GENERAL ASSEMBLY

NAIVE BAYES

Mason Gallo, Data Scientist

AGENDA

- Motivating example for today's class
- Conditional probability refresher
- Naive bayes implementation

OBJECTIVES

- Understand the formal definition of Naive Bayes
- Define the advantages and disadvantages of NB
- Implement NB in Python

MOTIVATING EXAMPLE: SMS SPAM DATASET

HOW CAN WE DETECT SMS SPAM?

HOW CAN WE DETECT SMS SPAM?

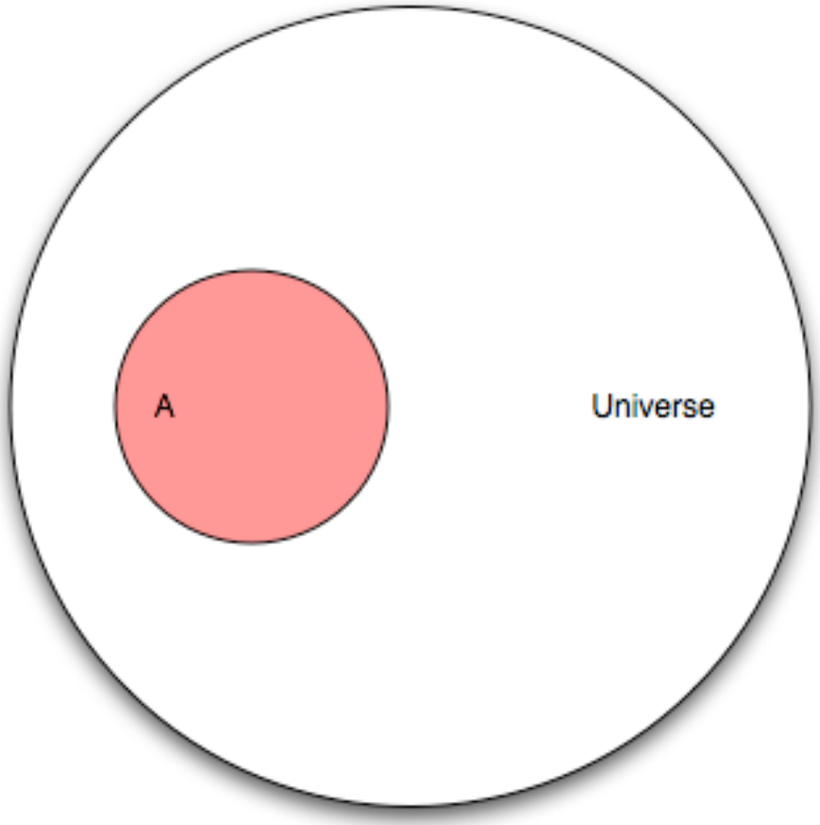
WE WILL LEARN HOW TO SIFT THROUGH THOUSANDS OF SMS FOR SPAM/HAM

DISCUSSION POINTS

- False positives and false negatives for spam
- His statistical approach to spam filtering
- How good was his prediction of the spam of the future?

