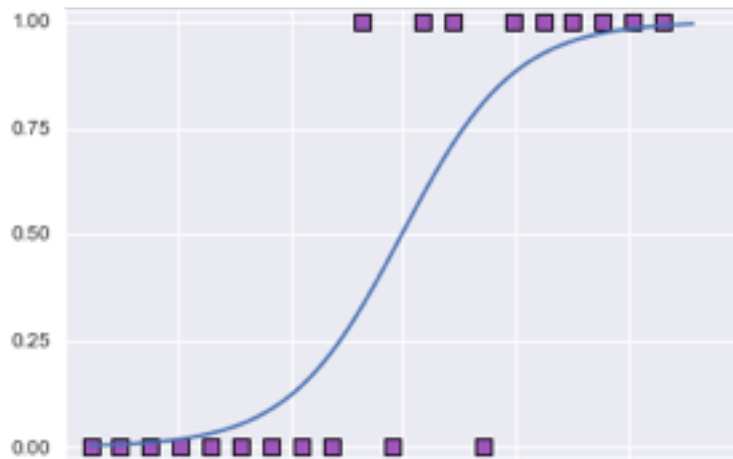


# INTRO to DATA SCIENCE

## LECTURE 11: BAYESIAN STATISTICS

- I. REGRESSION RECAP**
- II. LOGISTIC REGRESSION**
- III. INTERPRETING RESULTS**
- IV. DECISION BOUNDARIES**
- V. EVALUATING CLASSIFIERS**



any questions?

**DATA EXPLORATION**

**SUPERVISED LEARNING: REGRESSION**

**SUPERVISED LEARNING: CLASSIFICATION**

**UNSUPERVISED LEARNING**

**VARIOUS TOPICS**

**DATA EXPLORATION**

**SUPERVISED LEARNING: REGRESSION**

**SUPERVISED LEARNING: CLASSIFICATION**

**UNSUPERVISED LEARNING**

**VARIOUS TOPICS**

**LOGISTIC REGRESSION**

**NAIVE BAYES** (TODAY)

**RANDOM FORESTS**

**SUPPORT VECTOR MACHINES**

**COMPETITION**

**0. DEMO SAMPLE PROJECTS** (IDEAS DUE NEXT LESSON)

**I. PROBABILITY**

**II. BAYES' THEOREM**

**III. EXAMPLE: BAYSAIN COIN FLIPS** (OPTIONAL)

**IV. NAIVE BAYES**

# **INTRO TO DATA SCIENCE**

---

## **I. PROBABILITY**

*Q: What is a probability?*

*Q: What is a **probability**?*

*A: A number between 0 and 1 that characterizes  
the likelihood that some event will occur.*



*Q: What is a **probability**?*

*A: A number between 0 and 1 that characterizes the likelihood that some event will occur.*

*The probability of event  $A$  is denoted  $P(A)$ .*

*Q: What is a **probability**?*

*A: A number between 0 and 1 that characterizes the likelihood that some event will occur.*

*The probability of event  $A$  is denoted  $P(A)$ .*

*The space of all possible events is called the **sample space** and denoted by  $\Omega$*

*Q: What is a **probability**?*

*A: A number between 0 and 1 that characterizes the likelihood that some event will occur.*

*The probability of event  $A$  is denoted  $P(A)$ .*

*The space of all possible events is called the **sample space** and denoted by  $\Omega$*

$$P(\Omega) = 1$$

*Q: What is a **probability**?*

*A: A number between 0 and 1 that characterizes the likelihood that some event will occur.*

*The probability of event  $A$  is denoted  $P(A)$ .*

*The space of all possible events is called the **sample space** and denoted by  $\Omega$*

$$P(\Omega) = 1$$

$$P(\emptyset) = 0$$

*$\emptyset$  is the empty set  $\{ \}$*

*It makes sense to think of events of subsets in the sample space  $\Omega$*

*Two events  $A$  and  $B$  are **mutually exclusive** or **disjoint** if they are not overlapping:*

$$A \cap B = \emptyset$$

*Their intersection is an empty set*

*It makes sense to think of events of subsets in the sample space  $\Omega$*

*Two events  $A$  and  $B$  are **mutually exclusive** or **disjoint** if they are not overlapping:*

