Final exam: Statistical Natural Language Processing (SS 2020)

SfS / University of Tübingen

July 24, 2020

This exam has 4 questions on **??** pages (including this title page).

Question:	1	2	3	4	Total
Points:	9	12	10	11	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=35645` **before 08:30 CEST on July 25, 2020**.
- You can consult with any information source (books, notes, Internet resources). However you are **not allowed** to
  - discuss the solutions with each other, or get help from any other person
  - ask questions about the current exam on Q&A sites (like Quora or stackexchange.com).
- 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 (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.
- Some questions require data analysis using a computer. The data files can be obtained from the Moodle page of the course at `https://moodle.zdv.uni-tuebingen.de/mod/assign/view.php?id=35645` or on the course web page at `https://snlp2020.github.io/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 Probability and information theory

_____/8 (+2) p.

$\{x_1, \ldots, x_n\}$ represent a set of n *uniformly distributed* random variables. Each variable x_i is characterized by a different discrete uniform distribution defined on interval $[a_i, b_i]$.[1] Based on this information, answer the following questions.

[1] The random variable x_i takes integer values between a_i and b_i, inclusive. Assume $a_i \neq b_i$ for any of the random variables.

a. Write down the mean (μ) and the covariance matrix (Σ) of the joint distribution of first three variables (x_1, x_2 and x_3) in terms of a_i and b_i (for $i \in \{1, 2, 3\}$).

b. Write down the expression for the joint probability distribution $P(x_1, \ldots, x_n)$ in terms of the individual probability distributions $P(x_1), \ldots, P(x_n)$. Explain your answer briefly.

c. How would you determine the random variable(s) with the largest entropy?

d. Is the distribution of sum of all variables ($x_1 + x_2 + \ldots + x_n$) a uniform distribution? Explain your answer briefly.

e. (2p) A friend reports that he trained and tested a regression model predicting the last variable x_n from the others ($x_1, \ldots x_{n-1}$) using a large sample drawn from their joint distribution. His results indicate that the model's root mean squared error is 0.01. Is the result expected? How would you explain this result?

f. (+2p) Calculate the cross entropy $H(p, q)$ (in bits) of the discrete uniform distribution q specified with parameters $a = 0, b = 1$ relative to the distribution p estimated using MLE from the sample given in Figure **??**. Briefly explain the interpretation of your result.

1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 1 0 0 1 0 0 0

Figure 1: A sample drawn from distribution p.

Question 2 Classification

_____/12 p.

A multi-layer perceptron (MLP) model is able to achieve perfect training set accuracy on a data set with two features (x_1, x_2) for predicting a binary variable (y). If the network is trained on the same data with each feature individually, the training set accuracy of the model is the same as an appropriate trivial baseline.

a. Name a reason (property of the data set) that may cause the result described (100 % accuracy with two features, baseline accuracy when trained with individual features).

b. Create a data set with four instances consistent with the result described. You can either list four feature vectors (four 2D vectors or a 4×2 matrix), or draw the data points on the 2D plane with distinct symbols for each class label.

c. Write down the equation[2] of a logistic regression model that classifies the data set you created in step (**??**) above correctly. The only variables in your equation should be the features (x_1, x_2) and the outcome variable (y, if you need to specify it), the weight values should be numeric constants.

[2] The equation for the discriminant surface is sufficient. But you can also provide the complete logistic regression equation for $p(y|\mathbf{x})$.

d. Can you create a data set with three instances with the same property as the one you created in (**??**)? Explain your answer briefly.[3]

[3] A formal proof is not necessary, but you should either show that it is possible by an example data set, or clearly explain the impossibility.

e. The MLP trained using both features has a 5-dimensional hidden layer. Calculate the number of parameters of the MLP described above. Briefly explain each term in your calculation.

f. List at least four hyperparameters of the MLP model.