NAIVE BAYES PROBABILITY REFRESHER

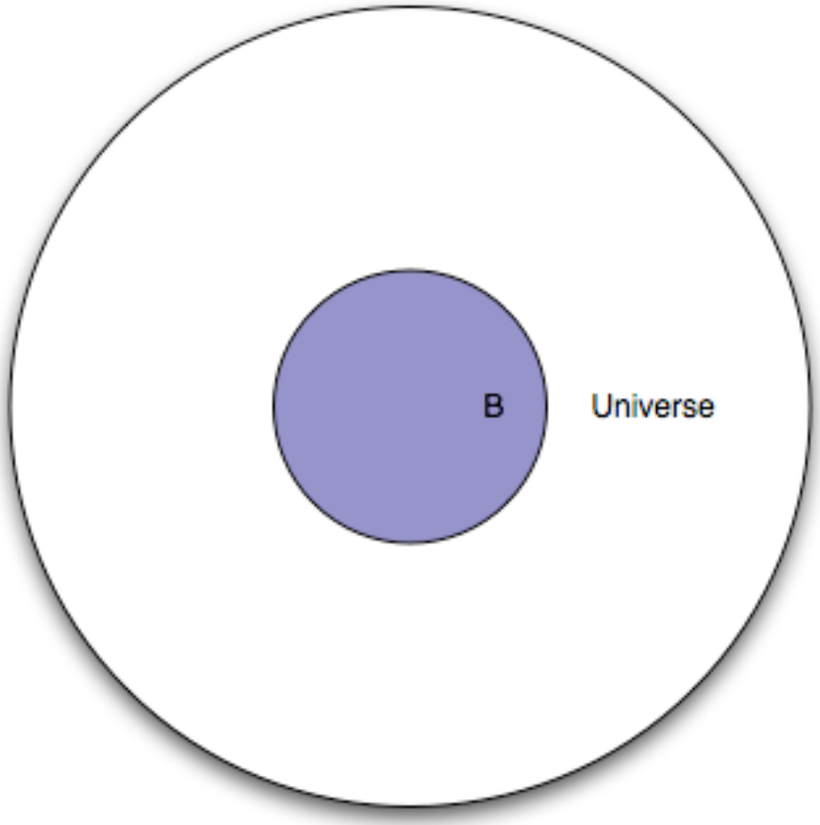
PROBABILITY



Let's now pretend that our universe involves a research study on humans. Event "A" is people in that study who have cancer.

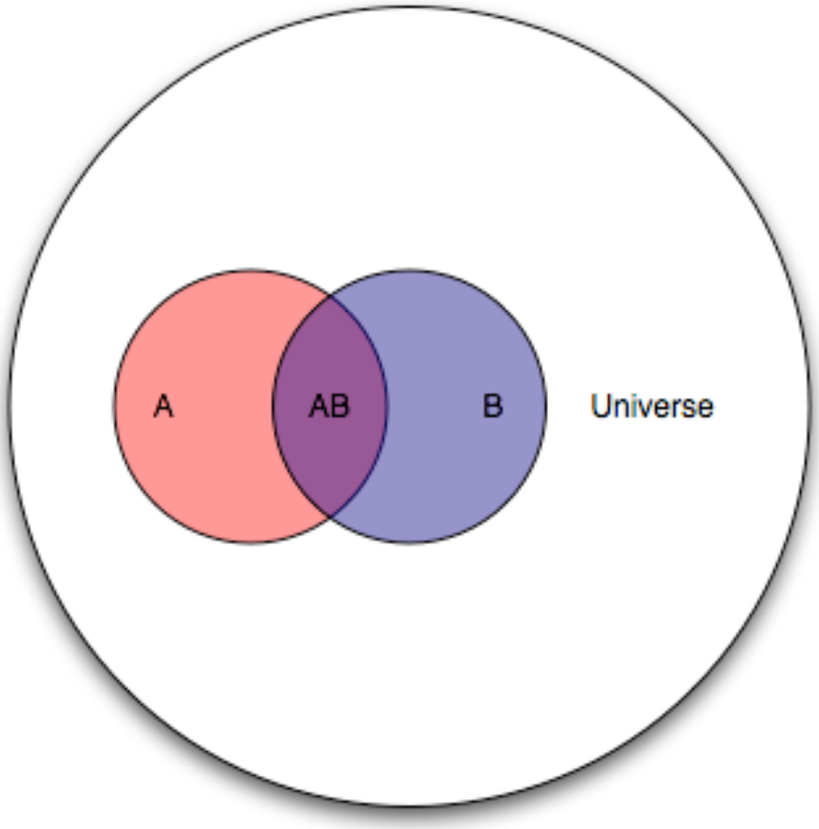
Q: If our study has 100 people and "A" has 25 people, what is the **probability** *of A?*
A: P(A) = 25/100

Q: What is the max probability of any event?
A: 1



This represents the same set of people, except everyone in the study is given a test. Event "B" is everyone in the study for whom the test is positive.

Q: What portion of the diagram represents the subset of people with a negative test?
A: The white area between the smaller circle and the larger circle.



Because "A" and "B" are events from the same study, we can show them together.