$$A \cap B = \emptyset$$

*Their intersection is an empty set*

$$P(\text{🎲} \text{ and } \text{🎲}) = 0$$

*you cannot have both 4 and 6 in one throw*

*It makes sense to think of events of subsets in the sample space  $\Omega$*

*If two events  $A$  and  $B$  are **mutually exclusive** or **disjoint**, i.e., if they are not overlapping, then we can add their probabilities*

$$P(A \text{ or } B) = P(A) + P(B)$$

*It makes sense to think of events of subsets in the sample space  $\Omega$*

*If two events  $A$  and  $B$  are **mutually exclusive** or **disjoint**, i.e., if they are not overlapping, then we can add their probabilities*

$$P(A \text{ or } B) = P(A) + P(B)$$

*If  $A$  and  $B$  are **not** mutually exclusive, then we have*

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$



*It makes sense to think of events of subsets in the sample space  $\Omega$*

*If two events  $A$  and  $B$  are **mutually exclusive***

$$P(\text{🎲} \text{ or } \text{🎲}) = P(\text{🎲}) + P(\text{🎲}) = 1/6 + 1/6 = 1/3$$

*It makes sense to think of events of subsets in the sample space  $\Omega$*

*If two events  $A$  and  $B$  are **mutually exclusive***

$$P(\text{⚪⚪ or ⚫⚫}) = P(\text{⚪⚪}) + P(\text{⚫⚫}) = 1/6 + 1/6 = 1/3$$

*If  $A$  and  $B$  are **not** mutually exclusive*

$$\begin{aligned} P(\text{7 or } \spadesuit) &= P(\text{7}) + P(\spadesuit) - P(\text{7 of } \spadesuit) \\ &= 1/13 + 1/4 - 1/52 = 11/26 \end{aligned}$$

*Two events  $A$  and  $B$  are called **independent** if their joint probability is the product of their individual probabilities:*

$$P(A \text{ and } B) = P(A) P(B)$$

*We often write  $A \perp B$*

*Two events  $A$  and  $B$  are called **independent** if their joint probability is the product of their individual probabilities:*

*two throws*     $P(\text{⚰ and ⚰}) = P(\text{⚰}) P(\text{⚰}) = 1/6 \times 1/6 = 1/36$

*Q: Suppose event  $B$  has occurred. What quantity represents the probability of  $A$  **given** this information about  $B$ ?*

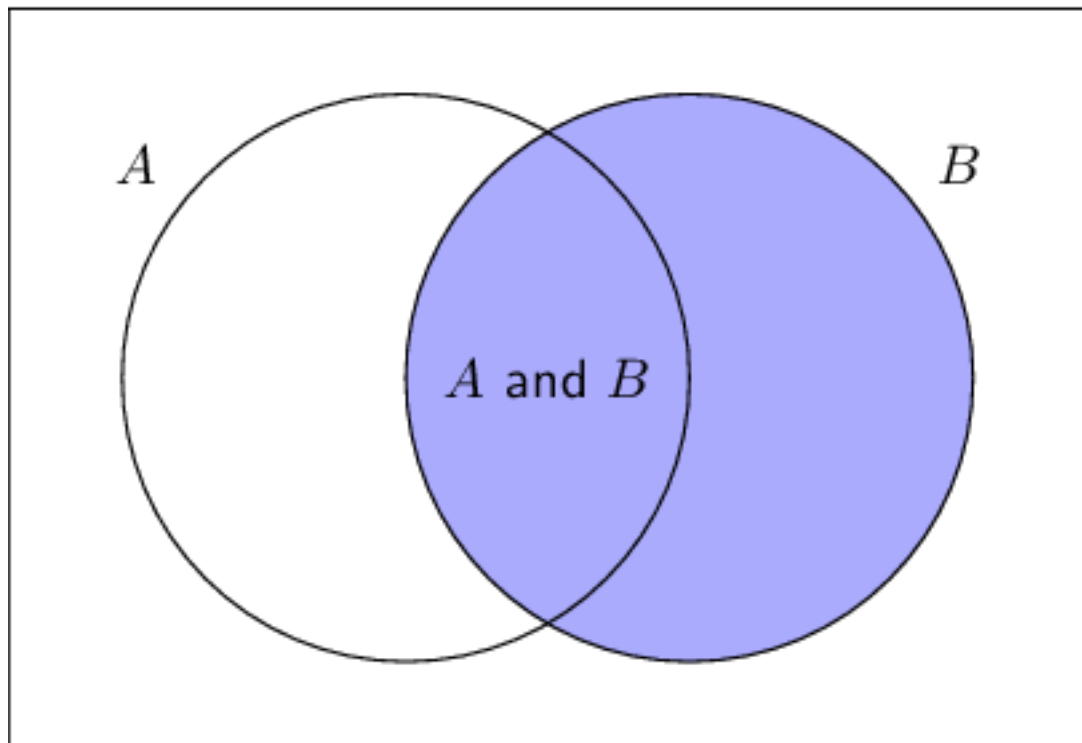
*Q: Suppose event  $B$  has occurred. What quantity represents the probability of  $A$  **given** this information about  $B$ ?*

