Bayesian Statistics

Data Science Immersive

FLATIRON SCHOOL

Conditional Probability

- Agenda today…
  - Review independent probability
  - Learn Dependent Probability
    - Theorems
    - Examples
  - Bayes’ Theorem & Bayesian Statistics
    - Derivation
    - Examples
  - Naive bayes classification and NLP

After today, you'll be able to...

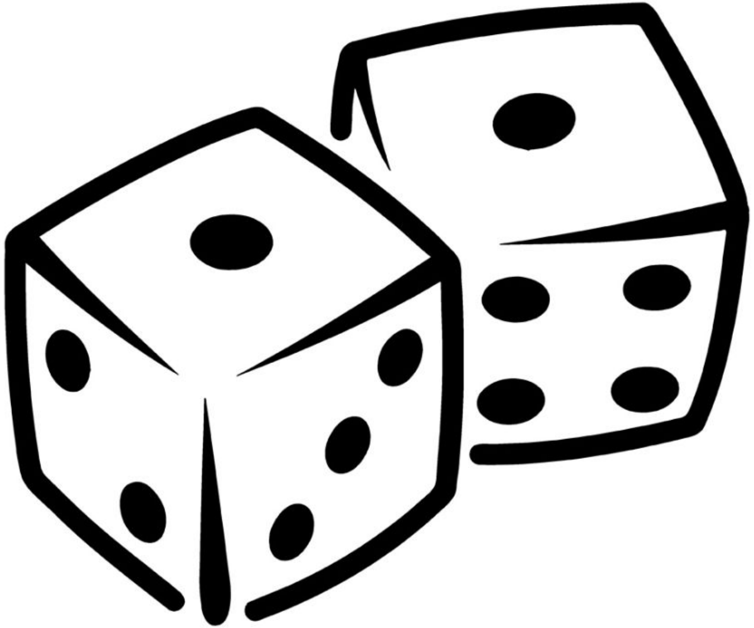
- Understand and explain the difference between independent and dependent events
- Gain an intuitive understanding of why conditional probability is necessary
- Calculate and compute conditional probabilities in given context
- Calculate conditional probabilities using Bayes' theorem
- Explain the difference between frequentist framework and Bayesian

Review

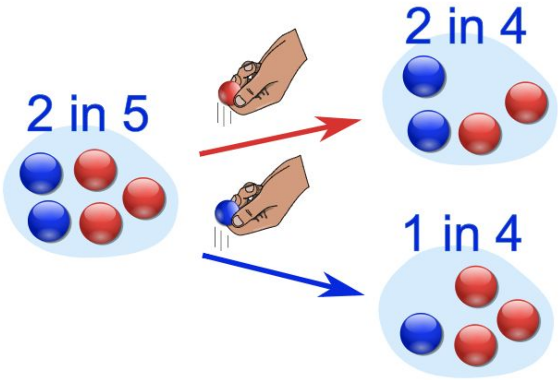
- Event
  - An **event** is the outcome of a random experiment
- Sample space
  - A **sample space** is a collection of every single possible outcome in a trial
- Independent probability is calculated by event divided by all the possible events in the sample space
  - The occurrence of one event does not affect, or dependent on the outcome of another

