Machine Learning Toolbox for Text Analysis: Probability, stat and all that

Jessy Li, The University of Texas at Austin
Original slides from Byron Wallace
Machine learning vs statistics?

“Machine learning without statistics is pure nonsense” —Eli Upfal, Brown
Today: a very basic overview/recap of stat & probability

• **Not** a replacement for a proper intro to probability theory and stats

• Will try to convey the very basic intuition behind some of the ideas

• **Warning**: half knowledge can be dangerous
Graphically

A: people with cancer
¬A: people w/o cancer

\[ P(A) = \frac{|A|}{|U|} \]
Graphically

B: people tested positive in a screening test
¬B: people tested negative

\[ P(B) = \frac{|B|}{|U|} \]
Putting them together...

\[ p(A \cap B) = \frac{|A \cap B|}{|U|} \]

|union of all events| = |U|:
P|union of all events| = 1
The definition

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A probability space has three components:</td>
</tr>
<tr>
<td>1 A sample space ( \Omega ), which is the set of all possible outcomes of the random process modeled by the probability space;</td>
</tr>
</tbody>
</table>
The definition

**Definition**
A probability space has three components:

1. A sample space \( \Omega \), which is the set of all possible outcomes of the random process modeled by the probability space;
2. A family of sets \( \mathcal{F} \) representing the allowable events, where each set in \( \mathcal{F} \) is a subset of the sample space \( \Omega \);
The definition

<table>
<thead>
<tr>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A probability space has three components:</td>
</tr>
<tr>
<td>1 A sample space ( \Omega ), which is the set of all possible outcomes of the random process modeled by the probability space;</td>
</tr>
<tr>
<td>2 A family of sets ( \mathcal{F} ) representing the allowable events, where each set in ( \mathcal{F} ) is a subset of the sample space ( \Omega );</td>
</tr>
<tr>
<td>3 A probability function ( \text{Pr} : \mathcal{F} \rightarrow \mathbb{R} ), satisfying the definition below.</td>
</tr>
</tbody>
</table>
Probability function

Definition

A probability function is any function \( \Pr : \mathcal{F} \to \mathbb{R} \) that satisfies the following conditions:

1. For any event \( E \), \( 0 \leq \Pr(E) \leq 1 \);
A probability function is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies the following conditions:

1. For any event $E$, $0 \leq \Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$;
A **probability function** is any function $\Pr : \mathcal{F} \to \mathbb{R}$ that satisfies the following conditions:

1. For any event $E$, $0 \leq \Pr(E) \leq 1$;
2. $\Pr(\Omega) = 1$;
3. For any finite or countably infinite sequence of pairwise mutually disjoint events $E_1, E_2, E_3, \ldots$

$$
\Pr \left( \bigcup_{i \geq 1} E_i \right) = \sum_{i \geq 1} \Pr(E_i). 
$$
Tossing a coin

\[ \Omega = \{ H, T \} \]

\[ F = 2^\Omega = 2^2 = 4 \text{Events} \]

\[ F = \{ \emptyset, \{ H \}, \{ T \}, \{ H, T \} \} \]

\[ \Pr(\emptyset) = 0 \]
\[ \Pr(\{ H \}) = 0.5 \]
\[ \Pr(\{ T \}) = 0.5 \]
\[ \Pr(\{ H, T \}) = 1 \]
Tossing a fair coin twice

\[ \Omega = \{HH, HT, TH, TT\} \]

\[ F = 2^{\Omega} = 2^4 \text{ Events} \]

\[ \Pr(\{\} ) = 0 \]
\[ \Pr(\{HH\}) = \Pr(\{H\}) \Pr(\{H\}) = 0.5 \times 0.5 = 0.25 \]
\[ \Pr(\{HT\}) = \Pr(\{TH\}) = \Pr(\{TT\}) = 0.25 \]
\[ \Pr(\{HT, TT\}) = \Pr(\{HH, TH\}) = 0.5 \]
\[ \Pr(\{HH, HT\}) = \Pr(\{TH, TT\}) = 0.5 \]

