Today: Probability Review

• The big picture
• Events and Event spaces
• Random variables
• Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
• Structural properties, e.g., Independence, conditional independence
• Maximum Likelihood Estimation
The Big Picture

Model i.e. Data generating process

Observed Data

Estimation / learning / Inference / Data mining

Probability
Probability

- Counting
- Basics of probability
- Conditional probability
- Random variables
- Discrete and continuous distributions
- Expectation and variance
- Tail bounds and central limit theorem
- ……
Statistics

- Maximum likelihood estimation
- Bayesian estimation
- Hypothesis testing
- Linear regression
- [Machine learning]
- ……
Probability as frequency

• Consider the following questions:
  
  – 1. What is the probability that when I flip a coin it is “heads”? **We can count \( \sim 1/2 \)**
  
  – 2. Why?

  – 3. What is the probability of Blue Ridge Mountains to have an erupting volcano in the near future? **could not count**

**Message:** *The frequentist view is very useful, but it seems that we can also use domain knowledge to come up with probabilities.*

Adapt from Prof. Nando de Freitas’s review slides
Probability as a measure of uncertainty

• Imagine we are throwing darts at a wall of size 1x1 and that all darts are guaranteed to fall within this 1x1 wall.

• What is the probability that a dart will hit the shaded area?

Adapt from Prof. Nando de Freitas’s review slides
Probability as a measure of uncertainty

• Probability is a measure of certainty of an event taking place.

• i.e. in the example, we were measuring the chances of hitting the shaded area.

\[
\text{prob} = \frac{\# \text{Red Boxes}}{\# \text{Boxes}}
\]

Its area is 1

Adapt from Prof. Nando de Freitas’s review slides
Today: Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation
Probability

*Probability* is the formal study of the laws of chance. Probability allows us to *manage uncertainty*.

The *sample space* is the set of all outcomes. For example, for a die we have 6 outcomes:

\[ O_{\text{die}} = \{1,2,3,4,5,6\} \]

**O:**

Elementary Event “Throw 2”

The elements of **O** are called *elementary events.*
Probability

• *Probability allows us to measure many events.*

• **The events are subsets of the sample space O.** For example, for a die we may consider the following events: e.g.,

  \[
  \text{GREATER} = \{5, 6\} \\
  \text{EVEN} = \{2, 4, 6\}
  \]

• **Assign probabilities to these events:** e.g.,

  \[
  P(\text{EVEN}) = 1/2
  \]
Sample space and Events

- **O**: Sample Space,
  - result of an experiment / set of all outcomes
  - If you toss a coin *twice* $O = \{HH,HT,TH,TT\}$

- **Event**: a subset of $O$
  - First toss is head $= \{HH,HT\}$

- **S**: event space, a set of events:
  - Contains the empty event and $O$
Axioms for Probability

• Defined over \((O,S)\) s.t.
  • \(1 \geq P(a) \geq 0\) for all \(a\) in \(S\)
  • \(P(O) = 1\)

• If \(A, B\) are disjoint, then
  • \(P(A \cup B) = p(A) + p(B)\)
Axioms for Probability

\[ P(O) = \sum P(B_i) \]
OR operation for Probability

• We can deduce other axioms from the above ones
  • Ex: \( P(A \cup B) \) for non-disjoint events

\[
P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)
\]
Theorems from the Axioms

- \( 0 \leq P(A) \leq 1 \),
- \( P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B) \)

From these we can prove:

\[
P(\text{not } A) = P(\sim A) = 1 - P(A)
\]
Another important theorem

- $0 \leq P(A) \leq 1$,
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

$$P(A) = P(A \cap B) + P(A \cap \sim B)$$

$P(\text{ Intersection of A and B})$
Another important theorem

- $0 \leq P(A) \leq 1$,
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

From these we can prove:

\[
P(A) = P(A \cap B) + P(A \cap \neg B)
\]
Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation
From **Events to Random Variable**

- Concise way of specifying attributes of outcomes
- Modeling students (Grade and Intelligence):
  - \( O = \) all possible students *(sample space)*
  - What are events *(subset of sample space)*
    - Grade_A = all students with grade A
    - Grade_B = all students with grade B
    - HardWorking,Yes = … who works hard
  - Very cumbersome

