Machine Learning
Maximum Likelihood Estimation

What is machine learning?

Programming computers to learn.

Why machine learning?

- It may not be possible to encode the task in the form of rules.
- There may be too much knowledge to encode.
- Different users/different environments may require different behaviors.
- The environment may change over time.
- There may be relationships hidden in large sets of data that are hard to see for humans.

Variants of machine learning

- supervised learning
- unsupervised learning
- reinforcement learning

Authorship Attribution

Austen:

Emma Woodhouse, handsome, clever, and rich, with a comfortable home and happy disposition, seemed to unite some of the best blessings of existence; and had lived nearly twenty-one years in the world with very little to distress or vex her.

She was the youngest of the two daughters of a most affectionate, indulgent father; and had, in consequence of her sister's marriage, been mistress of his house from a very early period. Her mother had died too long ago for her to have more than an indistinct remembrance of her caresses; and her place had been supplied by an excellent woman as governess, who had fallen little short of a mother in affection.

Authorship Attribution

Chesterton:

The flying ship of Professor Lucifer sang through the skies like a silver arrow; the bleak white steel of it, gleaming in the bleak blue emptiness of the evening. That it was far above the earth was no expression for it; to the two men in it, it seemed to be far above the stars. The professor had himself invented the flying machine, and had also invented nearly everything in it. Every sort of tool or apparatus had, in consequence, to the full, that fantastic and distorted look which belongs to the miracles of science. For the world of science and evolution is far more nameless and elusive and like a dream than the world of poetry and religion; since in the latter images and ideas remain themselves eternally, while it is the whole idea of evolution that identities melt into each other as they do in a nightmare.

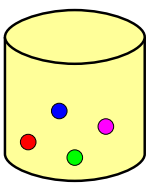
Authorship Attribution

Who wrote these passages? Austen or Chesterton?

1) The family of Dashwood had long been settled in Sussex. Their estate was large, and their residence was at Norland Park, in the centre of their property, where, for many generations, they had lived in so respectable a manner as to engage the general good opinion of their surrounding acquaintance. The late owner of this estate was a single man, who lived to a very advanced age, and who for many years of his life, had a constant companion and housekeeper in his sister.

2) The suburb of Saffron Park lay on the sunset side of London, as red and ragged as a cloud of sunset. It was built of a bright brick throughout; its sky-line was fantastic, and even its ground plan was wild. It had been the outburst of a speculative builder, faintly tinged with art, who called its architecture sometimes Elizabethan and sometimes Queen Anne, apparently under the impression that the two sovereigns were identical. It was described with some justice as an artistic colony, though it never in any definable way produced any art.

Excursion: Basic Probability

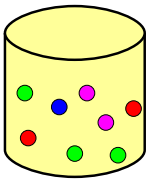


If we draw a ball without looking, what is the probability of drawing a green one?

A red one?

A red one or a green one?

Excursion: Basic Probability

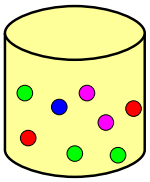


If we draw a ball without looking, what is the probability of drawing a green one?

A red one?

A red one or a green one?

Excursion: Basic Probability



If we draw a ball without looking, what is the probability of drawing first a red one and then a green one?

A green one and then a red one?

A red one and a green one in either order?

The probability of words

Given

- a collection of texts by author A, and
- a word.

Can we say anything about how likely it is that this word is produced/used by author A?

Estimating unigram probabilities

Idea: use a text collection to estimate how likely an author is to use different words.

Text collection by author A:

- # of words: 299,294
- # of times 'the' occurs: 8531
- # of times 'laughed' occurs: 19
- # of times 'house' occurs: 189

Estimating unigram probabilities

Idea: use a text collection to estimate how likely an author is to use different words.

Text collection by author A:

- # of words: 299,294
- # of times 'the' occurs: 8531
- # of times 'laughed' occurs: 19
- # of times 'house' occurs: 189

P(unigram) = Count(unigram) / # of words

Using unigrams – likelihood of a sentence

Given what we know about unigrams, how likely is it that author A wrote the following sentence?

"His attachment to them all increased."

- # of words: 299,294
- frequency of 'his': 1808
- frequency of 'attachment': 73
- frequency of 'to': 8046
- frequency of 'them': 706
- frequency of 'all': 1378
- frequency of 'increased': 20

Bigrams – looking at two words at a time

What is the probability of author A using the word 'attachment' after the word 'his'?

- frequency of 'his': 1808
- frequency of 'his attachment': 9

P(attachment | his) = Count(his attachment) / Count(his)
= 9 / 1808

Using bigrams – likelihood of a sentence

Given what we know about bigrams, how likely is it that author A wrote the following sentence?

"His attachment to them all increased."

- # of words: 299,294
- frequency of 'his': 1808
- frequency of 'his attachment': 9
- frequency of 'attachment to': 17
- frequency of 'to them': 46
- frequency of 'them all': 41
- frequency of 'all increased': 1

- frequency of 'attachment': 73
- frequency of 'to': 8046
- frequency of 'them': 706
- frequency of 'all': 1378

Author attribution

Given

·a collection of texts by author A

·a collection of texts by author B

·a text passage

Was this text passage written by author A or B?

Computing with probabilities

- They get very small/close to 0.
- The product of very small numbers if an even smaller number.
- Python cannot represent arbitrarily small numbers.

Computing with probabilities

- They get very small/close to 0.
- The product of very small numbers if an even smaller number.
- Python cannot represent arbitrarily small numbers.

➔ Use the logarithm of probabilities instead.

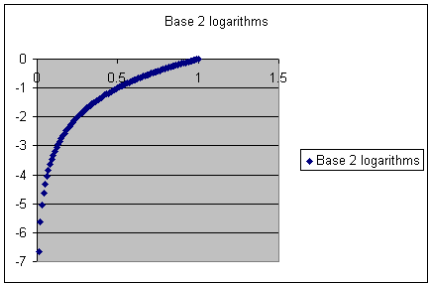
Logarithm

$\log_2 8 = ?$

$2^? = 8$

$\log (a \cdot b) = \log(a) + \log(b)$

Base 2 logarithms



Log Probabilities

P(his attachment...increased) =

P(his) · P(attachment | his) · P(to | attachment) · ... · P(increased | all)

log(P(his attachment...increased)) =

log(P(his)) + log(P(attachment | his)) + log(P(to | attachment)) + ... + log(P(increased | all))

Taking a step back

What subtasks to we need to solve to write a program that decides which one of two authors wrote a given passage.

Given:

- several files containing texts by author A
- several files containing texts by author B
- a file containing text passages that need to be classified

Designing the main algorithm

Given:

- v_size: an integer; the size of our vocabulary
- training_A: a list of files containing texts by author A
- training_B: a list of files containing texts by author B
- test_passage: a string
- count_frequencies: a function that takes a list of files and returns 1) a dictionary with unigram frequencies, 2) a dictionary with bigram frequencies, 3) the total number of words in those files
- string_prob: a function that takes a string, dictionaries with unigram and bigram frequencies and the total number of words for one author as well as the vocabulary size and computes the string's probability.

Wanted:

- An algorithm that builds unigram and bigram models for two authors and then checks whether the given passage was more likely written by author A or by author B.

Programming

- Download text_classification.py and development.zip and start working on the homework assignment.