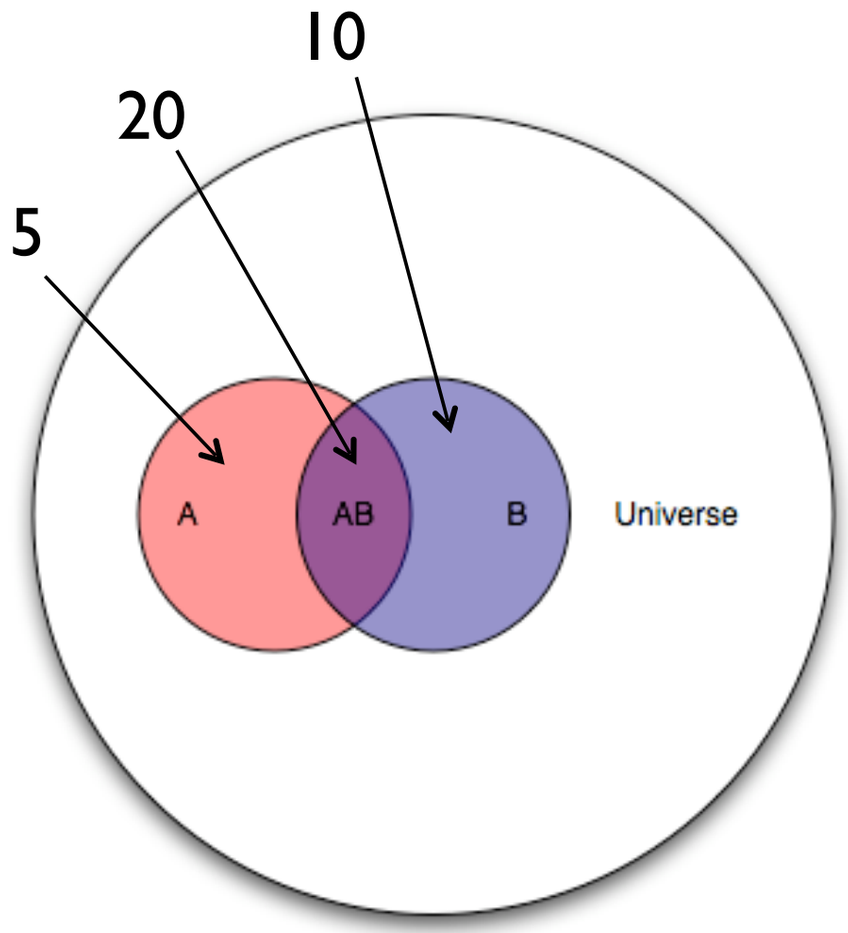
Q: How would you describe the "cancer status" and "test status" of people in each area of the diagram?

