spans above, compute the following span-level evaluation metrics:

precision	recall	F1-score

Task 9

Below is a sentence with named entity spans highlighted and annotated with their type, for example “PER” for a span indicating an entity of type “person”.

The Seattle Seahawks are an American football team.
ORG GPE

Convert these named entity annotations to BIO notation by filling out the table below. Assume that each row of the table corresponds to a single token, so that each row is annotated with exactly one tag.

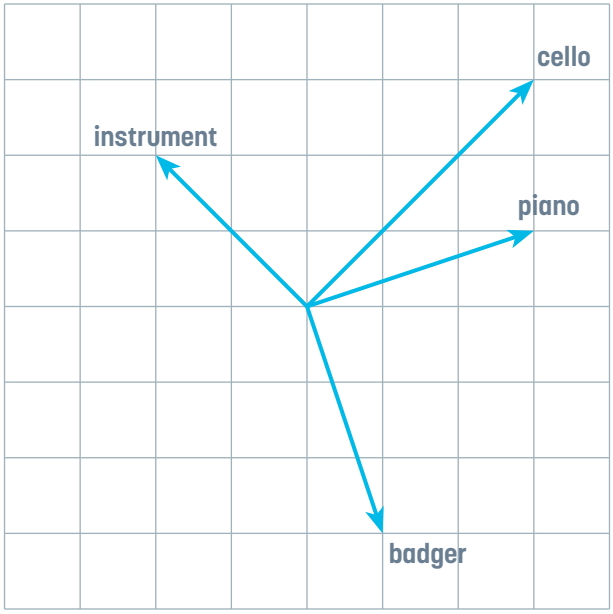
token	tag
The	
Seattle	
Seahawks	
are	
an	
American	
football	
team.	

Task 10

State the distributional hypothesis (in one sentence).

Task 11

Below is a plot of word embeddings in a two-dimensional vector space. Each vector is shown as an arrow from the origin.



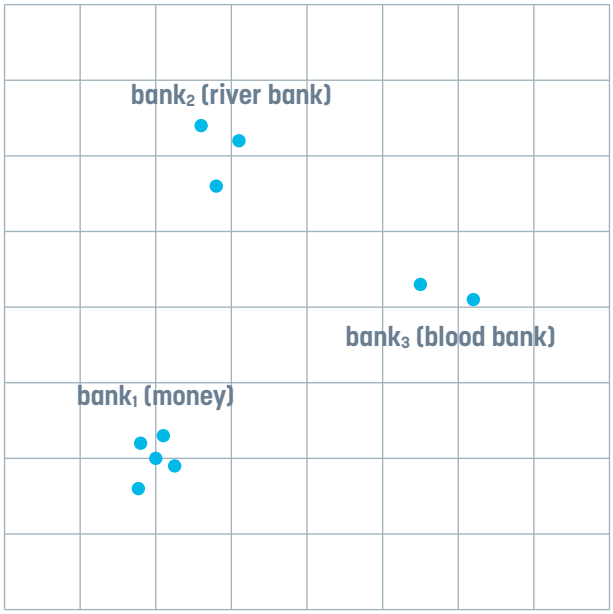
Given the embeddings as shown above, match the word pairs listed below (1–4) to the cosine similarity values (a–d) of their respective embeddings.

1. piano – cello
2. cello – instrument
3. instrument – badger
4. badger – badger

a) −0.895
b) 0.895
c) 0
d) 1

Task 12

Below is a plot of word embeddings in a two-dimensional vector space. Each vector is shown as a point in the vector space.



Embeddings for the word "bank"

Given the plot above and the text annotations, what model(s) could have been used to produce these embeddings? Mark all that apply.

- ☐ continuous bag-of-words
- ☐ skip-gram with negative sampling
- ☐ n-gram language model, e.g. bigram model
- ☐ masked language model, e.g. BERT
- ☐ generative language model, e.g. Gemma

Task 13

Given the following sentence and its tokenized representation as produced by a subword tokenizer, what is the fertility score of the tokenizer on this sentence?

sentence: *The cumulative tool response length reaches a preset threshold*

tokens: `The cum ##ulative tool response length reaches a pres ##et threshold`

Task 14

Which of these properties and descriptors apply to a BERT model? Mark all that apply.

