Probability theory yields mathematical tools to deal with uncertain events.
Probability theory yields mathematical tools to deal with uncertain events.

Used everywhere nowadays and its importance is growing.
Probability and Statistics

- Probability $\neq$ Statistics

- Probability: Known distributions $\Rightarrow$ what are the outcomes?

- Statistics: Known outcomes $\Rightarrow$ what are the distributions?
Counting

- Many **basic** probability problems are counting problems.

Example: Assume there are 1 man and 2 women in a room. You pick a person randomly. What is the probability $P_1$ that this is a man?

If you pick two persons randomly, what is the probability $P_2$ that these are a man and woman?

Answer: You have the possible outcomes: (M), (W1), (W2) so $P_1 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{\text{# men}}{\text{# men} + \text{# women}} = \frac{1}{3}$.

To compute $P_2$, you can think of all the possible events: (M,W1), (M,W2), (W1,W2) so $P_2 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{2}{3}$.
Many **basic** probability problems are counting problems.

*Example*: Assume there are 1 man and 2 women in a room. You pick a person randomly. What is the probability $P_1$ that this is a man?

If you pick two persons randomly, what is the probability $P_2$ that these are a man and woman?

**Answer**: You have the possible outcomes: (M), (W1), (W2) so $P_1 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{\text{# men}}{\text{# men} + \text{# women}} = \frac{1}{3}$.

To compute $P_2$, you can think of all the possible events: (M,W1), (M,W2), (W1,W2) so $P_2 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{2}{3}$. 
Counting

- Many **basic** probability problems are counting problems.
- *Example*: Assume there are 1 man and 2 women in a room. You pick a person randomly. What is the probability $P_1$ that this is a man?

If you pick two persons randomly, what is the probability $P_2$ that these are a man and woman?

Answer: You have the possible outcomes: (M), (W1), (W2) so $P_1 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{\text{# men}}{\text{# men} + \text{# women}} = \frac{1}{3}$.

To compute $P_2$, you can think of all the possible events: (M,W1), (M,W2), (W1,W2) so $P_2 = \frac{\text{# "successful" events}}{\text{# events}} = \frac{2}{3}$. 
Many **basic** probability problems are counting problems.

*Example*: Assume there are 1 man and 2 women in a room. You pick a person randomly. What is the probability $P_1$ that this is a man? If you pick two persons randomly, what is the probability $P_2$ that these are a man and woman?

*Answer*: You have the possible outcomes: (M), (W1), (W2) so

$$P_1 = \frac{\# \text{ “successful” events}}{\# \text{ events}} = \frac{\# \text{ men}}{\# \text{ men} + \# \text{ women}} = \frac{1}{3}.$$ 

To compute $P_2$, you can think of all the possible events: (M,W1), (M,W2), (W1,W2) so

$$P_2 = \frac{\# \text{ “successful” events}}{\# \text{ events}} = \frac{2}{3}.$$
Sample Space

Definition
The *sample space* $S$ of an experiment (whose outcome is uncertain) is the set of all possible outcomes of the experiment.
Example (child): Determining the sex of a newborn child in which case \( S = \{ \text{boy, girl} \} \).
Sample Space

- **Example (child):** Determining the sex of a newborn child in which case \( S = \{\text{boy}, \text{girl}\} \).
- **Example (horse race):** Assume you have an horse race with 12 horses. If the experiment is the order of finish in a race, then

\[
S = \{\text{all } 12! \text{ permutations of } (1, 2, 3, ..., 11, 12)\}.
\]
Sample Space

- *Example (child)*: Determining the sex of a newborn child in which case $S = \{\text{boy}, \text{girl}\}$.

- *Example (horse race)*: Assume you have an horse race with 12 horses. If the experiment is the order of finish in a race, then

  $$S = \{\text{all 12! permutations of } (1, 2, 3, ..., 11, 12)\}.$$ 

- *Example (coins)*: If the experiment consists of flipping two coins, then the sample space is

  $$S = \{(H, H), (H, T), (T, H), (T, T)\}.$$
Sample Space

- **Example (child):** Determining the sex of a newborn child in which case \( S = \{ \text{boy}, \text{girl} \} \).