*A: The intersection of  $A$  &  $B$  divided by region  $B$ .*

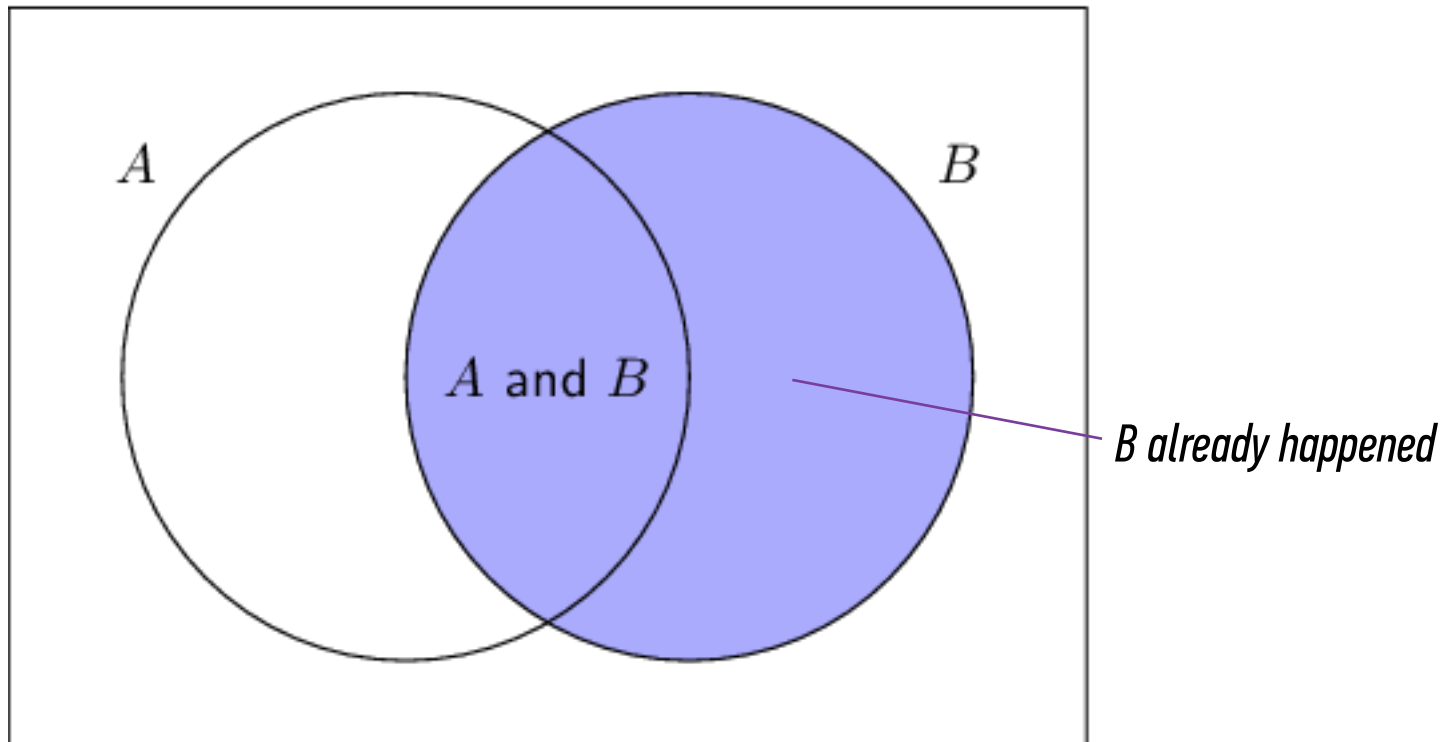
*Q: Suppose event  $B$  has occurred. What quantity represents the probability of  $A$  **given** this information about  $B$ ?*

