Machine Learning

Lecture 3

Justin Pearson[1]

2020

[1] http://user.it.uu.se/~justin/Teaching/MachineLearning/index.html

Today's plan — Classification

- Revision basic probability
- Bayes Theorem
- Naive Bayes Classification

Probability

What does the probability of an event tell us?

- The probability of a fair coin toss coming up heads is 0.5.
- The probability of getting four of a kind in poker is 0.000240.
- The probability of nuclear war is 0.0039^2.

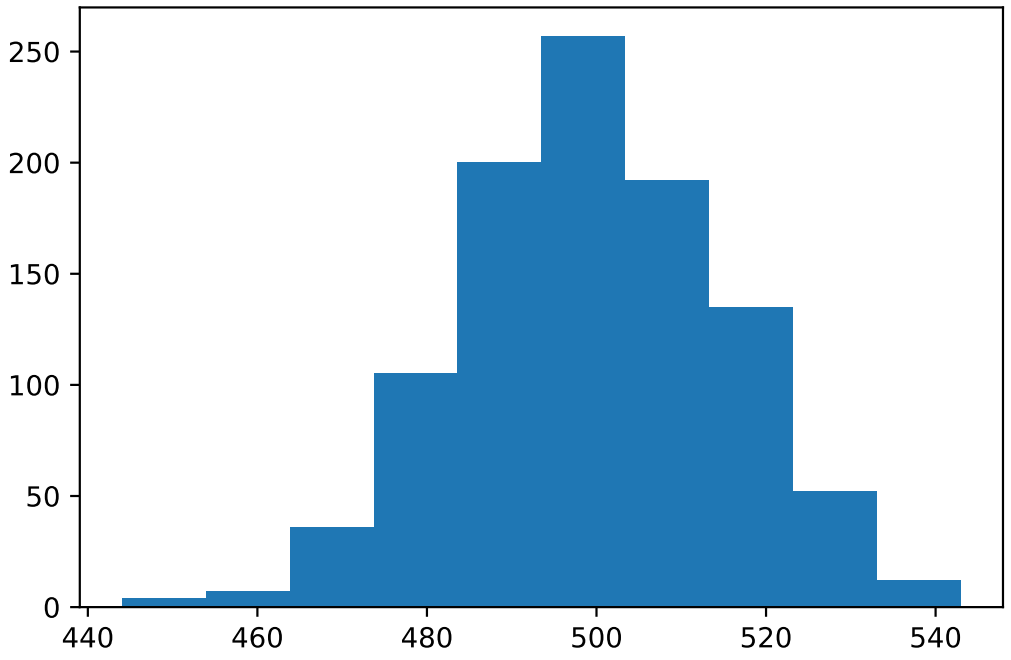
The first two statements tell us something about the frequency events occur, while it is not clear what the last statement actually tells us.

[2] https://marginalrevolution.com/marginalrevolution/2019/07/what-is-the-probability-of-a-nuclear-war.html

Probability — Subjectivist or Frequentist

There is a lot of debate:

Frequentist If you repeat an experiment enough times then the probability tells you something about the number of outcomes. If I toss a coin 1000 times then I expect around 500 heads and 500 tails.



Probability — Subjectivist or Frequentist

There is a lot of debate:

Subjectivist Some how the probability measures your subjective belief in a statement. The axioms of probability gives you constitutions for logically consistent beliefs.

Probability for machine learning

- Build classifiers that estimate the probability of something falling into a class. Is my mail Spam or not. If the probability is high enough then classify the email as spam.

Probability — Experiments, sample spaces and events

Mathematically probability is a way of modelling the world.

- An *experiment* produces exactly one out of several possible *outcomes*
- The set of all possible outcomes is called the *sample space*
- A subset of the sample space is called an *event*.

Probability — Experiments, sample spaces and events

Consider four six-sided dice[3]

For an experiment you roll all four dice.

[3]Picture taken from https://commons.wikimedia.org/wiki/File:6sided_dice.jpghttps://commons.wikimedia.org/wiki/File:6sided_dice.jpg

Probability — Experiments, sample spaces and events

Sample Space The set $\{(r, g, b, p) \mid r, g, b, p \in \{1, \ldots 6\}\}$ the tuple representing the values of the four dice.