- **Example (horse race):** Assume you have an horse race with 12 horses. If the experiment is the order of finish in a race, then

\[
S = \{ \text{all 12! permutations of } (1, 2, 3, \ldots, 11, 12) \}.
\]

- **Example (coins):** If the experiment consists of flipping two coins, then the sample space is

\[
S = \{ (H, H), (H, T), (T, H), (T, T) \}.
\]

- **Example (lifetime):** If the experiment consists of measuring the lifetime (in years) of your pet then the sample space consists of all nonnegative real numbers: \( S = \{ x; 0 \leq x < \infty \} \).
Any subset $E$ of the sample space $S$ is known as an event; i.e. an event is a set consisting of possible outcomes of the experiment.
Events

- Any subset $E$ of the sample space $S$ is known as an event; i.e. an event is a set consisting of possible outcomes of the experiment.
- If the outcome of the experiment is in $E$, then we say that $E$ has occurred.
Events

- Example (child): The event $E = \{\text{boy}\}$ is the event that the child is a boy.
Events

- *Example (child):* The event $E = \{\text{boy}\}$ is the event that the child is a boy.

- *Example (horse race):* The event $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ is the event that the race was won by horse 7.
Events

- **Example (child)**: The event $E = \{\text{boy}\}$ is the event that the child is a boy.
- **Example (horse race)**: The event $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ is the event that the race was won by horse 7.
- **Example (coins)**: The event $E = \{(H, T), (T, T)\}$ is the event that a tail appears on the second coin.
Events

- **Example (child):** The event $E = \{\text{boy}\}$ is the event that the child is a boy.

- **Example (horse race):** The event $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ is the event that the race was won by horse 7.

- **Example (coins):** The event $E = \{(H, T), (T, T)\}$ is the event that a tail appears on the second coin.

- **Example (lifetime):** The event $E = \{x : 3 \leq x \leq 15\}$ is the event that your pet will live more than 3 years but won’t live more than 15 years.
Given events $E$ and $F$, $E \cup F$ is the set of all outcomes either in $E$ or $F$ or in both $E$ and $F$.

$E \cup F$ occurs if either $E$ or $F$ occurs.

$E \cup F$ is the union of events $E$ and $F$. 
Example (coins): If we have $E = \{(H, T)\}$ and $F = \{(T, H)\}$ then $E \cup F = \{(H, T), (T, H)\}$ is the event that one coin is head and the other is tail.
Union of Events

- **Example (coins):** If we have $E = \{(H, T)\}$ and $F = \{(T, H)\}$ then $E \cup F = \{(H, T), (T, H)\}$ is the event that one coin is head and the other is tail.

- **Example (horse race):** If we have
  
  $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ and
  $F = \{\text{all outcomes in } S \text{ finishing with a 3}\}$ then $E \cup F$ is the event that the race was won by horse 7 and/or the last horse was horse 3.
Union of Events

- **Example (coins):** If we have \( E = \{ (H, T) \} \) and \( F = \{ (T, H) \} \) then \( E \cup F = \{ (H, T), (T, H) \} \) is the event that one coin is head and the other is tail.

- **Example (horse race):** If we have
  
  \[ E = \{ \text{all outcomes in } S \text{ starting with a } 7 \} \]
  
  \[ F = \{ \text{all outcomes in } S \text{ finishing with a } 3 \} \]

  then \( E \cup F \) is the event that the race was won by horse 7 and/or the last horse was horse 3.

- **Example (lifetime):** If \( E = \{ x : 0 \leq x \leq 10 \} \) and
  
  \[ F = \{ x : 15 \leq x < \infty \} \]

  then \( E \cup F \) is the event that your pet will die before 10 or will die after 15.
Intersection of Events

Given events $E$ and $F$, $E \cap F$ is the set of all outcomes which are both in $E$ and $F$.

$E \cap F$ is also denoted as $EF$. 
Intersection of Events

- **Example (coins):** If we have \( E = \{(H, H), (H, T), (T, H)\} \) (event that one H at least occurs) and \( F = \{(H, T), (T, H), (T, T)\} \) (even that one T at least occurs) then \( E \cap F = \{(H, T), (T, H)\} \) is the event that one H and one T occur.
Intersection of Events

- **Example (coins):** If we have $E = \{(H, H), (H, T), (T, H)\}$ (event that one $H$ at least occurs) and $F = \{(H, T), (T, H), (T, T)\}$ (event that one $T$ at least occurs) then $E \cap F = \{(H, T), (T, H)\}$ is the event that one $H$ and one $T$ occur.

- **Example (horse race):** If we have $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ and $F = \{\text{all outcomes in } S \text{ starting with a 8}\}$ then $E \cap F$ does not contain any outcome and is denoted by $\emptyset$. 

