Version: 2021-11-04

Assignment 4: Sequence Learning and Maximum Likelihood Estimation

Machine Learning

Fall 2021

Learning Objectives

- Practice implementing generative models
- Spend some time on context and ethics around text classification

If this course were several semesters long, we'd love to offer you an opportunity to create from scratch your own generative models based on the probability learning you've done. And if that is something you're really excited about, we encourage you to pick that topic for your project! But because of time constraints, we're going to move from theory to implementation on generative models, such as Naïve Bayes.

1 Text Classification with Bag of Words

Next we'll be applying Naïve Bayes to the task of classifying text. You may have seen some elements of this before, but now you'll know how it works.

External Resource(s) (60 minutes)

This will be done in the Assignment 3 companion notebook.

2 99% Invisible Context on Enron Corpus (50 minutes)

Exercise 1

Please listen to this *99% Invisible* podcast episode (actually it's a *Brought to You By* episode, but we heard it through *99% Invisible*).

Exercise 2

Lawsuits and public records acts result in many emails being made public. In fact, some of Carrie's emails were made public due to a freedom of information act request. Technically, Olin as an institution has the ability to read any emails sent to or from and olin.edu address (though there is no evidence that this is ever done).

Please reflect. Do you consider your emails private or public? How would you feel if your emails were released? Are emails different than other information that might be used as a training set, such as your tweets, instagram posts, or discord comments?

3 Mini-Project: ML and text

Topics covered were to be - Identify writer - Distinguish gender of writer - distinguish between boss and subordinate