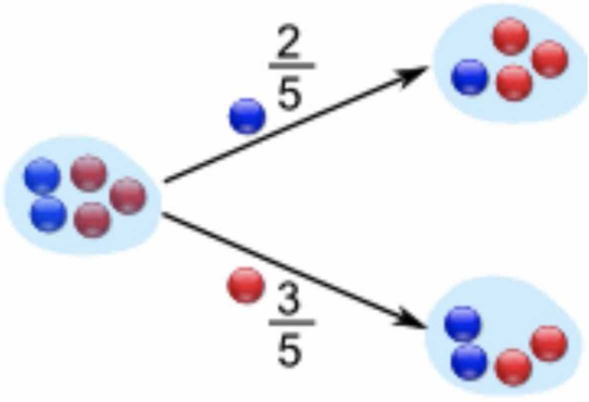
Review

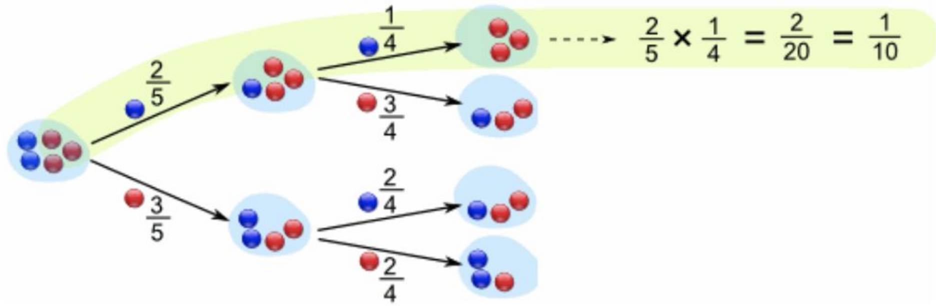
- Independent probability of an event is calculated as P(A) divided by **all** possible events.
- Some statistical distributions make such assumptions about events
  - Poisson distribution
  - Binomial distribution



Conditional Probability

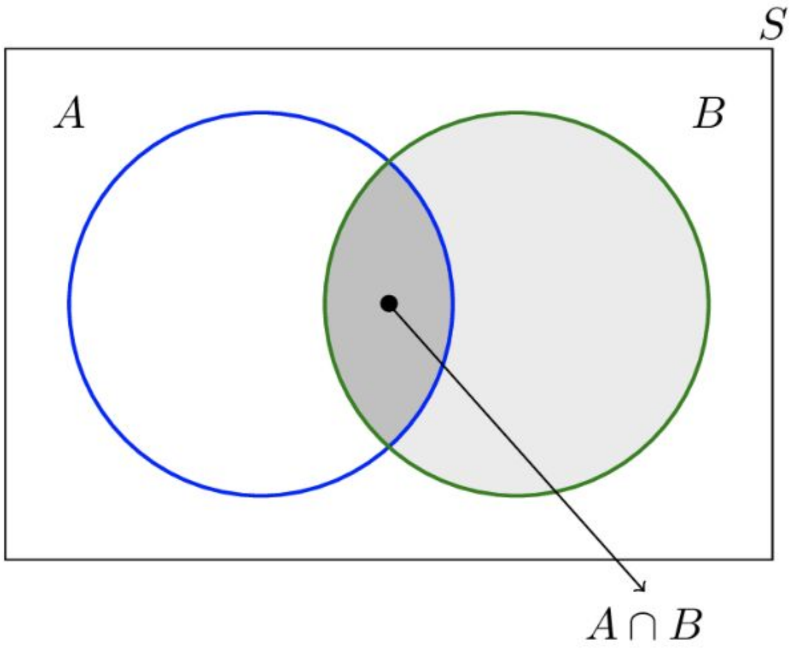






Conditional Probability

- Conditional Probability emerges in the examination of experiments where a result of a trial may influence the results of the upcoming trials. For example:
  - The probability of drawing an Ace given already drew an Ace
  - The probability of being in a good mood given the weather is nice



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

Example 1

The percentage of adults who are men and alcoholic is 2.25%. What is the probability of being an alcoholic given that someone is a male?

Example 2

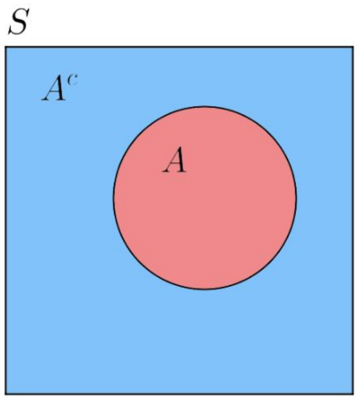
You are going to visit your distant cousins who recently had two children. You are told that at least one of them is a girl. What is the probability of both of them being girls?

Example 3

You are going to visit your distant cousins who recently had two children. You are told that the older one is a girl. What is the probability of both of them being girls?

Theorems of Conditional Probability

1. P(A') + P(A) = 1



Sample Space S, event A, and complement A^c

2. The Product Rule

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

The product rule is useful when the conditional prob is easy to compute but the intersection is not

Bayes’ Theorem

- Bayes’ Theorem underlies the foundation of Bayesian Inference, an incredibly powerful way in which statistics are probability are computed
- It is derived from conditional probability

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

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

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

Bayes’ Theorem

P(A) is called the **prior**; this is the probability of our hypothesis without any additional prior information. It could also be a belief we have prior to seeing the data.