Janyl Jumadinova

Basics of Probability

February 24–26, 2020
Intersection of Events

**Example (coins):** If we have $E = \{(H, H), (H, T), (T, H)\}$ (event that one H at least occurs) and $F = \{(H, T), (T, H), (T, T)\}$ (even that one T at least occurs) then $E \cap F = \{(H, T), (T, H)\}$ is the event that one H and one T occur.

**Example (horse race):** If we have $E = \{\text{all outcomes in } S \text{ starting with a 7}\}$ and $F = \{\text{all outcomes in } S \text{ starting with a 8}\}$ then $E \cap F$ does not contain any outcome and is denoted by $\emptyset$.

**Example (lifetime):** If we have $E = \{x : 0 \leq x \leq 5\}$ and $F = \{x : 10 \leq x < 15\}$ then $E \cap F = \{x : 3 \leq x \leq 5\}$ is the event that your pet will die between 10 and 15.
For any event $E$, $E^c$ denote the *complement* set of all outcomes in $S$ which are not in $E$.

Hence we have $E \cup E^c = S$ and $E \cap E^c = \emptyset$. 
For any event $E$, $E^c$ denote the complement set of all outcomes in $S$ which are not in $E$. Hence we have $E \cup E^c = S$ and $E \cap E^c = \emptyset$.

For any two events $E$ and $F$, we write $E \subset F$ is all the outcomes of $E$ are in $F$. 
Axioms of Probability

Consider an experiment with sample space $S$. For each event $E$, we assume that a number $P(E)$, the *probability* of the event $E$, is defined and satisfies the following 3 axioms.

- **Axiom 1**
  \[ 0 \leq P(E) \leq 1 \]

- **Axiom 2**
  \[ P(S) = 1 \]

- **Axiom 3**
  For any sequence of mutually exclusive events $\{E_i\}_{i \geq 1}$, i.e., $E_i \cap E_j = \emptyset$ when $i \neq j$, then
  \[ P(\bigcup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i) \]
Axioms of Probability

Consider an experiment with sample space $S$. For each event $E$, we assume that a number $P(E)$, the *probability* of the event $E$, is defined and satisfies the following 3 axioms.

- **Axiom 1**
  
  $$0 \leq P(E) \leq 1$$
Axioms of Probability

- Consider an experiment with sample space $S$. For each event $E$, we assume that a number $P(E)$, the *probability* of the event $E$, is defined and satisfies the following 3 axioms.

- **Axiom 1**

  
  
  
  
  
  $0 \leq P(E) \leq 1$

- **Axiom 2**

  
  
  
  
  
  $P(S) = 1$
Consider an experiment with sample space $S$. For each event $E$, we assume that a number $P(E)$, the *probability* of the event $E$, is defined and satisfies the following 3 axioms.

- **Axiom 1**
  
  $$0 \leq P(E) \leq 1$$

- **Axiom 2**
  
  $$P(S) = 1$$

- **Axiom 3.** For any sequence of mutually exclusive events $\{E_i\}_{i \geq 1}$, i.e. $E_i \cap E_j = \emptyset$ when $i \neq j$, then

  $$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$
Properties

- Proposition: $P(E^c) = 1 - P(E)$.
Properties

- **Proposition**: $P(E^c) = 1 - P(E)$.
- **Proposition**: If $E \subset F$ then $P(E) \leq P(F)$.
Properties

- **Proposition**: \( P(E^c) = 1 - P(E) \).
- **Proposition**: If \( E \subset F \) then \( P(E) \leq P(F) \).
- **Proposition**: We have \( P(E \cup F) = P(E) + P(F) - P(E \cap F) \).
Conditional Probabilities

• **Conditional Probability.** Consider an experiment with sample space $S$. Let $E$ and $F$ be two events, then the conditional probability of $E$ given $F$ is denoted by $P(E|F)$ and satisfies if $P(F) > 0$

$$P(E|F) = \frac{P(E \cap F)}{P(F)}$$
**Conditional Probabilities**

- **Conditional Probability.** Consider an experiment with sample space $S$. Let $E$ and $F$ be two events, then the conditional probability of $E$ given $F$ is denoted by $P(E|F)$ and satisfies if $P(F) > 0$

  $P(E|F) = \frac{P(E \cap F)}{P(F)}$

- **Intuition:** If $F$ has occurred, then, in order for $E$ to occur, it is necessary that the occurrence be both in $E$ and $F$, hence it must be in $E \cap F$. Once $F$ has occurred, $F$ is the new sample space.
Conditional Probabilities