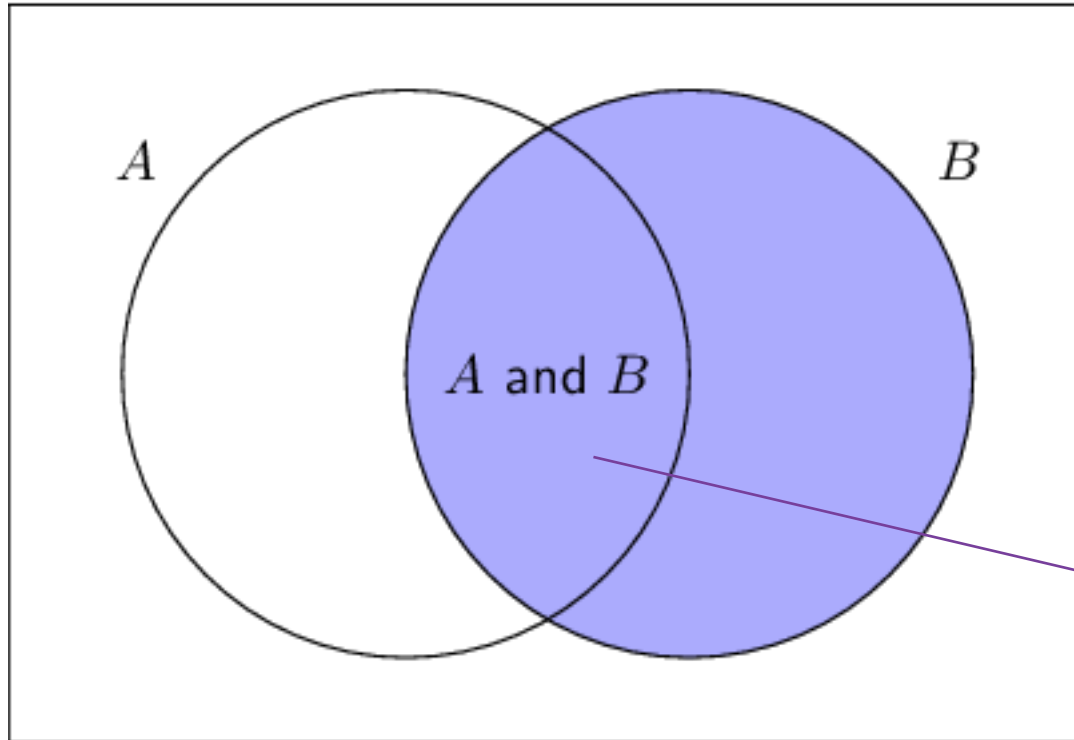
*A: The intersection of  $A$  &  $B$  divided by region  $B$ .*

*This is called the **conditional probability** of  $A$  given  $B$ , written  $P(A|B) = P(AB) / P(B)$ .*









*B already happened*

*The probability of A  
is now given by  $A \cap B$   
(or  $AB$ )*

*Two **conditional probability** is the probability of some event A, given the occurrence of some other event B.*

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

*Two **conditional probability** is the probability of some event  $A$ , given the occurrence of some other event  $B$ .*

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

*It follows that if  $P(A \mid B) = P(A)$  if and only if  $A \perp B$*

---

**INTRO TO DATA SCIENCE**

---

# ***QUIZ QUESTION***

*Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

*Which is more probable?*

*1) Linda is a bank teller.*

*2) Linda is a bank teller and active in the feminist movement.*

*Q: What does it mean for two events to be independent?*

*Q: What does it mean for two events to be **independent**?*

*A: Information about one does not affect the probability of the other.*



*Q: What does it mean for two events to be **independent**?*

*A: Information about one does not affect the probability of the other.*

*This can be written as  $P(A|B) = P(A)$*

*Q: What does it mean for two events to be **independent**?*

*A: Information about one does not affect the probability of the other.*

*This can be written as  $P(A|B) = P(A)$*

*And we have  $P(A \text{ and } B) = P(A) P(B)$*

*Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

*Which is more probable?*

*1) Linda is a bank teller.*

*2) Linda is a bank teller and active in the feminist movement.*

*Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

*Which is larger?*

*1)  $P(\text{bank teller})$*

*2)  $P(\text{bank teller and feminist movement})$*

*Linda is 31 years old, single, outspoken and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