P(B) is called the **marginal likelihood**; this is the total probability of observing the evidence. In many applications of Bayes Rule, this usually serves as normalization constant, which we will explain in further detail.

P(B|A) is called the **likelihood**; this is the probability of observing the new evidence, given our initial hypothesis.

P(A|B) is called the **posterior**; this is what we are trying to estimate.

Bayes' Theorem

Why Bayes'?

- Bayes' theorem allows us to accommodate degrees of belief into the equation, which accounts for uncertainty
- Has practical implication in natural language processing--i.g. Naive bayes for classification or topic modeling, such as Latent Dirichlet Allocation
- Stochastic methods for parameter estimation, i.g. Markov Chain Monte Carlo
- Network analysis or graph theory
- And many more applications in philosophy, cognitive sciences, even legal systems

Bayes' Theorem - Practice 1

If $P(A) = \frac{1}{2}$ and $P(B) = \frac{1}{2}$ and $P(B|A) = 1/3$, find:

a. $P(A \text{ and } B)$

b. $P(A \text{ or } B)$

c. $P(A|B)$

Example 1

Assume that the probability of having lung cancer is 2%, and the probability of being a smoker is 15%. Empirically, we know that 20% of the people who have lung cancer are smokers. What is the probability of having lung cancer given you are a smoker?

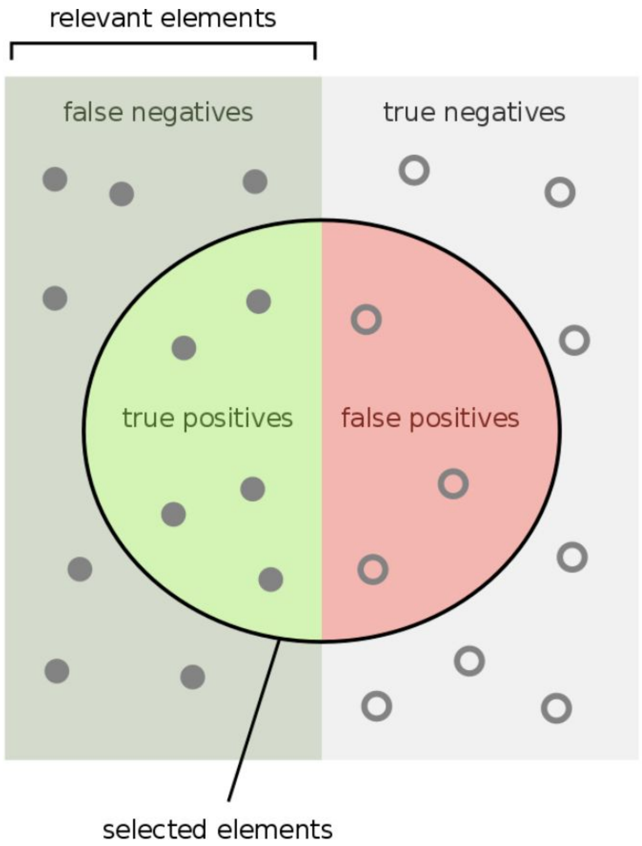
Example 2 - False Positives

- You tested positive for a disease with 5% false positive rate
- The probability of getting a positive test when you have the disease is 90%
- We know in the population 1% of people have this disease
- What is the probability of having this diseases given that you tested positive?

Example 2 Continued - Bayesian updating

What is the probability of you having the disease given that you tested positive twice?