- *Equally likely outcomes.* In this case, we have

\[
P(E|F) = \frac{\# \text{ outcomes in } E \cap F}{\# \text{ outcomes in } F} = \frac{\# \text{ outcomes in } E \cap F}{\# \text{ outcomes in } S} \cdot \frac{\# \text{ outcomes in } F}{\# \text{ outcomes in } S} = \frac{P(E \cap F)}{P(F)}.
\]
Independence

Events $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$
Independence

- Events $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$
- Equivalent to $P(A|B) = P(A)$
Independence

- Events $A$ and $B$ are independent iff $P(A \cap B) = P(A)P(B)$
- Equivalent to $P(A|B) = P(A)$
- One event occurring does not effect the probability of another occurring
The Multiplication Rule

Let $E_1, E_2, \ldots, E_n$ be a sequence of events, then we have

$$P(E_1 \cap E_2 \cap \cdots \cap E_n) = P(E_1) P(E_2|E_1) \times P(E_3|E_1 \cap E_2) \cdots P(E_n|E_1 \cap \cdots \cap E_{n-1})$$
**Example:** You have a box with 3 blue marbles, 2 red marbles, and 4 yellow marbles. You are going to pull out one marble, record its color, put it back in the box and draw another marble. What is the probability of pulling out a red followed by a blue?
Example

- **Example**: You have a box with 3 blue marbles, 2 red marbles, and 4 yellow marbles. You are going to pull out one marble, record its color, put it back in the box and draw another marble. What is the probability of pulling out a red followed by a blue?

- **Example**: Consider the same box of marbles. However, we are going to pull out the first marble, leave it out and then pull out the second marble. What is the probability of pulling out a red marble followed by a blue marble?
A random variable is a function $R : S \rightarrow \mathbb{R}$
Random Variables

- A random variable is a function $R : S \rightarrow R$
- Domain of $R$ is the sample space $S$
Random Variables

- A random variable is a function $R : S \rightarrow R$
- Domain of $R$ is the sample space $S$
- Range of $R$ is the real line
Example: **Discrete Random Variable**

*Experiment*: flip 10 coins

*Desired outcome*: the number of heads

*We care about*: the number of heads that appear among 10 tosses (not the probability of getting a particular sequence of heads and tails)
Example: **Discrete Random Variable**

*Experiment:* flip 10 coins

*Desired outcome:* the number of heads

*We care about:* the number of heads that appear among 10 tosses (not the probability of getting a particular sequence of heads and tails)

Probability of a random variable $R$ taking on some specific value $k$ is:

$$P(R = k) = P(\{s : R(s) = k\})$$, with $R(s)$ - number of heads occurring after $s$ tosses
Random Variables

Example: **Continuous Random Variable**

$R(s)$ - random variable indicating the amount of time it takes for a fast food burger to decay
Example: **Continuous Random Variable**

$R(s)$ - random variable indicating the amount of time it takes for a fast food burger to decay.

Probability that $R$ takes on a value between two real constants $a$ and $b$ is:

$$P(a \leq R \leq b) = P(\{s : a \leq R(s) \leq b\})$$
A probability distribution is a summary of probabilities for the values of a random variable.

- It is a list/table/equation that links all possible outcomes of a random variable to their corresponding probability values.
A probability distribution is a summary of probabilities for the values of a random variable.

- It is a list/table/equation that links all possible outcomes of a random variable to their corresponding probability values.

- **Mean** is the arithmetical average value of the data.
- **Median** is the middle value of the data.
- **Mode** is the most frequently occurring value of the data.
- **Expected value** of some a random variable $X$ with respect to a distribution $P(X=x)$ is the mean value of $X$ when $x$ is drawn from $P$.
- **Variance** is the measure of variability in the data from the mean value.
Probability Distribution

**Binomial**: the random variable can have only two outcomes.

```python
import numpy as np
n=100  # number of trials
p=0.5  # probability of success
s=1000 # size
np.random.binomial(n,p,s)
```
**Probability Distribution**

**Binomial**: the random variable can have only two outcomes.

```python
import numpy as np
n=100 # number of trials
p=0.5 # probability of success
s=1000 # size
np.random.binomial(n,p,s)
```

**Uniform**: equal likelihood.

```python
import numpy as np
np.random.uniform(low=1, high=10,size=100)
```
Normal (Gaussian): most common.