*Which is larger?*

*1)  $P(\text{bank teller})$*

*2)  $P(\text{bank teller}) \times P(\text{feminist movement})$*

# II. BAYES' THEOREM

*Recall the* **conditional probability**

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

*Recall the* **conditional probability**

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

*We can rewrite that as*  $P(AB) = P(A|B) P(B)$



*Recall the* **conditional probability**

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

*We can rewrite that as*  $P(AB) = P(A|B) P(B)$

*As well as*  $P(AB) = P(B|A) P(A)$

*Recall the **conditional probability***

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

*We can rewrite that as*  $P(AB) = P(A|B) P(B)$

*As well as*  $P(AB) = P(B|A) P(A)$

*It follows that* 
$$P(A \mid B) = \frac{P(B|A) P(A)}{P(B)}$$

*This result is called* **Bayes' Theorem**

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

*This result is called **Bayes' Theorem***

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

*It means you can swap conditional probabilities*

*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?*

*This result is called **Bayes' Theorem***

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

*It means you can swap conditional probabilities*

$$\begin{array}{l} \text{In a movie it's raining. What's the} \\ \text{chance the movie is shot in Holland?} \end{array} = \frac{P(\text{raining in Holland}) P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

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

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

*This term is the **posterior probability** of A. It's the probability of A after 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?*

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



*This term is the **posterior probability** of A. It's the probability of A after 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.*

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

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

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

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

*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.*

*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?*

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

*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.*

*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?*

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

*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.*

$$\begin{aligned} &\text{In a movie it's raining. What's the} \\ &\text{chance the movie is shot in Holland?} \end{aligned} = \frac{P(\text{raining in Holland}) P(\text{random movie shot in Holland})}{P(\text{raining anywhere in the world})}$$

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



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

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

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

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



*Coefficients of regression*

*Class labels of samples*

*Student proficiency and question difficulty*

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

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


*Data points in Euclidean space*

*List of labeled samples*

*Student responses*

*Starting out with a prior belief of the parameters  $\beta$  ...*

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



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

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

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

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

*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?*

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

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

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

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



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

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

*As a final remark, Bayes' Theorem offers a “wormhole” between two different “interpretations” of probability*

*As a final remark, Bayes' Theorem offers a “wormhole” between two different “interpretations” of probability*

*The **frequentist** interpretation regards an event's probability as its limiting frequency across a very large number of trials*

*As a final remark, Bayes' Theorem offers a “wormhole” between two different “interpretations” of probability*

*The **frequentist** interpretation regards an event's probability as its limiting frequency across a very large number of trials*

*The **Bayesian** interpretation regards an event's probability as a “degree of belief,” which applies even to events that haven't occurred yet*

*If this sounds crazy to you, don't worry...we won't dwell on the theoretical details.*

*If this sounds crazy to you, don't worry...we won't dwell on the theoretical details.*

*If this sounds interesting: this a good direction to head if you're serious in becoming a rock star data scientist.*

### ***Method***

---

### ***Predictions***

*The **frequentist** interpretation*

*point estimates*

*The **Bayesian** interpretation*

*distributions*

# **III. BAYESIAN COIN FLIPS**

## **EXAMPLE** (SIT BACK & RELAX)



---

**INTRO TO DATA SCIENCE**

---

*(A FREQUENTIST)*  
**QUIZ QUESTION**

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?  $3/4$ , Why?

Problem:

We observe the following coin flips:

HTHHTHT

What is  $P(X = \text{Heads})$  ?

Problem:

We observe the following coin flips:

HTHHTHT

What is  $P(X = \text{Heads})$  ?  $4/7$ , Why?

Problem:

We observe the following coin flips:

HTHHTHT

What is  $P(X = \text{Heads})$  ? 4/7, Why?

With the classical method,

$$P(X = \text{head}) = \frac{\# \text{ heads}}{\# \text{ tosses}}$$

Problem:

We observe the following coin flips:

HTHHTHT

What is  $P(X = \text{Heads})$  ? 4/7, Why?