Events There are a large number of events. The event that sum of all 4 dice is 36 is the set $\{(r, g, b, p) \mid r, g, b, p \in \{1, \ldots 6\}\}\{(r, g, b, p) \mid r, g, b, p \in \{1, \ldots 6\}, r + g + b + p = 36\}$

Probability — Experiments, sample spaces and events

A probability model assigns probabilities to events. Given a sample space S, a probability distribution is a mapping from events (subsets of S) to the interval $[0, \ldots, 1]$ such that

- For any event A then $P(A) \geq 0$
- $P(S) = 1$.
- For any two disjoint sets A and B, $P(A \cup B) = P(A) + P(B)$
- If your sample space is infinite then for any infinite sequence of disjoint sets $A_1, \ldots, A_2, \ldots$

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots$$

Most of the time you can think of events, but sometimes you have to worry about the sample space.

Probability — Experiments, sample spaces and events

For the 4 dice example if our dice are fair then

- For any (r, g, b, p) with $r, g, b, p \in \{1, \ldots 6\}$ the probability of the event $P(\{(r, g, b, p)\})$ is $\frac{1}{6^4}$.
- By the axioms of probability the probability of any event, that is subset of S for the 4 dice example follows from the probability of $P(\{(r, g, b, p)\})$ by taking unions.

Probability — Experiments, sample spaces and events

Non-discrete example.

- Experiment, measure somebody's BMI.
- The is a continuous variable. The sample space is all positive real numbers.
- An Experiment $P(15 \leq x \leq 20)$ the probability that a value is between 15 and 20.

Continuous probability distributions are modelled probability density functions and cumulative distribution functions.

Dependent and Independent Events

Given two events A and B what is the probability of $P(A \cap B)$

- Since we have a probability model on the sample space, in theory we can just calculate it. For the 4 dice example we just intersect the sets of events.

We want a nice formula.

Independent Events

- Suppose I toss a coin twice what is the probability that I get two heads?
- $P(\text{first toss is a head}) \times P(\text{second toss is a head}) \times P) = \frac{1}{2} \times \frac{1}{2}$.

The two coin tosses are independent so we can multiply probabilities.

Dependent Events — Conditional Probability

We need a new quantity

$$P(A \mid B)$$

The Probability that A occurs given that B has happened.

Dependent Events — Conditional Probability

There are lots of ways to motivate and define it, but we can take as an axiom that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

The Probability that A occurs given that B has happened.

Conditional Probability

It is common to re-arrange the formula

$$P(A \cap B) = P(A|B)P(B)$$

If A and B are independent then $P(A|B) = P(A)$ which gives $P(A \cap B) = P(A)P(B)$.

Bayes' Theorem

First

$$P(A \cap B) = P(B \cap A)$$

From

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

and

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \quad \Rightarrow \quad P(B \cap A) = P(A \cap B) = P(B|A)P(A)$$

Gives

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Bayes' Theorem

This rather innocent formula

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

is the bases for classification, a whole school of statistics and a tool to correct inconsistent reasoning.

Bayes' Theorem

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

Using terminology from statistics

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

Bayes' Theorem — Useful identity

$$P(A) = P(A|B)P(B) + P(A|\overline{B})P(\overline{B})$$

For a set of events B_i

$$P(A) = \sum_i (P(A|B_i)P(B_i) + P(A|\overline{B}_i)P(\overline{B}_i))$$

Note that for an event B the notation $\overline{B}$ is the complement event when B does not happen that is $S \setminus B$.

Bayes' Theorem — Example

Suppose we are testing for cancer a population, but the probability of cancer is quite low 0.01. We have a test that is not perfect

- True positive, probability test says there is cancer given there is cancer $P(T \mid C) = 0.90$.
- False positive, probability that the tests says that cancer given that there is cancer given that there is no cancer $P(T \mid \overline{C}) = 0.10$.

Given that our test is positive what is the probability that there is cancer. So

$$P(C \mid T) = \frac{P(T \mid C)P(C)}{P(T)} = \frac{0.90 \times 0.01}{P(T)}$$

Bayes' Theorem — Example

Given that our test is positive what is the probability that there is cancer. So

$$P(C \mid T) = \frac{P(T \mid C)P(C)}{P(T)} = \frac{0.99 \times 0.01}{P(T)}$$

We still need to work out the probability that our test is positive.

- There is cancer and the test is positive,
  $P(C)P(T|C) = 0.01 \times 0.90 = 0.009$
- There is no cancer, but the test is still positive,
  $P(\overline{C})P(T|\overline{C}) = (1 - P(C)) * P(T|\overline{C}) = 0.99 \times 0.10 = .099$

So the probability that the test says there is cancer regardless if the patient has cancer is $0.009 + 0.099 = 0.108$. So

$$P(C \mid T) = \frac{0.90 \times 0.01}{0.108} \approx 0.08$$

This is much lower than you might think. The false negatives are contributing quite a lot.