A: Pink: cancer, negative test

Purple: cancer, positive test

Blue: no cancer, positive test

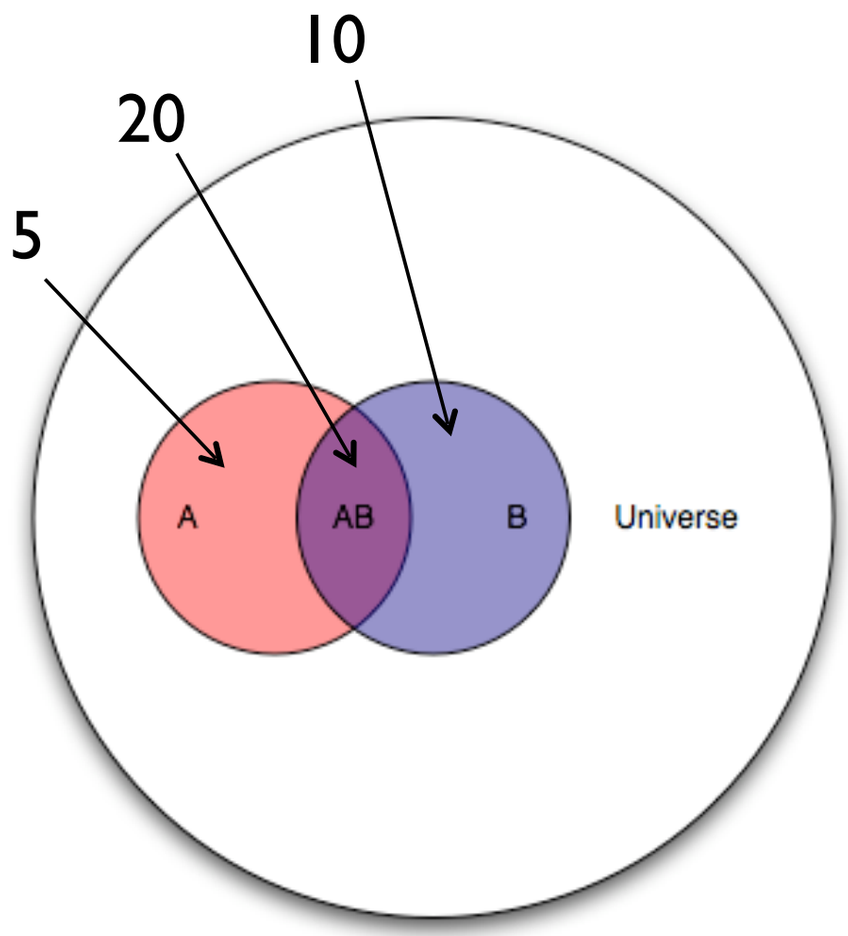
White: no cancer, negative test



The purple section is known as the **intersection** *of A and B, denoted as P(AB).*

Thinking of this test as a classifier for predicting cancer, draw the confusion matrix.

n=100	**Predicted: NO**	**Predicted: YES**
Actual: NO	65	10
Actual: YES	5	20

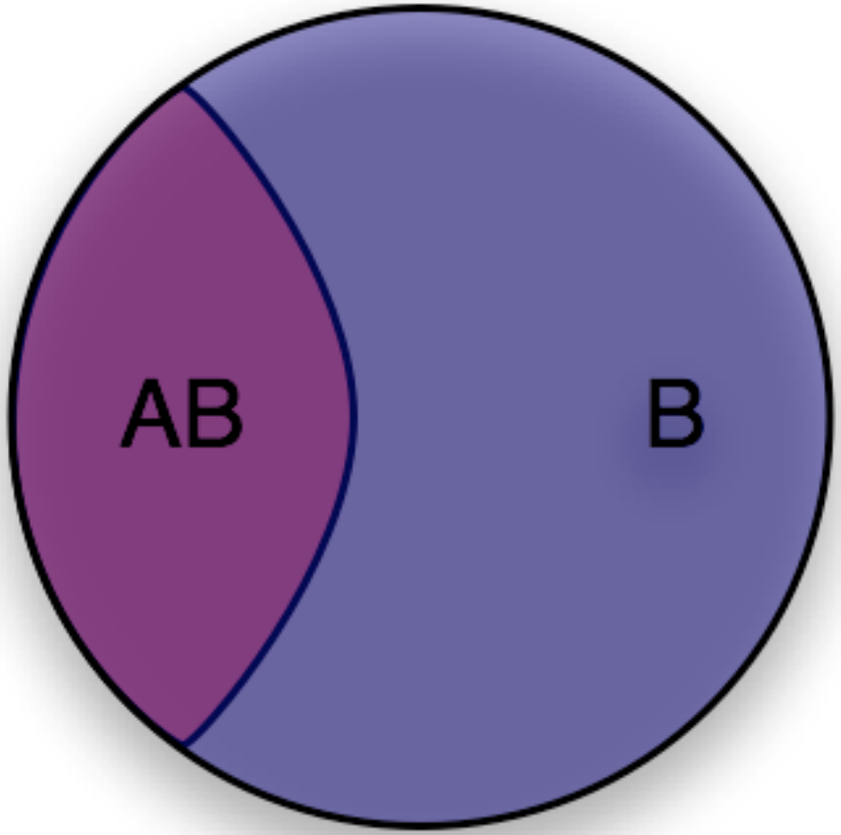


Q: Let's pick an arbitrary person from this study. If you were told their test result was positive, what is the probability they actually have cancer?