With the classical method,

$$P(X = \text{head}) = \frac{\# \text{ heads}}{\# \text{ tosses}}$$

... which is not so reliable  
with little data

Problem:

We observe the following coin flips:

H

What is  $P(X = \text{Heads})$  ?



Problem:

We observe the following coin flips:

H

What is  $P(X = \text{Heads})$  ? Exactly 1.

Why do you care?

Many problems are binary and are estimated using counts...

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 1: Sample 100 people and ask if they support a politician.

23 say Yes

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 1: Sample 100 people and ask if they support a politician.  
23 say Yes - Is the correct prediction  $23/100$ ?

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 2: Sample 100 people and ask *which* politician they support  
3 say Trump

Why do you care?

Many problems are binary and are estimated using counts...

Ex. 2: Sample 100 people and ask *which* politician they support  
3 say Trump - Is the correct prediction  $P(\text{Trump}) = 3/100$ ?

For the frequentist method, you need a lot of data to succeed

Let's try the Bayesian approach

Let's try the Bayesian approach



- This is going to be a lot of high-level math to illustrate Bayes*
- *If you'd like to fully understand, see the enclosed notebook*
  - *If you're fine with hand-waiving: sit back and relax*



$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

We don't not estimate the probability  $P(H)$  directly

$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

We don't not estimate the probability  $P(H)$  directly, but we ask:

*Given the observed tosses, what is the chance that this probability  $P(H)$  is equal to some value  $p$ ?*

$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

We don't not estimate the probability  $P(H)$  directly, but we ask:

*Given the observed tosses, what is the chance that this probability  $P(H)$  is equal to some value  $p$ ?*

We look for which  $p$  the probability of  $P(H) = p$  is the most likely.

$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

$$P(p | D) = \frac{P(D | p) \times P(p)}{P(D)}$$

let's clean up notation

$$P(P(H) = p | \text{tosses}) = \frac{P(\text{tosses} | H) \times P(P(H) = p)}{P(\text{tosses})}$$

$$P(p | D) = \frac{P(D | p) \times P(p)}{P(D)}$$

let's clean up notation

$$\max_p P(p | D) = \max_p P(D | p) P(p)$$

look for which  $p$  the probability of  $P(H) = p$  is the **most likely**

*What is the **prior distribution** of  $p$ ?*

$$\max_p P(p|D) = \max_p P(D|p)P(p)$$

*What is the **prior distribution** of  $p$ ?*

*Let's pick a simple Beta distribution*

$$P(p) = 6 p (1 - p)$$





*What is the **likelihood function** of  $p$ ?*

$$\max_p P(p|D) = \max_p P(D|p)P(p)$$

*What is the **likelihood function** of  $p$ ?*

*This is the reversed question: Given any probability  $p$ , what is the chance I'd see the observed coin tosses  $D$ ?*

*What is the **likelihood function** of  $p$ ?*

*This is the reversed question: Given any probability  $p$ , what is the chance I'd see the observed coin tosses  $D$ ?*

*That's the binomial distribution*

$$P(D|p) = \binom{N}{n} p^n (1 - p)^{N-n}$$

$$\max_p P(p|D) = \max_p P(D|p)P(p)$$

$$\max_p P(p|D) = \max_p P(D|p)P(p) = \max_p \underbrace{\binom{N}{n} p^n (1-p)^{N-n}}_{\text{likelihood function}} \underbrace{6p(1-p)}_{\text{prior belief}}$$

$$\max_p P(p|D) = \max_p P(D|p)P(p) = \max_p \underbrace{\binom{N}{n} p^n (1-p)^{N-n}}_{\text{likelihood function}} \underbrace{6p(1-p)}_{\text{prior belief}}$$

$$\frac{d}{dp} P(p|D) = 0$$

derivative is zero at maximum

$$\max_p P(p|D) = \max_p P(D|p)P(p) = \max_p \underbrace{\binom{N}{n} p^n (1-p)^{N-n}}_{\text{likelihood function}} \underbrace{6p(1-p)}_{\text{prior belief}}$$

$$\frac{d}{dp} P(p|D) = 0$$

derivative is zero at maximum

$$p = \frac{n+1}{N+2}$$

solution follows algebraically

$$p = \frac{n+1}{N+2} = \begin{cases} 1/2 & \text{if no coins have been tossed yet } (N=0) \end{cases}$$