$\sigma = 1$
Bayesian approach provides mathematical rule explaining how you should change your existing beliefs in the light of new evidence.
Bayes’s theorem

- *posterior* = \( \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \)
Bayes’s theorem

- posterior = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}

- \( P(R = r | e) = \frac{P(e | R = r)P(R = r)}{P(e)} \)

- \( P(R = r | e) \): probability that random variable \( R \) has value \( r \) given evidence \( e \)
Bayes’s theorem

- **posterior** = \( \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}} \)
- \( P(R = r | e) = \frac{P(e | R = r)P(R = r)}{P(e)} \)
- \( P(R = r | e) \): probability that random variable \( R \) has value \( r \) given evidence \( e \)
- The denominator is just a normalizing constant (called **marginal likelihood**) that ensures the posterior adds up to 1; it can be computed by summing up the numerator over all possible values of \( R \), i.e.,
  \[
  P(e) = P(R = 0, e) + P(R = 1, e) + \ldots = 
  \sum_r P(e | R = r)P(R = r)
  \]
Naive Bayes Algorithm

- Simple ("naive") classification method based on Bayes rule
- Relies on very simple representation of document:
  - e.g. "bag of words"
**Text Classification**

**Input:**
- document $d$
- fixed set of classes $C = \{c_1, c_2, \ldots, c_j\}$

**Output:** predicted class $c \in C$
Supervised Learning for Text Classification

**Input:**
- document \( d \)
- fixed set of classes \( C = \{c_1, c_2, ..., c_j\} \)
- A training set of \( m \) hand-labeled documents \( (d_1, c_1), ..., (d_m, c_m) \)

**Output:** a learned classifier \( \gamma : d \to c \)
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.
Naive Bayes Algorithm

\[ y(\cdot) = C \]

x love xxxxxxxxxxxxxxxxxxx sweet
xxxxxxxx satirical xxxxxxxxxx
xxxxxxxxxxxxx great xxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxx fun xxxxx
xxxxxxxxxxxxxxxxxx whimsical xxxxx
romantic xxxx laughing
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx recommend xxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx xx several xxxxxxxxxxxxxxxxxxxxxx
xxxxx happy xxxxxxxxxx again
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
Naive Bayes Algorithm

\[
\gamma(\text{great, love, recommend, laugh, happy, ...}) = C
\]

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>great</td>
<td>2</td>
</tr>
<tr>
<td>love</td>
<td>2</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>laugh</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Naive Bayes Algorithm

For a document $d$ and a class $c$

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$
Naive Bayes Algorithm

\[
c_{MAP} = \arg\max_{c \in C} P(c | d)
\]

\[
= \arg\max_{c \in C} \frac{P(d | c)P(c)}{P(d)}
\]

\[
= \arg\max_{c \in C} P(d | c)P(c)
\]

MAP is “maximum a posteriori” = most likely class

Bayes Rule

Dropping the denominator
Naive Bayes Algorithm

\[ c_{MAP} = \arg\max_{c \in C} P(d \mid c)P(c) \]

\[ = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c) \]

Document d represented as features x1..xn
Naive Bayes Algorithm

\[ c_{MAP} = \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n | c)P(c) \]

- \(O(|X|^n \cdot |C|)\) parameters
- How often does this class occur?
- Could only be estimated if a very, very large number of training examples was available.
- We can just count the relative frequencies in a corpus.
Binarized (Boolean feature) Multinomial Naive Bayes

Intuition:

- Word occurrence may matter more than word frequency
- The occurrence of the word *fantastic* tells us a lot
- The fact that it occurs 5 times may not tell us much more
Binarized (Boolean feature) Multinomial Naive Bayes

**Intuition:**
- Word occurrence may matter more than word frequency
- The occurrence of the word *fantastic* tells us a lot
- The fact that it occurs 5 times may not tell us much more

**Boolean Multinomial Naive Bayes**
Clips all the word counts in each document at 1
Multinomial Naive Bayes

Assumption:

- **Bag of Words**: assume position does not matter.
- **Conditional Independence**: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class $c$.

$$P(x_1, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \cdots P(x_n|c)$$
Multinomial Naive Bayes

**Assumption:**
- Bag of Words: assume position does not matter.
- Conditional Independence: Assume the feature probabilities $P(x_i|c_j)$ are independent given the class $c$.

$$P(x_1, \ldots, x_n|c) = P(x_1|c) \cdot P(x_2|c) \cdots P(x_n|c)$$

**Applying Multinomial Naive Bayes Classifiers to Text Classification**
- features are generated from a simple multinomial distribution.
- multinomial distribution: probability of observing counts among a number of categories.