...
Rolling a die

\[ \Omega = \{1,2,3,4,5,6\} \]

\[ F = 2^\Omega = 2^6 \text{ Events} \]

\[ \Pr(\{\ \}) = 0 \]

\[ \Pr(\{1\}) = \Pr(\{2\}) = \Pr(\{3\}) = \Pr(\{4\}) = \Pr(\{5\}) = \Pr(\{6\}) = \frac{1}{6} \]

\[ \Pr(\{1,2\}) = \Pr(\{1,3\}) = \Pr(\{1,4\}) = \Pr(\{1,5\}) = \Pr(\{1,6\}) = \frac{2}{6} \]

...
Rolling dice

Roll a die twice and consider the two events
A = \{ \text{sum is even} \},
B = \{ 2^{\text{nd}} \text{ outcome is even} \}.
What is P(A), P(B)
A (small) roulette wheel is labelled with the integers from 0 to 12. Each number is equally likely to come up. You spin the wheel twice.

1. What is the sample space $\Omega$?
2. What is the probability that you observe two 7’s in a row?
Conditional probability

We have two coins, coin $A$ is a fair coin, coin $B$ has probability $2/3$ to come up HEAD. We chose a coin at random and got HEAD. What is the probability that we chose coin $A$?
Conditional probability

We have two coins, coin $A$ is a fair coin, coin $B$ has probability $2/3$ to come up HEAD. We chose a coin at random and got HEAD. What is the probability that we chose coin $A$?

$E_1 =$ the event ”Chosen coin $A$”.

$E_2 =$ the event ”outcome is HEAD”.

The conditional probability that we chose coin $A$ given that the outcome is HEAD is denoted

$$Pr(E_1 \mid E_2).$$
Conditional probability

- Given that the test is positive for a randomly selected individual, what is the probability that said individual has cancer?

- Given that we are in region B, what is the probability that we are in region AB?
If we make region B our new Universe, what is the probability of A?

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

\[ = \frac{|A \cap B|/|U|}{|B|/|U|} \]

\[ = \frac{P(A \cap B)}{P(B)} \]
Calculating conditional probabilities

Definition

The **conditional probability** that event $E_1$ occurs given that event $E_2$ occurs is

$$\Pr(E_1 \mid E_2) = \frac{\Pr(E_1 \cap E_2)}{\Pr(E_2)}.$$

The conditional probability is only well-defined if $\Pr(E_2) > 0$.

By conditioning on $E_2$ we restrict the sample space to the set $E_2$. Thus we are interested in $\Pr(E_1 \cap E_2)$ “normalized” by $\Pr(E_2)$. 
Rolling dice

Roll a die twice and consider the two events
A = \{ \text{sum is even} \},
B = \{ \text{2}\textsuperscript{nd} outcome is even} \}.
What is P(A), P(B), P(A|B)?
Example: the unfair coin

We are given 2 coins. One is a fair coin $A$, the other coin, $B$ has probability $2/3$ for HEAD.
We choose a coin at random, i.e. each coin is chosen with probability $1/2$.
Given that we got head, what is the probability that we chose the fair coin $A$???
Define a sample space of ordered pairs \((\text{coin, outcome})\). The sample space has four points

\[
\{(A, h), (A, t), (B, h), (B, t)\}
\]

\[
Pr((A, h)) = Pr((A, t)) = 1/4
\]

\[
Pr((B, h)) = 1/2 * 2/3 = 1/3
\]

\[
Pr((B, t)) = 1/2 * 1/3 = 1/6
\]

Define two events:

\(E_1 = \text{“Chose coin } A\”\).