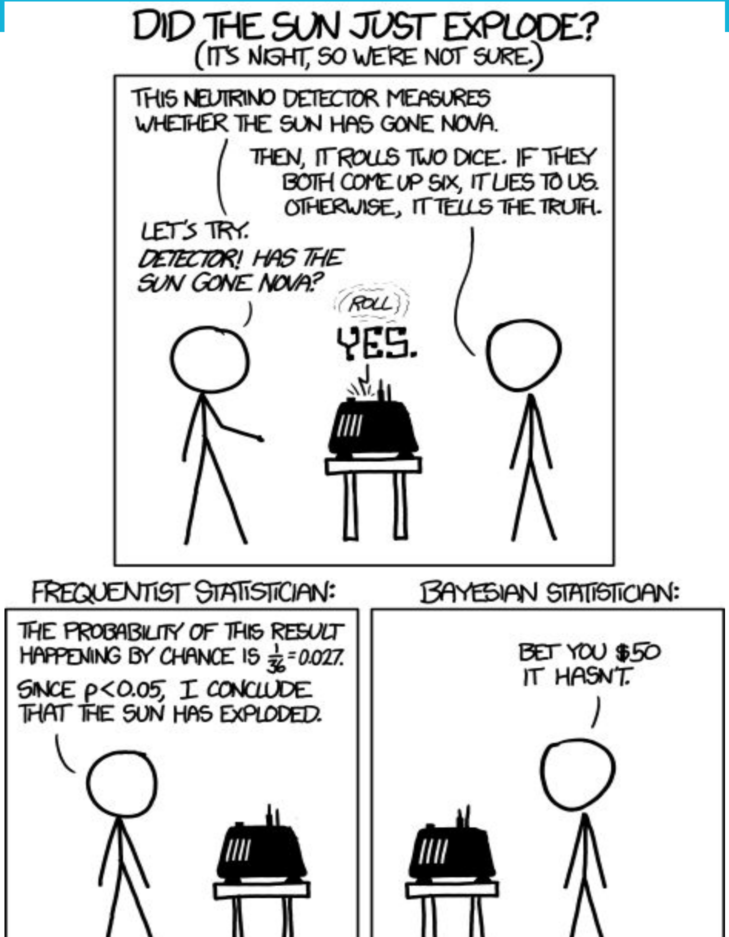
Sensitivity vs. Specificity



Frequentist vs. Bayesian

- Frequentist statisticians rely on the imaginary sampling of an infinite population and derive a probability value that summarizes the result of the experiment
  - Inference made by a frequentist statistician only depends on the frequency of events, or samples observed
- Bayesian statisticians not only rely on actual evidence observed, but also on beliefs

Frequentist vs. Bayesian



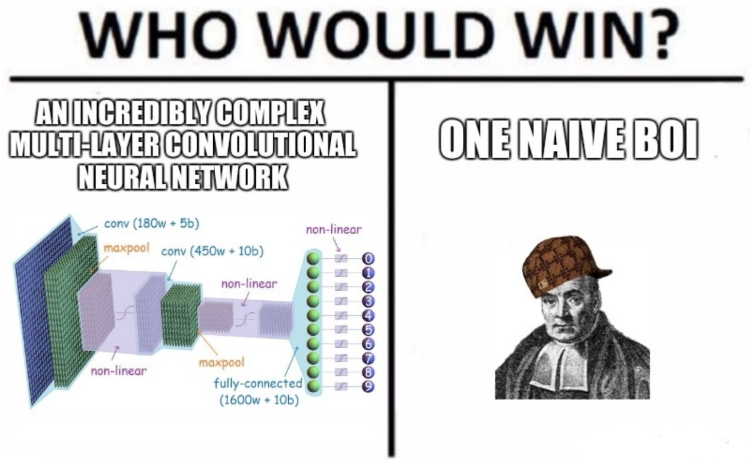
Naive Bayes

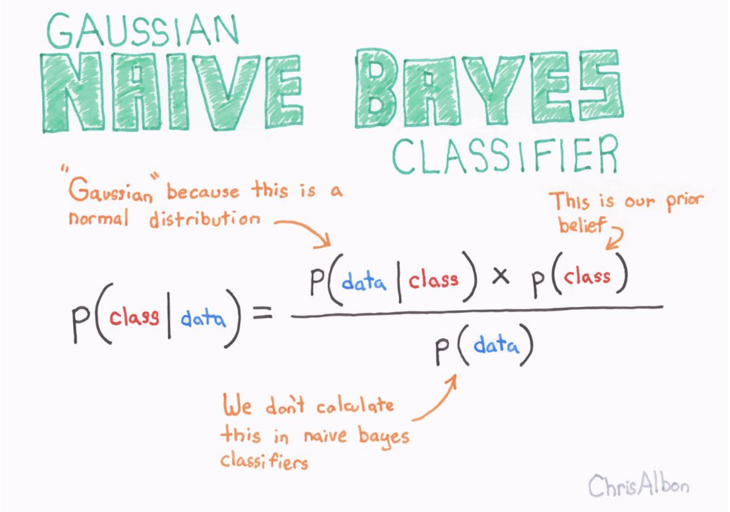
Let's review:

- What is the difference between supervised vs unsupervised learning?
- What is the difference between regression vs classification?