Bayes' Theorem — Example

Another way of thinking about this. Suppose you have 1000 patients.

- If the probability of cancer is 0.01 then 10 patients are expected to have cancer.
- Of the 10 patients that have cancer the test will report positive on 9 cases.
- Of the 990 patients who do not have cancer the test will report positive on $0.1 * 990 = 99$ patients.
- If you have a positive test case then it is one of the $9 + 99$ patients and the probability that one of them has cancer is $\frac{9}{9+99} \approx 0.08$.

Spam Detection — Naive Bayes

We are going to build a classifier for emails. It is to decide if an email is spam or not.

- The training set must be a set of emails that are classified as spam or ham (non-spam).

We have a number of common words that appear emails and we use the training set to estimate the probability that a particular word appears in a spam or non spam email.

Spam Detection — Naive Bayes

Given email you want to decide if it is spam or not. One way of doing look at words that appear in the mail. Suppose that we only consider if the email contains the word "Prince" or not.
If we receive an email that contains the word "Prince" what is the probability that it is spam or not. Using Bayes' theorem

$$P(\text{Spam} \mid \text{Prince}) = \frac{P(\text{Prince} \mid \text{Spam})P(\text{Spam})}{P(\text{Prince})}$$

Spam Detection — Naive Bayes

So for our spam detector to work we need

- $P(\mathrm{Spam})$ the probability an email is Spam.
- $P(\mathrm{Prince} \mid \mathrm{Spam})$ in all our Spam email the probability that the word "Prince" occurs.
- $P(\mathrm{Prince})$ the probability that the word "Prince' occurs in any email. We could calculate this with

$$P(\mathrm{Prince}) = P(\mathrm{Prince}|\mathrm{Spam})P(\mathrm{Spam}) + P(\mathrm{Prince}|\overline{\mathrm{Spam}})P(\overline{\mathrm{Spam}})$$

All these quantities can be estimated from your training set by counting the occurrences of words.

Spam Detection — Naive Bayes

Given

$$P(\text{Spam} \mid \text{Prince}) = \frac{P(\text{Prince} \mid \text{Spam})P(\text{Spam})}{P(\text{Prince})}$$

If we look at

$$P(\overline{\text{Spam}} \mid \text{Prince}) = \frac{P(\text{Prince} \mid \overline{\text{Spam}})P(\overline{\text{Spam}})}{P(\text{Prince})}$$

The term $P(\text{Prince})$ appears for both spam and non spam and is effectively a normalising constant.
Computing

$$P(\text{Prince} \mid \text{Spam})P(\text{Spam}) \quad \text{and} \quad P(\text{Prince} \mid \overline{\text{Spam}})P(\overline{\text{Spam}})$$

and looking at which is bigger tells us if the email is spam or not spam.

Spam Detection — Naive Bayes

A single word is not going to do so well. We need to consider multiple words.

- Three test words “Prince”, “Viagra” , “Linear”.
- If I receive a message that contains “Prince” and “Linear” but not “Viagra” what is the probability that it is spam?

Spam Detection — Naive Bayes

Using Bayes' theorem

$$P(\text{Spam} \mid \text{Prince}, \overline{\text{Viagra}}, \text{Linear}) \propto P(\text{Prince}, \overline{\text{Viagra}}, \text{Linear} \mid \text{Spam})P(\text{Spa}$$

So the question is how do we compute

$$P(\text{Prince}, \overline{\text{Viagra}}, \text{Linear} \mid \text{Spam})$$

With one word we could just compute the occurrences of the words in various messages.

Spam Detection — Naive Bayes

An unrealistic assumption is to assume that the probability a word appears in a message is independent of the probability that a different word appears in a message.
Thus in our case $P(\mathrm{Prince}, \overline{\mathrm{Viagra}}, \mathrm{Linear} \mid \mathrm{Spam})$ equals

$$P(\mathrm{Prince} \mid \mathrm{Spam})P(\overline{\mathrm{Viagra}} \mid \mathrm{Spam})P(\mathrm{Linear} \mid \mathrm{Spam})$$

There are lots of technical reasons why this naive assumption is not so naive.