- ☐ autoregressive
- ☐ encoder model
- ☐ decoder model
- ☐ masked language model
- ☐ generative language model
- ☐ instruction fine-tuned

Task 15

A large language model is typically trained in several steps. Below is an example from a dataset used for one of these steps.

input: *Why is the sky blue?*

choices:
A. *The sky appears blue because molecules in Earth's atmosphere scatter shorter wavelengths of sunlight more strongly than longer wavelengths.*
B. *It is because of Rayleigh scattering.*
C. *The sky is blue because of a phenomenon known as Rayleigh scattering.*

conciseness: B > C > A

helpfulness: A > C > B

What is the dataset from which the above example is taken most likely used for?

- ○ pre-training
- ○ instruction fine-tuning
- ○ preference alignment
- ○ few-shot prompting

Solutions begin on the next page!

Solutions

Task 1

- ○ To make the training process faster by reducing the amount of data the model sees.
- ● **To estimate how well the classifier generalizes to new, unseen data.**
- ○ To guarantee that the model achieves a higher accuracy.
- ○ To allow the model to learn from its mistakes during evaluation.

Task 2

$$\frac{50}{55 + 12 + 50}$$

Task 3

precision with respect to *"news"*	recall with respect to *"culture"*
$\frac{505}{505+0+1+10}$	$\frac{320}{1+4+320+45}$

Task 4

- ☐ $P(\text{cat} \mid \text{sleeps})$
- ☑ $P(\text{cat} \mid \text{My})$
- ☑ $P(\text{My} \mid \text{<BOS>})$
- ☐ $P(\text{<EOS>} \mid \text{cat sleeps})$
- ☑ $P(\text{<EOS>} \mid \text{sleeps})$
- ☐ $P(\text{sleeps} \mid \text{<EOS>})$
- ☐ $P(\text{My cat} \mid \text{sleeps})$
- ☐ $P(\text{cat} \mid \text{<BOS> My})$
- ☑ $P(\text{sleeps} \mid \text{cat})$

Task 5

P(*claims*)	P(*report* | *claims*)
$\frac{1550}{2000000}$	$\frac{29}{1550}$

Task 6

P(*report*)	P(*claims* | *report*)
$\frac{2672+1}{2000000+1\times40000}$	$\frac{190+1}{2672+1\times40000}$

Task 7

- [x] **The perplexity will always be a value ≥ 1.**
- [ ] The perplexity will always be lower than the entropy.
- [ ] A high probability score corresponds to a high perplexity value.
- [x] **Perplexity is an example for an intrinsic evaluation metric.**

Task 8

precision	recall	F1-score		
$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{2+3}$	*or*	$2 \times \frac{\frac{1}{2} \times \frac{1}{3}}{\frac{1}{2}+\frac{1}{3}}$

Task 9

token	tag
The	**O**
Seattle	**B-ORG**
Seahawks	**I-ORG**
are	**O**
an	**O**
American	**B-GPE**
football	**O**
team.	**O**

Task 10

Examples of accepted answers:

- **Words with similar distributions have similar meanings.**
- **The distributional hypothesis states that we can learn something about the meaning of words by looking at their context.**

Task 11

- **1.b)**
- **2.c)**
- **3.a)**
- **4.d)**

Task 12

- [ ] continuous bag-of-words
- [ ] skip-gram with negative sampling
- [ ] n-gram language model, e.g. bigram model
- [x] **masked language model, e.g. BERT**
- [x] **generative language model, e.g. Gemma**

Task 13

$$\frac{11}{9}$$

Task 14

- [ ] autoregressive
- [x] **encoder model**
- [ ] decoder model
- [x] **masked language model**
- [ ] generative language model
- [ ] instruction fine-tuned

Task 15

- ○ pre-training
- ○ instruction fine-tuning
- ● **preference alignment**
- ○ few-shot prompting