Final exam: Statistical Natural Language Processing (SS 2021)

SfS / University of Tübingen

July 30, 2021

This exam has 5 questions on 6 pages (including this title page).

Question:	1	2	3	4	5	Total
Points:	6	6	10	10	8+2	42
Reached:						

Please read the information below carefully.

- This is a take-home exam. You are required to submit it electronically through Moodle at `https://moodle.zdv.uni-tuebingen.de/mod/assign/view.php?id=105676` **before 08:30 CEST on July 31, 2021**.
- You can consult any information source (books, notes, Internet resources). However you are **not allowed** to
  - discuss the solutions with each other, or get help from another person
  - ask questions about the current exam on Q&A sites (like Quora or stackexchange.com).
- None of the questions explicitly require the use of computers. However, you are free to use computers or calculators to compute or verify your answers.
- Submit your solutions via Moodle, as a **single PDF file**. Submissions in other formats, or submissions containing multiple documents will not be checked.
- Do not forget to write your full name and student id (Matrikelnr.) on the first page of your submission.
- Do not use the same page for multiple answers, each page should contain the (partial) solution of a single question. Naturally, you can use multiple pages for an answer.
- Indicate the question number fully and clearly on each page used for the answer of each question.
- You are recommended to typeset your answers using a computer. You can also use pen and paper for writing (part of) your answers, and scan and submit the electronic (PDF) file. In any case, make sure all your answers are readable.
- Answer the questions **briefly, and directly**. You may lose points if you write long answers with irrelevant information. Questions that ask you to "briefly explain" something require short (1-3 sentence) explanations, not a full page of text.
- You are required to submit a separate, one-page anti-plagiarism statement, which you can find on the Moodle page for this exam.
- We will hold an online session for your questions between 12:15–13:45 (usual lab hours). You are welcome to ask any clarification questions during this session.

Question 1 Correlation and regression

_____/6 p.

Answer the following questions based on the data presented in Figure 1.

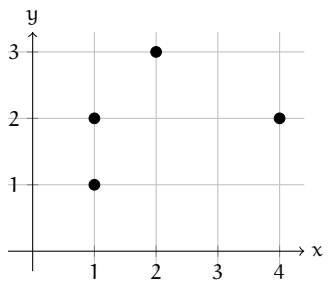

Figure 1: Data set for Question 1.

a. (1p) Calculate the correlation between the variables x and y.

b. (1p) Write down the equation for the least squares regression applied to this data set (predicting y from x).

c. (2p) Write down the objective function of L2 regularized regression (Ridge regression) *on this training set* in terms of w_0 (the intercept) and w_1 (the coefficient of x, or slope), and λ (the regularization strength).

d. (2p) Define a neural network architecture/model that can be used for the same problem (predicting y from x), but expected to perform better on this data set.[1]

[1] You are not required to define the weights of the model, Once trained properly (or maybe not so properly, since we want it to overfit), the neural model should, in principle, be able to learn weights that would result in a lower RMSE than the least squares regression on this training set.

Question 2 Linear classifiers

_____/6 p.

a. Fully define a *perceptron*, *logistic regression*, or *naive Bayes* classifier that classifies the data set in Table 1 perfectly. Make sure to assign values to all parameters of the model.

b. Given your model in (a), what is the precision, recall and F1 score of the model on the data set in Table 2.

c. If it is possible, calculate the cross-entropy of your classifier on the test set given in Table 2. If you cannot calculate the cross entropy, explain the reason briefly.

Table 1: The training set for Question 2.

x	y
2.11	+
−0.11	−
2.78	+
1.98	−
3.88	+
9.71	−
6.02	−
3.20	+

Table 2: The test set for Question 2

x	y
4.88	+
0.73	−
3.92	−
−2.12	+

Question 3 Sequence labeling

_____/10 p.

In a sequence labeling task, two alternative neural network models are considered. The first network, shown in Figure 2, is a simple feed-forward network that takes the current, previous and next items in the sequence as input and outputs the categorical label for the current item in the sequence.

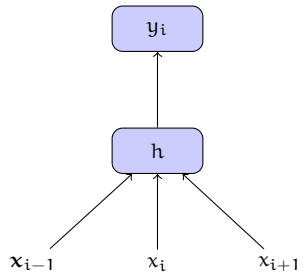

Figure 2: A feed-forward network used for sequence labeling. Bias/intercept terms are not shown in the figure.

The second model is a simple recurrent network (SRN), which is shown in Figure 3. In this model, a simple (without any gating mechanism) RNN model is trained to predict the label at each step, based on the current input and the internal representation of the RNN at the previous time step.

The input for the task is a variable sequence of scalar values (a single real number at each time step), and the output is a categorical variable with three possible values. The input is padded with special values indicating the beginning and end of sequence. Both networks use a single dense (fully connected) layer for the classification task.