Naive Bayes — Continuous Variables

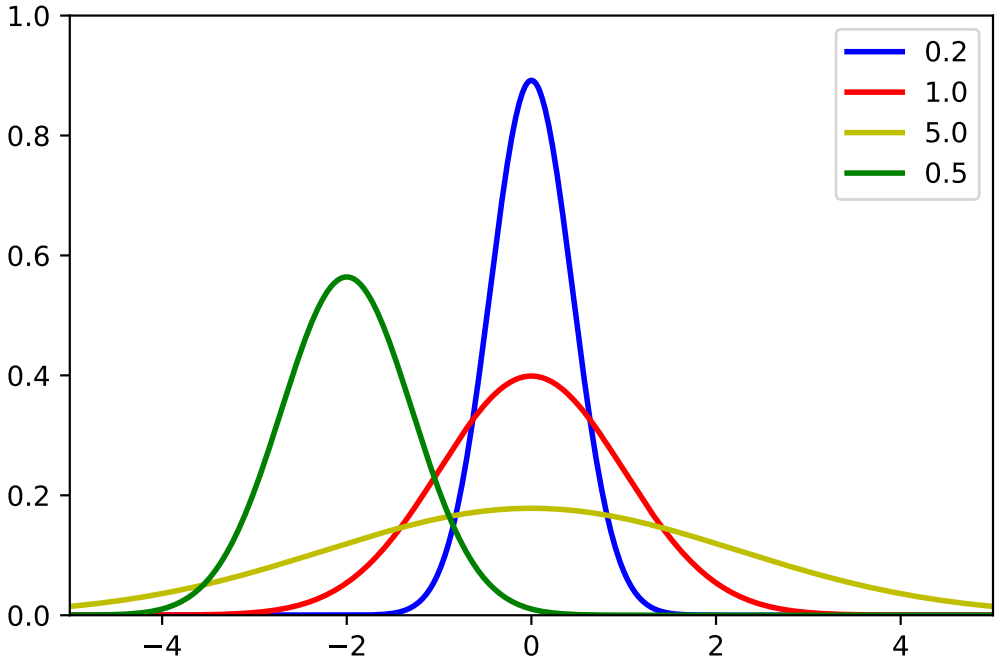
Suppose your data set has a continuous variable x we want to compute the probability that the date belongs to some class given the variable has value a. Again Bayes' theorem applies

$$P(C \mid x = a) \propto P(x = a \mid C)P(C)$$

So the question is how do we compute $P(x = a \mid C)$ and how do we define the classes?

Classes defined by Gaussian distributions

A Gaussian distribution is a good model of lots of types of data.



A normal distribution is defined by a mean μ and a variance σ

$$P(x = a) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(a-\mu)^2}{2\sigma^2}}$$

Classes for Continuous Variables

Each class could be defined by a different normal distribution with a different mean and variance. So let our class C have mean and variance μ_c, σ_c then

$$P(x = a \mid C) = \frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(a-\mu_c)^2}{2\sigma_c^2}}$$

$$P(C \mid x = a) \propto \left(P(C \mid x = a) = \left(\frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-\frac{(a-\mu_c)^2}{2\sigma_c^2}} \right) \right)$$

You will explore this in the lab.

Extra Material if we have time

- Zero probabilities – Zero count
- Very small numbers, taking logs.

Zero counts — Pseudocounts

In our spam filter we have to keep track of how many times words appear in our training set. Suppose our test words contain the word "zaphod", but none of training spam or non-spam emails contain the word spam. This means that

- We will estimate that $P(\text{zaphod} \mid \text{Spam})$ is zero.
- If "zaphod" appears in an email when computing

  $$P(\text{Spam} \mid \ldots \text{zaphod} \ldots) \propto P(\text{Spam})(\cdots \times P(\text{zaphod} \mid \text{Spam} \cdots \times)$$

  The $P(\text{zaphod} \mid \text{Spam})$ term forces the product to be 0. Even if lots of other spam words appear in the email we will not classify it as spam.