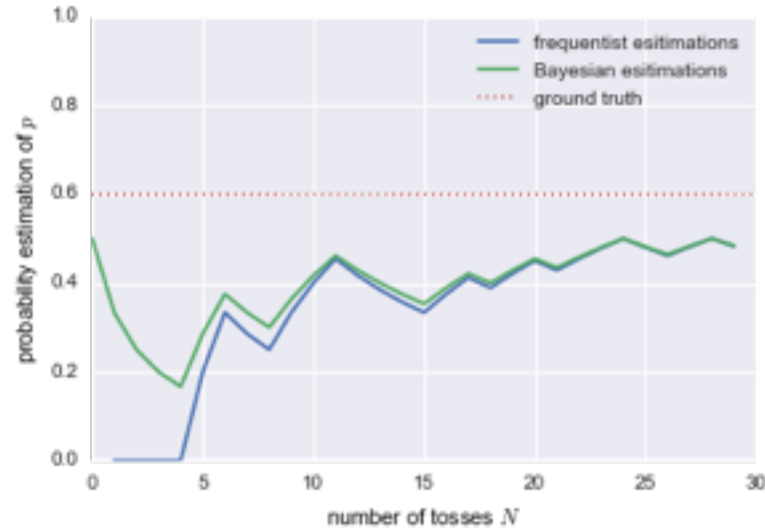


$$p = \frac{n+1}{N+2} = \begin{cases} 1/2 & \text{if no coins have been tossed yet } (N=0) \\ \rightarrow n/N & \text{if many coins have been tossed (i.e., frequentist)} \end{cases}$$

*In this case, you could interpret this as the classical approach, after having added two imaginary coin flips at the beginning. We count one more head, and one more tail:*

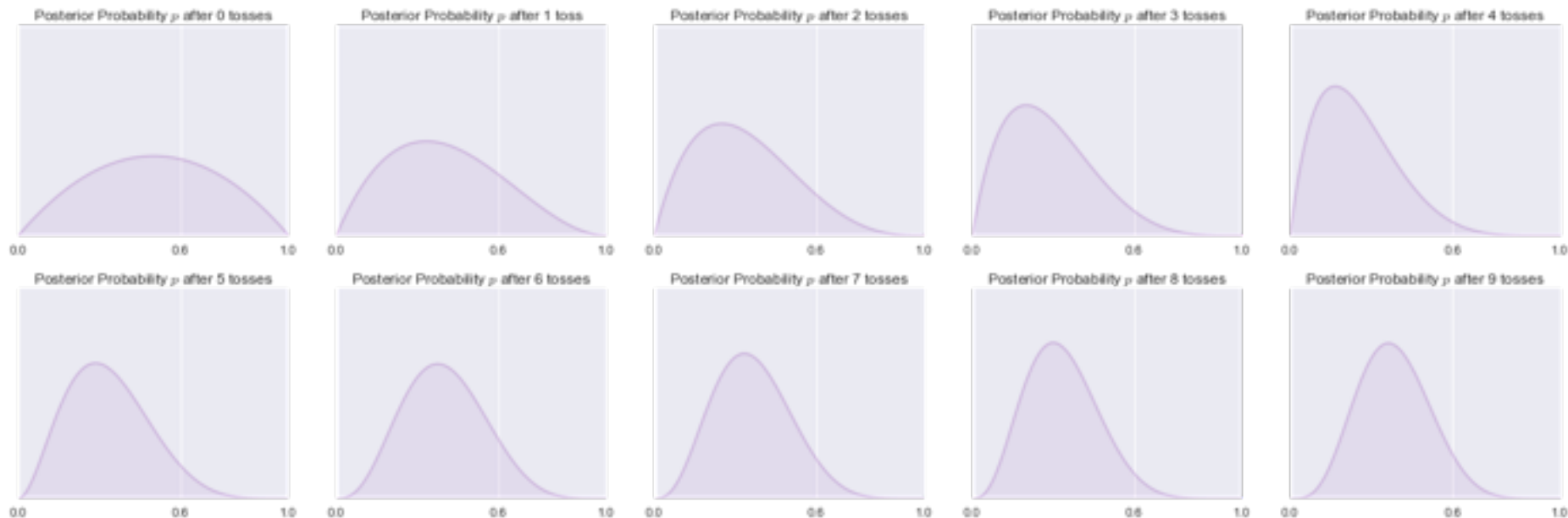
[HT]HTHHTHT

$$p = \frac{n+1}{N+2} = \begin{cases} 1/2 & \text{if no coins have been tossed yet } (N=0) \\ \rightarrow n/N & \text{if many coins have been tossed (i.e., frequentist)} \end{cases}$$