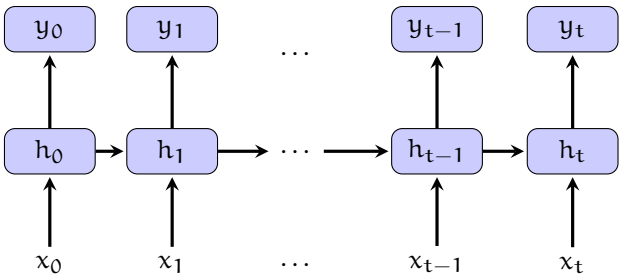

Figure 3: An unrolled representation of the sequence labeling RNN. Bias/intercept terms are not shown in the figure.

a. Calculate the number of parameters for each network in terms of number of hidden units h, and the number of time steps t.

b. After training both models on the same data set, you find that the classification performance of the feed-forward network is much better than the recurrent network. List two possible reasons why the RNN (which is particularly designed for working with sequences) may perform worse than the feed-forward network (based on the definitions of both networks above).

c. A colleague trains a simple hidden Markov model (HMM) for the task above, and finds out that it performs better than both networks. Again, this goes against your expectations that the neural networks to perform better. Suggest a reason why the networks above would not perform better than an HMM.

d. Suggest modifications to the RNN definition above (or the way to train it) that may resolve the issues you indicated in (b) and (c).

e. List the parameters of the HMM used for this task.[2]

[2] The task is rather different than the HMMs we covered in the class whose observed sequences were categorical variables. In this problem the observed sequence is a numeric variable. Assume that the conditional distribution of the observed variable given the latent category, $p(x|y)$, is a normal distribution.

Question 4 Classification

_____/10 p.

The following is a report of a machine learning system built for 'bot detection' on a social media platform. Given the posts of a social media user, the task of the model is to predict whether it is an automated system (a bot) or not. The report is supposed to inform the potential users of the system, and it should be complete enough to allow others to fully re-implement this bot detection system.

> The model we use for bot detection is a deep neural network with five layers of convolution/pooling followed by a classification layer. The input is presented to the network as an $t \times d$ matrix where t is the number of tokens in the input, and d is the embedding dimension. We use $t = 100$, padding documents shorter than 100 tokens with zero vectors, and truncating the longer documents. We use 50 convolutions of size 3×3 at each layer, and ReLu as the activation function at all convolutional layers. Each convolution layer is followed by a pooling layer. We use softmax activation at the output layer. The network is trained for 20 epochs with mean squared error loss using Adam optimizer.
>
> The model is trained on the posts of 10 000 users collected from the social media platform, and labeled as 'bot' or 'human' manually. We remove the stop words and punctuation from the input. We use pre-trained word2vec embeddings.
>
> Since the class distribution is not balanced, we report weighted precision, recall and F1 score in Table 3. As the scores indicate, the model shows near perfect classification performance in this task.

Table 3: Weighted precision, recall and F1-score of the bot detection system.

Precision	Recall	F1-score
0.95	0.96	0.98

List five wrong or questionable practices, or missing information in this report. For each problem you list, suggest a correction, or state what should be specified in the report.

Question 5 Unsupervised learning

_____/8 (+2) p.

A word-context matrix of 6 words is factorized using truncated SVD to the following matrices (the subscript indicates the dimensionality remains after truncation).

$$U_2 = \begin{bmatrix} -0.04 & 0.58 \\ -0.05 & 0.58 \\ -0.32 & 0.52 \\ -0.49 & -0.15 \\ -0.62 & -0.12 \\ -0.51 & -0.13 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 8.16 & 0. \\ 0. & 4.82 \end{bmatrix} V_2^T = \begin{bmatrix} -0.07 & -0.14 & -0.62 & -0.77 \\ 0.71 & 0.68 & -0.15 & -0.06 \end{bmatrix}$$

a. Based on this decomposition, compute the best approximation to the original data matrix.You are encouraged to use a computer to make the calculations required in this question (and the ones below).[3] You can use the following text to copy and paste the data as numpy arrays.

```
U = np.array([[-0.04, 0.58], [-0.05, 0.58], [-0.32, 0.52], [-0.49, -0.15], [-0.62, -0.12], [-0.51, -0.13]])
S = np.array([[8.16, 0.0], [0.0, 4.82]])
VT=np.array([[-0.07, -0.14, -0.62, -0.77], [0.71, 0.68, -0.15, -0.06]])
```

b. Describe a neural network architecture that could be used to obtain similar representations for words as the representations of the above SVD decomposition.

c. Assuming a two-way clustering of the words based on the dendrogram in Figure 4, calculate the silhouette score for w_3 using L1 distance as the distance metric.

d. Cluster the words into two clusters using k-means algorithm based on their dense representations of the SVD decomposition above. Start from two random centroids within the range of the data, and for each iteration of the k-means algorithm, list the centroid locations.

e. Given an unseen word has the original context vector of $(0, 2, 2, 2)$, what is this word's representation in the two-dimensional embedding space defined by this decomposition, and which cluster would it be assigned by the clustering solution in (d).

[3] You are only required to submit the answers. You are not required to submit/include the programs/scripts you use for the solutions. It is also perfectly fine if you use a calculator, manually calculate all necessary quantities. In any case, showing the formulas you use and the intermediate steps increase your chances of getting partial credit.

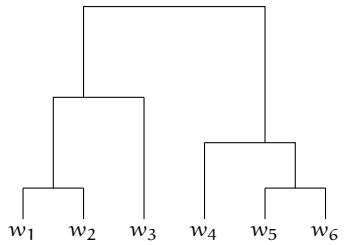

Figure 4: A dendrogram for clustering the words in the data set (used for SVD above). The labels at the leaves correspond to the index of the words (e.g., in the U_2 matrix above).