- Need “functions” that maps from \( O \) to an attribute space \( T \).
  - \( P(H = \text{YES}) = P(\{\text{student} \in O : H(\text{student}) = \text{YES}\}) \)
Random Variables (RV)

O

H: hardworking

Yes
No

G: Grade

A
B
A+

P(H = Yes) = P( \{all students who is working hard on the course\})

• “functions” that maps from O to an attribute space T.
Notations

• P(A) is shorthand for P(A=true)
• P(~A) is shorthand for P(A=false)
• Same notation applies to other binary RVs: P(Gender=M), P(Gender=F)
• Same notation applies to multivalued RVs: P(Major=history), P(Age=19), P(Q=c)
• Note: upper case letters/names for variables, lower case letters/names for values
Discrete Random Variables

• Random variables (RVs) which may take on only a \textbf{countable} number of \textbf{distinct} values

• X is a RV with arity \( k \) if it \textbf{can take on exactly one value} out of \( \{x_1, ..., x_k\} \)
Probability of Discrete RV

• Probability mass function (pmf): $P(X = x_i)$

• Easy facts about pmf
  - $\sum_i P(X = x_i) = 1$
  - $P(X = x_i \cap X = x_j) = 0$ if $i \neq j$
  - $P(X = x_i \cup X = x_j) = P(X = x_i) + P(X = x_j)$ if $i \neq j$
  - $P(X = x_1 \cup X = x_2 \cup \ldots \cup X = x_k) = 1$
e.g. Coin Flips

• You flip a coin
  – Head with probability $p$, e.g. $=0.5$

• You flip a coin for $k$, e.g., $=100$ times
  – How many heads would you expect
e.g. Coin Flips cont.

- You flip a coin
  - Head with probability $p$
  - Binary random variable
  - Bernoulli trial with success probability $p$

- You flip a coin for $k$ times
  - How many heads would you expect
  - Number of heads $X$ is a discrete random variable
  - Binomial distribution with parameters $k$ and $p$
Discrete Random Variables

• Random variables (RVs) which may take on only a **countable** number of **distinct** values
  – E.g. the total number of heads $X$ you get if you flip 100 coins

• $X$ is a RV with arity $k$ if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  – E.g. the possible values that $X$ can take on are 0, 1, 2,..., 100
e.g., two Common Distributions

• **Uniform** \( X \sim U[1, ..., N] \)
  - \( X \) takes values 1, 2, ..., \( N \)
  - \( P(X = i) = \frac{1}{N} \)
  - E.g. picking balls of different colors from a box

• **Binomial** \( X \sim Bin(k, p) \)
  - \( X \) takes values 0, 1, ..., \( k \)
  - \( P(X = i) = \binom{k}{i} p^i (1 - p)^{k-i} \)
  - E.g. coin flips \( k \) times
Today: Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties
  - Independence, conditional independence
Conditional / Joint / Marginal Probability

\[ P(A \text{ given } B) = \frac{P(A \text{ and } B)}{P(B)} \]

That is, in the frequentist interpretation, we calculate the ratio of the number of times both A and B occurred and divide it by the number of times B occurred.

For short we write: \( P(A|B) = \frac{P(AB)}{P(B)} \); or \( P(AB) = P(A|B)P(B) \), where \( P(A|B) \) is the **conditional** probability, \( P(AB) \) is the **joint**, and \( P(B) \) is the **marginal**.

If we have more events, we use the chain rule:

\[ P(ABC) = P(A|BC) P(B|C) P(C) \]

from Prof. Nando de Freitas’s review
If hard to directly estimate from data, most likely we can estimate

- 1. Joint probability
  - Use Chain Rule

- 2. Marginal probability
  - Use the total law of probability

- 3. Conditional probability
  - Use the Bayes Rule
(1). To calculate **Joint Probability:**

**Use Chain Rule**

- Two ways to use chain rules on joint probability

\[
P(A,B) = p(B | A)p(A)
\]

\[
P(A,B) = p(A | B)p(B)
\]
(2). To calculate **Marginal Probability**:

Use Rule of total probability (Event)

\[
p(A) = \sum P(B_i) P(A | B_i)
\]

\[
p(A) = p(A \cap B) + p(A \cap \neg B)
\]

\[
p(B_i \cap A)
\]

WHY ???
(2). To calculate Marginal Probability:
Use Rule of total probability (RV)

- Given two discrete RVs $X$ and $Y$, which take values in $\{x_1, \ldots, x_k\}$ and $\{y_1, \ldots, y_m\}$, We have

\[
P(X = x_i) = \sum_j P(X = x_i \cap Y = y_j) \\
= \sum_j P(X = x_i \mid Y = y_j)P(Y = y_j)
\]

\[
P(A) = P(A \cap B) + P(A \cap \sim B)
\]
(3). To calculate **Conditional Probability**: Use Bayes Rule

- \( P(X = x | Y = y) \) is the probability of \( X = x \), given the occurrence of \( Y = y \)

\[
P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}
\]
Bayes Rule

• X and Y are discrete RVs...

\[
P(\mathbb{X} = x \mid \mathcal{Y} = y) = \frac{P(\mathbb{X} = x \cap \mathcal{Y} = y)}{P(\mathcal{Y} = y)}
\]

\[
P(\mathbb{X} = x_i \mid \mathcal{Y} = y_j) = \frac{P(\mathcal{Y} = y_j \mid \mathbb{X} = x_i)P(\mathbb{X} = x_i)}{\sum_{k} P(\mathcal{Y} = y_j \mid \mathbb{X} = x_k)P(\mathbb{X} = x_k)}
\]
Bayes Rule

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

This is Bayes Rule

More General Forms of Bayes Rule

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

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

\[ P(A = a_1 | B) = \frac{P(B | A = a_1)P(A = a_1)}{\sum_i P(B | A = a_i)P(A = a_i)} \]
One Example: Joint

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set \( \{r, r, r, b\} \). What is the probability of drawing 2 red balls in the first 2 tries?

\[
P(B_1 = r, B_2 = r) =
\]
One Example: Joint

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set \{r, r, r, b\}. What is the probability of drawing 2 red balls in the first 2 tries?

\[
P(B_1 = r, B_2 = r) = P(B_1 = r) \cdot P(B_2 = r | B_1 = r)
\]

\[
P(B_1 = r) = \frac{3}{4}
\]

\[
P(B_1 = b) = \frac{1}{4}
\]

\[
P(B_2 = r | B_1 = r) = \frac{2}{3}
\]

Adapt from Prof. Nando de Freitas’s review slides
One Example: Joint

Assume we have a dark box with 3 red balls and 1 blue ball. That is, we have the set \{r,r,r,b\}. What is the probability of drawing 2 red balls in the first 2 tries?

\[
P(B_1 = r, B_2 = r) = P(B_1 = r) \cdot P(B_2 = r | B_1 = r)
\]

\[
= \frac{3}{4} \times \frac{2}{3} = \frac{1}{2}
\]

Adapt from Prof. Nando de Freitas’s review slides
One Example: Marginal

What is the probability that the 2nd ball drawn from the set \{r, r, r, b\} will be red?

Using marginalization, \( P(B_2 = r) = P(B_2 = r, B_1 = r) \) 
\[ + P(B_2 = r, B_1 = b) \]
One Example: Marginal

What is the probability that the 2\textsuperscript{nd} ball drawn from the set \{r, r, r, b\} will be red?

Using marginalization, \( P(B_2 = r) = P(B_2 = r \land B_1 = r) + P(B_2 = r \land B_1 = b) \)

\[
= P(B_1 = r) P(B_2 = r | B_1 = r) + P(B_1 = b) P(B_2 = r | B_1 = b)
\]

\[
= \frac{3}{4} \times \frac{2}{3} + \frac{1}{4} \times 1
\]
One Example: Conditional

\[ P(B_1 = r \mid B_2 = r) \]

\[ = \frac{P(B_2 = r \mid B_1 = r) \cdot P(B_1 = r)}{P(B_2 = r)} \]

\[ \Rightarrow \text{ Last page} \]

\[ \downarrow \text{Last paper} \]
Simplify Notation: Conditional Probability

\[ P(X = x \mid Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)} \]