$$p = \frac{n+1}{N+2} = \begin{cases} 1/2 & \text{if no coins have been tossed yet } (N=0) \\ \rightarrow n/N & \text{if many coins have been tossed (i.e., frequentist)} \end{cases}$$

*Bayes provides you distributions, rather than point estimates*



---

**INTRO TO DATA SCIENCE**

---

*(A BAYESIAN)*  
***QUIZ QUESTION***

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Well, it depends on...

*... on what?*

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Well, it depends on the prior

*Which one shall we take?*



Problem:

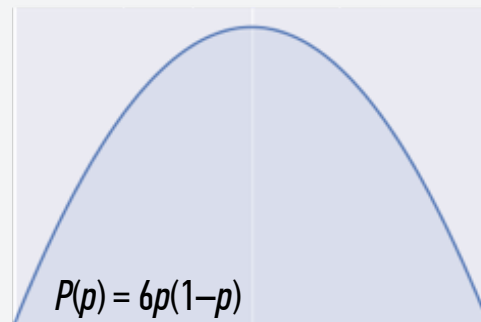
We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Well, it depends on the prior

- ▶ Let's take the Beta prior



Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?

Well, it depends on the prior

- ▶ Let's take the Beta prior

$$P(X = \text{head}) = \frac{\# \text{ heads} + 1}{\# \text{ tosses} + 2}$$

Problem:

We observe the following coin flips:

HTHH

What is  $P(X = \text{Heads})$  ?  $2/3$

Well, it depends on the prior

- ▶ Let's take the Beta prior

$$P(X = \text{head}) = \frac{\# \text{ heads} + 1}{\# \text{ tosses} + 2}$$

# **IV. NAIVE BAYES**

Confused?

Confused? Relax, it gets easier!

*Suppose we have a dataset with features  $x_1, \dots, x_n$  and class labels  $c$ .  
What can we say about classification using Bayes' theorem?*

*Suppose we have a dataset with features  $x_1, \dots, x_n$  and class labels  $C$ .  
What can we say about classification using Bayes' theorem?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

What is the chance that these words  
(and n-grams) have class label  $C$ ?




*Suppose we have a dataset with features  $x_1, \dots, x_n$  and class labels  $C$ . What can we say about classification using Bayes' theorem?*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$

*Bayes' theorem can help us to determine the probability of a record belonging to a class, given the data we observe.*

*However, the likelihood function can often be intractably difficult in practice to determine*

$$P(\text{class } C \mid \{x_i\}) = \frac{P(\{x_i\} \mid \text{class } C) \cdot P(\text{class } C)}{P(\{x_i\})}$$


$$P(\{x_i\} \mid C) = P(\{x_1, x_2, \dots, x_n\} \mid C)$$

What is the chance that a random sample from a given class C has exactly all these words (and n-grams)?

*So let's make a simplifying assumption. In particular, we assume that the features  $x_i$  are conditionally independent from each other:*

$$P(\{x_i\}|C) = P(x_1, x_2, \dots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \dots * P(x_n|C)$$



What is the chance that a random word (or n-gram) from a given class C is exactly word  $x_1$ ?

*So let's make a simplifying assumption. In particular, we assume that the features  $x_i$  are conditionally independent from each other:*

$$P(\{x_i\}|C) = P(x_1, x_2, \dots, x_n|C) \approx P(x_1|C) * P(x_2|C) * \dots * P(x_n|C)$$

*This “naive” assumption simplifies the likelihood function to make it tractable.*

*The Naive Bayes algorithm combines the probability of a class  $C$  overall with the probabilities of each individual feature appearing in class  $C$*

$$P(C|\{x_i\}) \sim P(C) \prod_i P(x_i|C)$$

---

**INTRO TO DATA SCIENCE**

---

**DISCUSSION**