Naive Bayes

- Naive Bayes is incredibly **POWERFUL** classification algorithm, especially used for nlp
- In short, Naive Bayes is classification algorithm using Bayes theorem with an **assumption of independence** between predictors





Naive Bayes - Our Data

Weather	Temperature	Humidity	Windy	Picnic
Sunny	High	Normal	Low	Y
Overcast	Mild	Normal	Low	Y
Sunny	Low	Normal	High	Y
Sunny	High	Normal	High	N
Overcast	Low	High	High	N
Sunny	Mild	High	Low	Y
Overcast	Mild	Normal	Low	Y
Rainy	Low	High	High	N
Rainy	High	Normal	High	N

Naive Bayes

Recall Bayes' Theorem...we are trying to find the probability of A, given that B is True:

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

In the context of classification, we can express it as - we want to find the probability of outcome Y occurring given some features:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Naive Bayes

Mathematically we can represent it as:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1)P(x_2)...P(x_n)}$$

Since the denominator remains constant, we can rewrite it as:

$$P(y|x_1, ..., x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y)$$

Naive Bayes

Weather Condition

	Yes	No	P(Y)	P(N)
Sunny	3	1	3/5	1/4
Overcast	2	1	2/5	1/4
Rainy	0	2	0/5	2/4
Total	5	4	1	1

Naive Bayes

Humidity

	Yes	No	P(Y)	P(N)
Normal	4	2	4/5	2/4
High	1	2	1/5	2/4
Total	5	4	1	1

Naive Bayes

Wind

	Yes	No	P(Y)	P(N)
High	1	4	1/5	4/4
Low	4	0	4/5	0/4
Total	5	4	1	1

Naive Bayes

Temperature

	Yes	No	P(Y)	P(N)
High	1	2	1/5	2/4
Mild	3	0	3/5	0/4
Low	1	2	1/5	2/4
Total	5	4	1	1

Naive Bayes

Therefore, if we were to calculate the probability of going out for a picnic when the day is = (sunny, hot, normal, no wind), we can express it as:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

And the probability of not going for picnic is:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Naive Bayes

P(Yes|Today) =

P(No|Today) =

We can compare the probability of P(Yes | today) and P(No | Today) and output a prediction for playing. Since Probability of Yes is greater than probability of No, the algorithm will predict that given the conditions of sunny, high temperature, no wind, and normal humidity, we will be going out for a picnic.

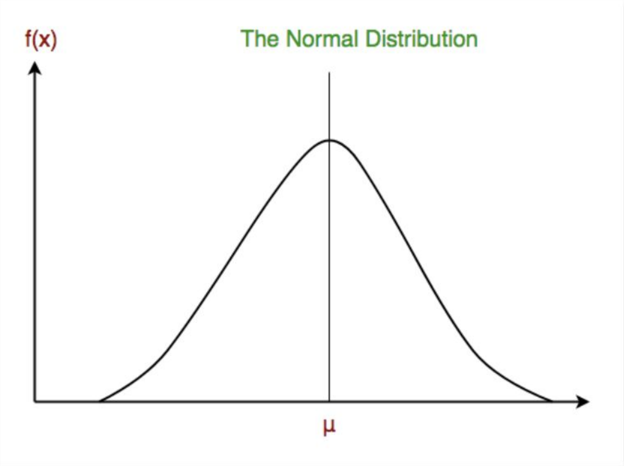
Naive Bayes

- Different types of Naive Bayes:
  - Gaussian Naive Bayes
  - Multinomial Naive Bayes
  - Bernoulli Naive Bayes

Naive Bayes

In **Gaussian Naive Bayes**, continuous values associated with each feature are assumed to be distributed according to a Gaussian/normal distribution.

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}\right)$$



Naive Bayes

Multinomial Naive Bayes: Feature vectors represent the frequencies with which certain events have been generated by a **multinomial distribution**. This is the event model typically used for document classification.

$$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$