\(E_2 = \text{“Outcome is head”}\).

\[
Pr(E_1 \mid E_2) = \frac{Pr(E_1 \cap E_2)}{Pr(E_2)} = \frac{1/4}{1/4 + 1/3} = 3/7.
\]
Independent events

**Definition**

Two events $E$ and $F$ are independent if and only if

$$\Pr(E \cap F) = \Pr(E) \cdot \Pr(F).$$

More generally, events $E_1, E_2, \ldots, E_k$ are mutually independent if and only if for any subset $I \subseteq [1, k],$

$$\Pr\left(\bigcap_{i \in I} E_i\right) = \prod_{i \in I} \Pr(E_i).$$
Pop quiz!

A fair coin was tossed 10 times and came up heads every time. What is the probability that it will come up tails next time?
• Some gotchas.
Monty Hall

• Suppose you’re on a game show, and you’re given the choice of three doors: Behind one door is a car; behind the others, goats.

• You pick a door, say No. 1, and the host, who knows what’s behind the doors, opens another door, say No. 3, which has a goat.

• He then says to you, “Do you want to pick door No. 2?”

• Is it to your advantage to switch your choice?

Simpson’s Paradox

- When a relationship is reversed at a higher level of data aggregation compared with the lower level.
In 1973, the University of California-Berkeley was sued for sex discrimination: they had accepted 44% of male applicants and only 35% of female applicants.

But.. the pooled and corrected data showed a "small but statistically significant bias in favor of women."

Theorem (Law of Total Probability)

Let $E_1, E_2, \ldots, E_n$ be mutually disjoint events in the sample space $\Omega$, and $\bigcup_{i=1}^{n} E_i = \Omega$, then

$$
\Pr(B) = \sum_{i=1}^{n} \Pr(B \cap E_i) = \sum_{i=1}^{n} \Pr(B \mid E_i) \Pr(E_i).
$$
Conditional probability

We have two coins, coin $A$ is a fair coin, coin $B$ has probability $2/3$ to come up HEAD. We chose a coin at random and got HEAD. What is the probability that we chose coin $A$?

$E_1 =$ the event ”Chosen coin $A$”.
$E_2 =$ the event ”outcome is HEAD”.

The conditional probability that we chose coin $A$ given that the outcome is HEAD is denoted

What is $P(E_2)$?
Conditional probability

• We looked at: Given that the test is positive for a randomly selected individual, what is the probability that said individual has cancer?

\[ P(A \mid B) = \frac{|A \cap B|}{|B|} \]

\[ = \frac{|A \cap B| / |U|}{|B| / |U|} \]

\[ = \frac{P(A \cap B)}{P(B)} \]
Conditional probability

- How about: Given that a randomly selected individual has cancer (event A), what is the probability that the test is positive for that individual (event AB)?
Theorem (Bayes’ Law)

Assume that $E_1, E_2, \ldots, E_n$ are mutually disjoint sets such that $\bigcup_{i=1}^n E_i = \Omega$, then

$$
\Pr(E_j \mid B) = \frac{\Pr(E_j \cap B)}{\Pr(B)} = \frac{\Pr(B \mid E_j) \Pr(E_j)}{\sum_{i=1}^n \Pr(B \mid E_i) \Pr(E_i)}.
$$
Bayes

Likelihood
Probability of collecting this data when our hypothesis is true

\[ P(H|D) = \frac{P(D|H) \cdot P(H)}{P(D)} \]

Prior
The probability of the hypothesis being true before collecting data

Posterior
The probability of our hypothesis being true given the data collected

Marginal
What is the probability of collecting this data under all possible hypotheses?
Drug testing example

• Assume .4% of the Texas population uses marijuana

• Drug test: 99% true positive results for drug users; 99% true negative results for non-users

• If a randomly selected individual is tested positive, what is the probability he or she is a user?
Drug testing example

\[ P(\text{User}|+) = \frac{P(+|\text{User})P(\text{User})}{P(+)} \]

\[ = \frac{P(+|\text{User})P(\text{User})}{P(+|\text{User})P(\text{User}) + P(+|\text{!User})P(\text{!User})} \]

\[ = \frac{0.99 \times 0.004}{0.99 \times 0.004 + 0.01 \times 0.996} \]

\[ = 28.4\% \]