A: 20/30

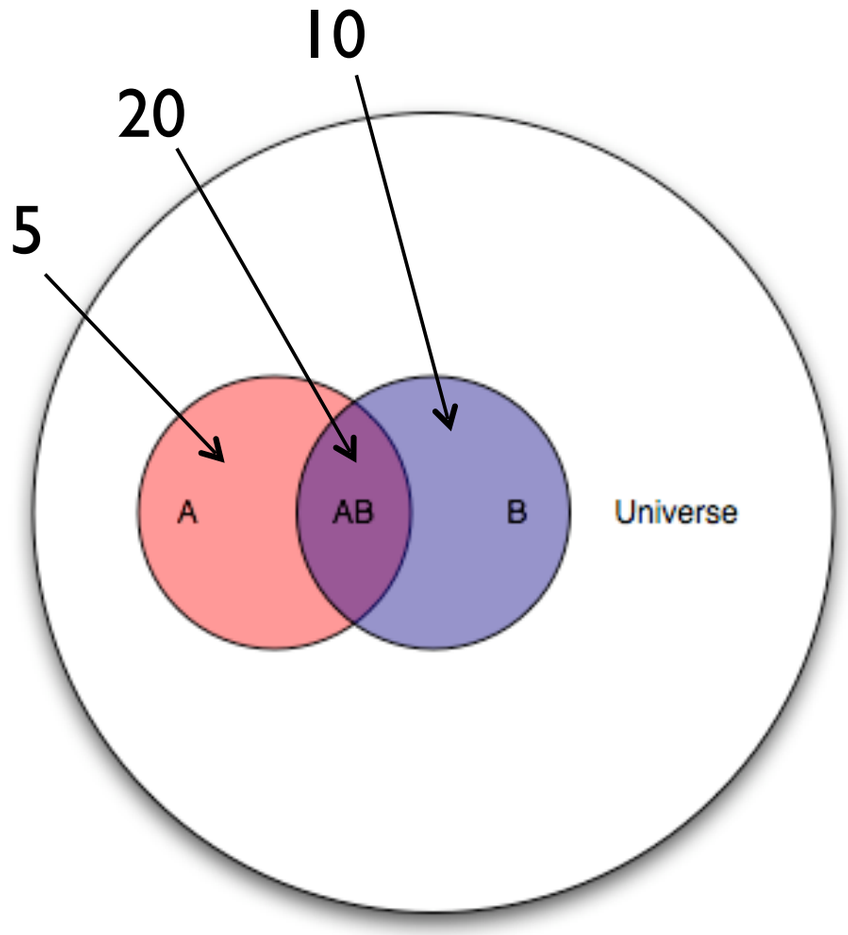
This is the **conditional probability of A given B**, *denoted as P(A|B).*

$P(A|B) = P(AB) / P(B) = (20/100) / (30/100)$



You can think of conditional probability as "changing the relevant universe." P(A|B) is a way of saying "Given that my entire universe is now B, what is the probability of A?"

This is also known as **transforming the sample space**.



Q: Let's pick another arbitrary person from this study. If you were told they have cancer, what is the probability they had a positive test result?

A: P(B|A) = P(AB) / P(A) = 20/25

BAYES' THEOREM

This result is called **Bayes' Theorem**

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

It means you can swap conditional probabilities

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\;P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM – TERMINOLOGY

Each term in this relationship has a name, and each plays a distinct role in any probability calculation (including ours).

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

BAYES' THEOREM – TERMINOLOGY

This term is the **posterior probability** *of A. It's the probability of A <u>after</u> the conditional data is taken into account.*

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

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\;P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM – TERMINOLOGY

This term is the **posterior probability** *of A. It's the probability of A <u>after</u> the conditional data is taken into account.*

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

The goal of any Bayesian computation is to find ("learn") the posterior distribution of a particular variable.

BAYES' THEOREM – TERMINOLOGY

This term is the **prior probability** *of A. It's the probability of A <u>before</u> any conditional data is taken into account.*

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

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\;P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM – TERMINOLOGY

This term is the **prior probability** *of A. It's the probability of A before any conditional data is taken into account.*

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

The value of the prior is often observed from general knowledge, the actual data, or even some desired scale or distribution.

BAYES' THEOREM – TERMINOLOGY

This term is the **likelihood** *function. This one swaps the conditional probabilities: it's the probability of your condition B, given A*

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

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\;P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM – TERMINOLOGY

This term is the **likelihood** *function. This one swaps the conditional probabilities: it's the probability of your condition B, given A*

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

The value of the likelihood function is observed from the actual data.

