2020-2021 1-GS Methods in AI research (INFOMAIR)

Announcements

Dashboard

Staff Information

Course Information

Course Schedule

Course Content (lecture slides etc.)

Team Project

Questionnaire and submitting project deliverables

My Grades

Review Test Submission: Quiz Lectures 4 and 5.

User	Otto Mättas
Course	2020-2021 1-GS Methods in AI research (INFOMAIR)
Test	Quiz Lectures 4 and 5.
Started	9/16/20 7:27 AM
Submitted	9/16/20 7:51 AM
Due Date	9/16/20 4:00 PM
Status	Needs Grading
Attempt Score	Grade not available.
Time Elapsed	24 minutes
Results Displayed	All Answers, Submitted Answers, Correct Answers, Feedback, Incorrectly Answered Questions

Question 1

0 out of 10 points

You like to train a machine learning system to predict whether a book will be come a "bestseller". You've collected a large dataset, and for each each book you have the following information:

- The author: You have 1000 unique authors in your dataset
- Has the author written a bestseller before? Yes or no
- Genre: {Crime, Fantasy, Historical Fiction, Science Fiction, Thriller}
- The number of pages of the book

Each book is one instance in your dataset. You first need to represent each book as a vector before training your machine learning model.
Each book will be represented as a [7]-dimensional vector. (Fill in the correct number.)

Selected Answer: 3

Correct Answer: 1,007

Answer range +/- 0 (1007 - 1007)

Response Feedback: Incorrect

You would apply one hot encoding for both the author and the genre, resulting in 1000 + 5 features.
The two other features (has the author written a bestseller before and the number of pages) can be stored in one dimension each.
Each book would therefore be represented with vectors of 1007 dimensions.
An alternative would be to turn the number of pages in a categorical variable, especially when you expect that the relation with the number of pages and becoming a bestseller is non-linear. Had you given this answer during an exam, that would have been marked correct as well.

Question 2

0 out of 10 points

You have a dataset where each instance is represented by 100 features. However, a large fraction of these features are noisy and not useful signals for making the classifications. Which of these two classifiers do you think would perform better?

Selected Answer: k-Nearest Neighbors

Answers: Logistic Regression

k-Nearest Neighbors

I expect both to perform similarly

Response Feedback: Incorrect! Logistic regression, because it can give some features more/less weight, while nearest neighbors weights each feature equally.

Question 3

Needs Grading

Describe a (1) task for which a bag of words representation is probably sufficient (2) and another task for which it is not.

Selected Answer: 1) Email spam detection by comparing to a known bag of (spam) words.
2) Intention detection in any text.

Correct Answer: We'll return to this in the live session.

Response Feedback: [None Given]

Question 4

0 out of 10 points

Calculate the Jaccard Similarity between these two sentences:

Bob and John just went to the grocery store
Bob bought the book at the auction

Provide your answer with 3 decimals.

Selected Answer: 0.154

Correct Answer: 0.154

Answer range +/- 0 (0.154 - 0.154)

Response Feedback: Incorrect!

union = {Bob, and, John, just, went, to, the, grocery, store, bought, book, at, auction} = 13
intersection = {Bob, the} = 2
The Jaccard similarity is 2/13

Question 5

0 out of 10 points

Calculate the cosine similarity between these two vectors:

a = [2, 1, 2, 0, 0]
b = [1, 1, 0, 1, 1]

Selected Answer: 1

Correct Answer: 0.5

Answer range +/- 0 (0.5 - 0.5)

Response Feedback: Incorrect!

3/(3*2) = 0.5

Question 6

10 out of 10 points

Stemming increases/reduces the number of word types in a dataset

Selected Answer: Reduces

Answers:

Increases

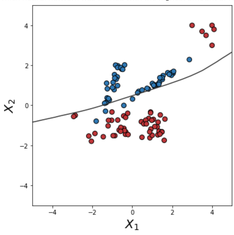
Reduces

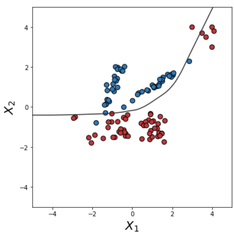
Response Feedback: Correct! With stemming different forms might be mapped to the same word type.

Question 7

10 out of 10 points

Here's a dataset on which a feed forward neural network was trained with two different values for regularization.
Which neural network was trained with the highest amount of regularization?





Selected Answer: Left

Answers: Left

Right

Response Feedback: Correct! The left one is less influenced by the red cluster at the top.

Question 8

Needs Grading

Are any aspects of the lecture material unclear, or do you have follow-up questions about this?
If I have your feedback in time AND if there is sufficient time to do so, I will try to address this during the live lecture that is associated with this question.

Leave this blank if you do not have any questions.

Selected Answer: Thanks!

Correct Answer: [None]

Response Feedback: [None Given]

Wednesday, November 4, 2020 10:34:45 PM CET

← OK