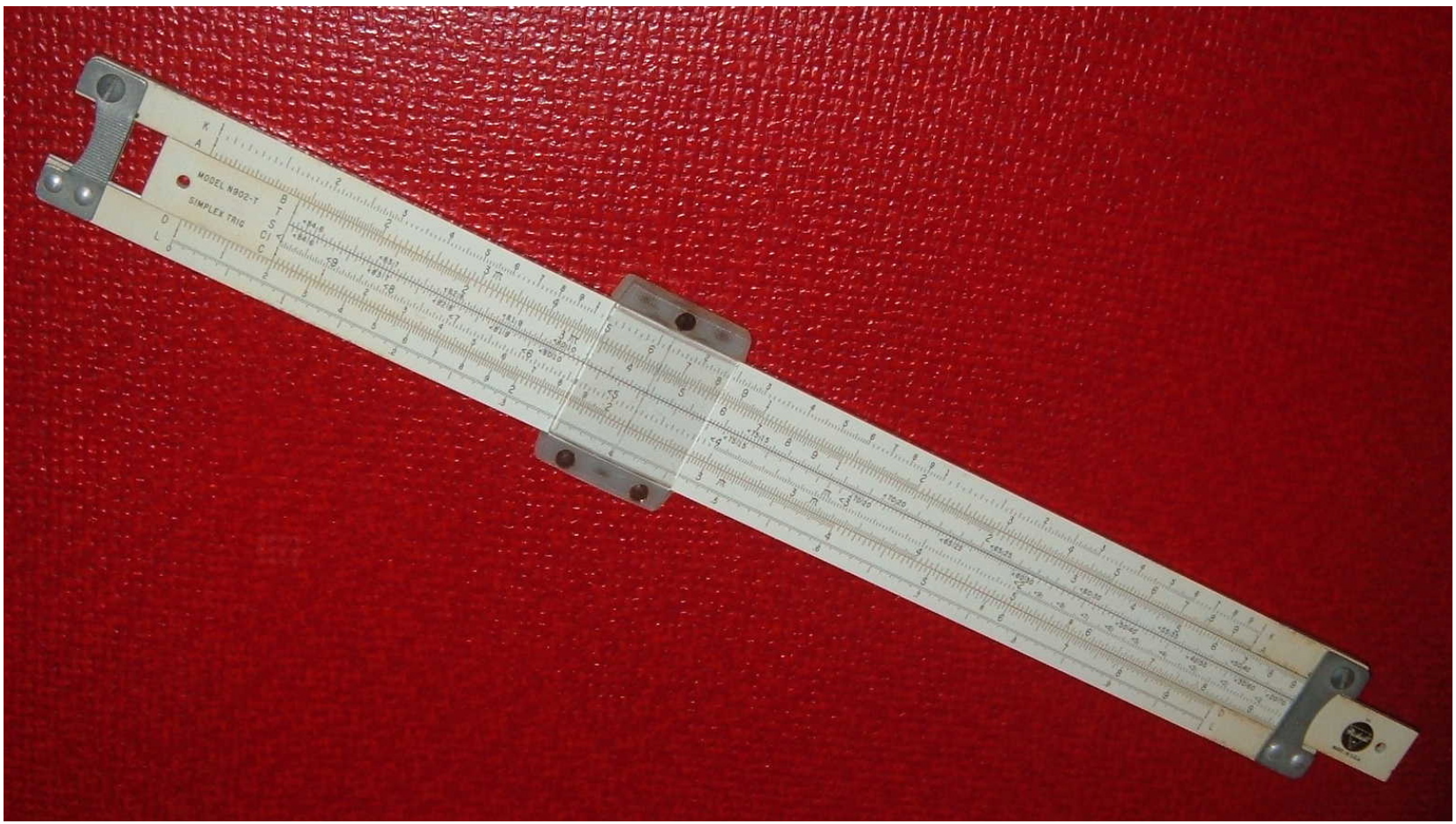
A common solution is to add 1 to every count. This stops the probabilities being 0. Although we might not estimate the probability correctly, we will not be too far off.

Taking Logarithms[4]

The most amazing identity in mathematics:

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

This formed the bases of a number of devices to make calculation easier.



[4] Picture from https://commons.wikimedia.org/wiki/File:Sliderule.PickettN902T.agr.jpg

Small numbers – Floating point

- Read "What every computer scientist should know about floating-point arithmetic"[5] by David Goldberg.
- Short story if you multiply together lots of small numbers errors can creep in.
- If your bank uses floating point numbers then maybe you should get a new bank.

[5] https://dl.acm.org/doi/10.1145/103162.103163

Working with Logs

Transform expression such as:

$$P(C)\Pi_i P(x_i \mid C)$$

into logarithms

$$\log\left(P(C)\Pi_i P(x_i \mid C)\right)$$

which is equal to

$$\log(P(C)) + \sum_i \log(P(x_i \mid C)$$