But we will always write it this way:

\[ P(x \mid y) = \frac{p(x, y)}{p(y)} \]

\[ P(\text{X=x true}) \rightarrow P(\text{X=x}) \rightarrow P(\text{x}) \]
Simplify Notation: Marginal

- We know $p(X, Y)$, what is $P(X=x)$?
- We can use the law of total probability, why?

\[
p(x) = \sum_y P(x, y) = \sum_y P(y)P(x \mid y)
\]

\[
p(x) = \sum_{y, z} P(x, y, z) = \sum_{z, y} P(y, z)P(x \mid y, z)
\]

\[
\sum_{y, z} p(y, z) = 1
\]
Simplify Notation: Conditional

- Bayes Rule
  \[ P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)} \]

- You can condition on more variables
  \[ P(x \mid y, z) = \frac{P(x \mid z)P(y \mid x, z)}{P(y \mid z)} \]
Simplify Notation: An Example

• We know that \( P(\text{rain}) = 0.5 \)

• If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

\[
P(\text{rain} \mid \text{wet}) = \frac{P(\text{rain})P(\text{wet} \mid \text{rain})}{P(\text{wet})}
\]
Simplify Notation: An Example

• We know that $P(\text{rain}) = 0.5$

• If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

$$P(\text{rain} \mid \text{wet}) = \frac{P(\text{rain})P(\text{wet} \mid \text{rain})}{P(\text{wet})}$$

$$P(\text{wet} \mid \text{rain})$$

$$P(x \mid y) = \frac{P(x)P(y \mid x)}{P(y)} = \frac{p(x,y)}{p(y)}$$
Simplify Notation: An Example

• We know that \( P(\text{rain}) = 0.5 \)
• If we also know that the grass is wet, then how this affects our belief about whether it rains or not?

\[
P(\text{rain} \mid \text{wet}) = \frac{P(\text{rain})P(\text{wet} \mid \text{rain})}{P(\text{wet})}
\]
Today : Probability Review

- The big picture
- Events and Event spaces
- Random variables
- Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
- Structural properties, e.g., Independence, conditional independence
- Maximum Likelihood Estimation
Independent RVs

• Intuition: X and Y are independent means that $X = x$ neither makes it more or less probable that $Y = y$

• Definition: X and Y are independent iff

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$
More on Independence

\[ P(X = x \cap Y = y) = P(X = x)P(Y = y) \]

\[ P(X = x \mid Y = y) = P(X = x) \]

\[ P(Y = y \mid X = x) = P(Y = y) \]

• E.g. no matter how many heads you get, your friend will not be affected, and vice versa
More on Independence

• X is independent of Y means that knowing Y does not change our belief about X. The following forms are equivalent:
  • \( P(X=x, Y=y) = P(X=x) \cdot P(Y=y) \)
  • \( P(X=x \mid Y=y) = P(X=x) \)

• The above should hold for all \( x_i, y_j \)
• It is symmetric and written as \( X \perp Y \)
Conditionally Independent RVs

• Intuition: X and Y are conditionally independent given Z means that once Z is known, the value of X does not add any additional information about Y.

• Definition: X and Y are conditionally independent given Z iff

\[
P(X = x \cap Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z)
\]

If holding for all \( x_i, y_j, z_k \) \( X \perp Y | Z \)
More on Conditional Independence

\[ P(X = x \cap Y = y | Z = z) = P(X = x | Z = z) P(Y = y | Z = z) \]

\[ P(X = x | Y = y, Z = z) = P(X = x | Z = z) \]

\[ P(Y = y | X = x, Z = z) = P(Y = y | Z = z) \]
Today Recap: Probability Review

• The big picture
• Events and Event spaces
• Random variables
• Joint probability, Marginalization, conditioning, chain rule, Bayes Rule, law of total probability, etc.
• Structural properties, e.g., Independence, conditional independence
• Maximum Likelihood Estimation (next class)
References

- Prof. Andrew Moore’s review tutorial
- Prof. Nando de Freitas’s review slides
- Prof. Carlos Guestrin recitation slides