BAYES' THEOREM – TERMINOLOGY

This term is a **normalization constant**. *It doesn't depend on A, and is generally ignored while optimizing for maximum probabilities.*

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

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\;P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM – TERMINOLOGY

This term is a **normalization constant**. *It doesn't depend on A, and is generally ignored while optimizing for maximum probabilities.*

> *For example, while running through countries to assess their weather and movie business to find the most likely one, the chance of "rain somewhere" is not relevant.*

In a movie it's raining. What's the chance the movie is shot in Holland?

$$= \frac{P(\text{raining in Holland})\ P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

BAYES' THEOREM

Many machine learning techniques use Bayesian statistics to estimate the parameters of their model

BAYES' THEOREM

Many machine learning techniques use Bayesian statistics to estimate the parameters of their model

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\ P(\beta)}{P(\text{data})}$$

BAYES' THEOREM

Many machine learning techniques use Bayesian statistics to estimate the parameters of their model

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\ P(\beta)}{P(\text{data})}$$

Coefficients of regression
Class labels of samples
Student proficiency and question difficulty

BAYES' THEOREM

Starting out with a prior belief of the parameters β *...*

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\, P(\beta)}{P(\text{data})}$$

What are reasonable coefficients?
What are common class labels?
How are student proficiencies generally distributed?

BAYES' THEOREM

... and updating the likelihood as new data comes in.

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\, P(\beta)}{P(\text{data})}$$

Given these parameters, are my data reasonable?
Given these proficiencies and difficulties, how likely are these seen student responses?

BAYES' THEOREM

Now you see why the normalization constant is generally ignored.

$$P(\beta|\text{data}) = \frac{P(\text{data}|\beta)\ P(\beta)}{P(\text{data})}$$

How likely is this data anyway?

BAYES' THEOREM – TERMINOLOGY

The idea of Bayesian inference, then, is to **update our beliefs** *about the distribution of A using the data ("evidence") at our disposal*

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\ P(\beta)}{P(\text{data})}$$

BAYES' THEOREM – MAXIMUM LIKELIHOOD ESTIMATOR

The **maximum likelihood estimator (MLE)** *finds the parameters that make the data most likely*

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\, P(\beta)}{P(\text{data})}$$

BAYES' THEOREM – MAXIMUM A POSTERIORI ESTIMATE

The **maximum a posteriori estimate (MAP)** *finds the parameters that are most likely, given the data and the prior*

$$P(\beta \mid \text{data}) = \frac{P(\text{data} \mid \beta)\ P(\beta)}{P(\text{data})}$$

NAIVE BAYES EXAMPLE NOTEBOOK

THE TERMS NO ONE EXPLAINS

HOW WE GROUP WORDS IN OUR DATA

N-GRAM STANDS FOR GROUPS OF N-WORD COMBOS

unigram (1-gram):

a	swimmer	likes	swimming	thus	he	swims

bigram (2-gram):

a swimmer	swimmer likes	likes swimming	swimming thus	...

trigram (3-gram):

a swimmer likes	swimmer likes swimming	likes swimming thus	...

NAIVE BAYES ASSUMES DISTRIBUTIONS FOR FEATURES

THREE MAJOR TYPES OF NAIVE BAYES:

- Multinomial: assumes features are counts

ex: feature indicates how many times ‘hello’ appears in a text

- Bernoulli (binomial): assumes features are 0/1 indicating presence of feature

ex: feature indicates whether or not ‘hello’ appears in a text

- Gaussian (normal): assumes features are normally-distributed

ex: feature indicates how much a person weighs

WHY IS IT CALLED NAIVE BAYES?!

NB NAIVELY ASSUMES ALL OF THE FEATURES ARE INDEPENDENT

- This means if someone searches for “Chicago Bulls”, you’ll get the same answer as “Bulls Chicago”

- In other words, we assume the order of the words doesn’t matter!

EVALUATING THE STRENGTHS AND WEAKNESSES OF NB

ADVANTAGES

- Works well with low amount of training data / high amount of features
- Fast to train and classify
- If assumptions are met, surprisingly accurate

DISADVANTAGES

- Can’t learn interactions between features
- Assumes independence of features
- Assumes discrete features or a particular distribution

LET'S CODE!