Questions from previous exams

1. Logistic regression

You are given a logistic regression model for three classes: *A*, *B* and *C*. Your current model parameters are $w = \{w_A, w_B, w_C\}$ where w_i is the weight vector for class *i*:

- $w_A = [1, 1.2, -2, 1.5, 1]$
- $w_B = [-2, 3, 1, 0, -2]$
- $w_C = [0, -3, 0, -2, 5]$

You are additionally given an example whose feature vector is $x = [0, 1, 0, 1, 1]$. Compute $p(i|x; w)$ for each of the classes *i*.

2. Softmax

A central formula in classification is based on the softmax function:

$$P(C_j|\vec{x}) = \frac{e^{\overrightarrow{w_j}\cdot\vec{x}}}{\sum_{i=1}^{k} e^{\overrightarrow{w_i}\cdot\vec{x}}}$$

Explain the formula, in particular

1. What does $P(C_i \mid \vec{x})$ express in a classification task?
2. What is C_i?
3. What is $\vec{w}_i$?
4. What is $\vec{x}$?
5. What is k?

(Observe that $\vec{x}$ can also be written as bold face $\mathbf{x}$ and similarly with $\vec{w}_i$.)

3(a) Approaches

In this course we have seen two ways of representing words as vectors based on their distribution in a corpus. The first is based on word-context matrices and the other is called "word embeddings". Describe shortly the main ideas of the two approaches. In particular, compare the two approaches with respect to the form of the vectors and how the vectors are derived.

3(b) Similarity

Assume the following simplified word vectors.

- girl - (1, 3)
- boy - (2,1)
- princess - (4, 5)

Using cosine for similarity, which word is closer to *girl*? Is it *boy* or *princess*? Explain how you find the answer.

3(c) Analogies

Word vectors may be used to study semantic properties of words. In particular, one may consider semantic analogies and, for example, ask what is related to *boy* as *princess* is to *girl*. We may symbolize the question as *girl:princess::boy:?*. A way to answer this question is called the *parallelogram method*. Using the example words, explain how the method works.

1(a) Part-of-speech tags

- What do we mean by “part-of-speech”?
- What does it mean that a text is tagged with part-of-speech?
- Why is it useful in natural language processing to part-of-speech-tag a text?

1(b) Hidden Markov models

- Given a sentence $w_1, w_2, \ldots, w_n$ and a tag set, what is the goal of a part-of-speech tagger?
- How does a hidden Markov model use the Bayes formula to solve this task?
- Which simplifying assumptions are made by the hidden Markov model?
- Give the formula for the hidden Markov model!

2(a) HMM tagger

Consider the two sentences

1. February made me shiver.
2. February gave me shiver.

and the two tag sequences

- a) NOUN VERB PRON VERB
- b) NOUN VERB PRON NOUN

It can be argued that the best tag sequence for (1) is (a) and for sentence (2) it is (b). Can a Hidden Markov Model (HMM) tagger be trained to assign tag sequence (a) to sentence (1) and tag sequence (b) to sentence (2)? State reasons for your answer. In case the answer is *yes*, which additional assumptions does the tagger have to fulfill?