Question 3 Dependency parsing

_____/10 p.

In a dependency parsing project, your task is to implement a scorer, a machine learning method that scores (head, dependent, label) triplets to be used in an MST parser. During parsing, your system is required to assign a score between 0 and 1 to any given triplet. The words in the triplet are given as word forms (no lemma, POS tag or other linguistic annotations). As well as the triplet to be scored, your system has access to the complete sequence of tokens (sentence).

Note that there is no single correct answer for any of the sub-questions. However, your answers to individual questions below should be consistent. Your choices do not have to be the best choices, but they have to be 'reasonable', and motivated well.

During training, your system has access to a treebank annotated according to Universal Dependencies. During test time you only have access to tokenized sentences without any morphosyntactic annotation.

a. Describe a machine learning model for solving this task. Make sure to explain the following clearly:

   - Features and feature representations you use.
   - The machine learning method you propose for the task.

   Your description should be detailed enough for a programmer to implement the model in a programming language, without any actual code. Briefly explain all choices you make.[4]

b. Describe the steps for tuning/training the model you defined. Make sure to clearly specify the data source you need to use, the tuning procedure and the evaluation method/metric. Your description should clearly specify the steps that need to be taken for getting the model ready for production.

c. Given the dependency tree in Figure **??**, list all training instances (e.g., dependency triplets) and the associated values to predict for the scorer you designed.[5]

d. Besides the treebank, you have a large unannotated corpus. Can you make use of the corpus for improving your scorer (and the parser)? Explain your answer briefly.

e. Can this parser be used with non-projective dependency trees? Briefly explain why, or why not.

[4] For example, 'I propose using a two layer ANN architecture, because . . . '

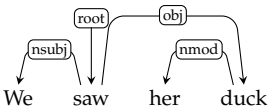

Figure 2: Dependency tree to be used for answering Question **??**c.

[5] Your answer may require different units for training instances (e.g., pairs of dependency triplets) depending on your modeling choices.

Question 4 Classifier evaluation

_____/10 p.

A social media company is interested in detecting offensive posts on their system. Since the volume of the posts makes it impossible to manually screen all posts, a machine learning model is planned for flagging (potentially) offensive posts, which are directed to people for final decision. The company is also considering a fully automated filtering solution if the method works well. Furthermore, the company is also interested in a classification of offensive posts as 'hate speech' (HS), 'cyber bullying' (CB), and other types of offensive statements (OTH).

A classifier for solving both problems jointly was developed by a third party. The classifier assigns a given post to one of the classes above (HS, CB, OTH) if the post is offensive with the given subtype, or to the class non-offensive (NON). You are provided with the predictions of the classifier, and the gold-standard labels annotated by humans. Your task is to evaluate the classifier, and write a **one-page**[6] report to help decision makers (e.g., company management) to decide adoption of the classifier in both tasks (catching any type of offensive post for screening, and classifying the offensive posts to one of the three sub categories).

[6] A standard one-page report is about 400 to 500 words. A slightly longer report is fine, if you feel you cannot fit some important points into one page. However, as noted on the cover page, you may lose points for an unnecessarily long report. Shorter but informative reports are definitely welcome.

Your report should include following information.

- Appropriate quantitative performance metrics
- A brief motivation for use of the particular metrics
- An assessment of how well the classifier performs for the purposes above, and its weaknesses.
- A clear statement (supported by quantitative measures) benefits and drawbacks of using this system for

filtering: the posts identified as offensive are rejected automatically, the posts identified as non-offensive are posted without further processing

flagging: the posts identified as offensive are forwarded to a human for further evaluation, the posts identified as non-offensive are posted without further processing

- Suggestions to improve the system for *filtering* or *flagging* as described above

The data for this question is available both on Moodle at `https://moodle.zdv.uni-tuebingen.de/mod/assign/view.php?id=35645` and on the course web page at `https://snlp2020.github.io/exam/`. The data is provided as a simple tab-separated text file where the first column lists the gold-standard labels (human annotations) and the second column lists the corresponding classifier predictions.