Modeling uncertainty

- Necessary component of almost all data analysis
- Approaches to modeling uncertainty:
  - Fuzzy logic
  - Possibility theory
  - Rough sets
  - **Probability (focus in this course)**

Facts about the world:
- John is 6 feet tall.
- Mary is 7 feet tall.

*...can have different interpretations:*
- **Probability**: \( P(\text{Mary is taller than John}) = 1 \)
- **Fuzzy Logic**: John is 0.8 Tall, Mary is 0.9 Tall
Probability

- Probability theory (*some disagreement*)
  - Concerned with interpretation of probability
  - 17th century: Pascal and Fermat develop probability theory to analyze games of chance
- Probability calculus (*universal agreement*)
  - Concerned with manipulation of mathematical representations
  - 1933: Kolmogorov states axioms of modern probability

Probability basics

- Basic element: **Random variable**
  - Mapping from a property of objects to a variable that can take one of a set of possible values
  - $X$ refers to random variable; $x$ refers to a value of that random variable
- Types of random variables
  - Discrete RV has a finite set of possible values; Continuous RV can take any value within an interval
  - **Boolean**: e.g., Warning (is there a storm warning? = <yes, no>)
  - **Discrete**: e.g., Weather is one of <sunny,rainy,cloudy,snow>
  - **Continuous**: e.g., Temperature
Probability basics

• **Sample space (S)**
  – Set of all possible outcomes of an experiment

• **Event**
  – Any subset of *outcomes* contained in the sample space S
  – When events *A* and *B* have no outcomes in common they are said to be *mutually exclusive*

Examples

<table>
<thead>
<tr>
<th>Random variable(s)</th>
<th>Sample space</th>
</tr>
</thead>
<tbody>
<tr>
<td>One coin toss</td>
<td>H, T</td>
</tr>
<tr>
<td>Two coin tosses</td>
<td>HH, HT, TH, TT</td>
</tr>
<tr>
<td>Select one card</td>
<td>2♥, 2♠, ..., A♣ (52)</td>
</tr>
<tr>
<td>Play a chess game</td>
<td>Win, Lose, Draw</td>
</tr>
<tr>
<td>Inspect a part</td>
<td>Defective, OK</td>
</tr>
<tr>
<td>Cavity and toothache</td>
<td>TT, TF, FT, FF</td>
</tr>
</tbody>
</table>
Axioms of probability

- For a sample space S with possible events As, a function that associates real values with each event A is called a probability function if the following properties are satisfied:
  1. $0 \leq P(A) \leq 1$ for every A
  2. $P(S) = 1$
  3. $P(A_1 \lor A_2 \ldots \lor A_{n\in S}) = P(A_1) + P(A_2) + \ldots + P(A_n)$

  if $A_1, A_2, \ldots, A_n$ are pairwise mutually exclusive events

Implications of axioms

- For any events A, B
  - $P(A) = 1 - P(\neg A)$
  - $P(\text{true}) = 1$ and $P(\text{false}) = 0$
  - If A and B are mutually exclusive then $P(A \land B) = 0$
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Probability distribution

- **Probability distribution** (i.e., probability mass function or probability density function) specifies the probability of observing every possible value of a random variable.
  - Discrete
    - Denotes probability that $X$ will take on a particular value:
      \[ P(X = x) \]
  - Continuous
    - Probability of any particular point is 0, have to consider probability within an interval:
      \[ P(a < X < b) = \int_a^b p(x) \, dx \]

Joint probability

- **Joint probability distribution** for a set of random variables gives the probability of every atomic event on those random variables.
  - E.g., \( P(\text{Weather, Warning}) = \) a \( 4 \times 2 \) matrix of values:
    
    |       | sunny | rainy | cloudy | snow |
    |-------|-------|-------|--------|------|
    | warning = Y | 0.005 | 0.08  | 0.02   | 0.02 |
    | warning = N  | 0.415 | 0.12  | 0.31   | 0.03 |
  - Every question about events can be answered by the joint distribution
Conditional probability

- **Conditional** (or posterior) probability:
  - e.g., \( P(\text{warning}=Y \mid \text{snow}=T) = 0.4 \)
  - Complete conditional distributions specify conditional probability for all possible combinations of a set of RVs:
    \[
    P(\text{warning} \mid \text{snow}) = \{P(\text{warning} = Y \mid \text{snow} = T), P(\text{warning} = N \mid \text{snow} = T)\},
    \{P(\text{warning} = Y \mid \text{snow} = F), P(\text{warning} = N \mid \text{snow} = F)\}
    \]
  - If we know more, then we can update the probability by conditioning on more evidence
    - e.g., if Windy is also given then \( P(\text{warning} \mid \text{snow}, \text{windy}) = 0.5 \)

**Conditional probability**

- **Definition of conditional probability:**
  \[
P(A \mid B) = \frac{P(A \land B)}{P(B)} \quad \text{if } P(B) > 0
\]
- **Product rule** gives an alternative formulation:
  \[
P(A \land B) = P(A \mid B)P(B) = P(B \mid A)P(A)
\]
- **Bayes rule** uses the product rule:
  \[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}
\]
Example

• Conditional probability:

\[ P(A|B) = \frac{P(A \land B)}{P(B)} \quad \text{if } P(B) > 0 \]

• Example: What is \( P(\text{sunny} \mid \text{warning} = Y) \)?

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>rainy</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>warning = Y</td>
<td>0.005</td>
<td>0.08</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>warning = N</td>
<td>0.415</td>
<td>0.12</td>
<td>0.31</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Conditional probability

• **Chain rule** is derived by successive application of product rule:

\[
P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})P(X_1, \ldots, X_{n-1}) \\
= P(X_n|X_1, \ldots, X_{n-1})P(X_{n-1}|X_1, \ldots, X_{n-2})P(X_1, \ldots, X_{n-2}) \\
= \ldots \\
= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1})
\]
Marginal probability

- **Marginal** (or unconditional) probability corresponds to belief that event will occur regardless of conditioning events.

  $$P(A) = \sum_{b \in B} P(A, b)$$

- Marginalization:

  $$= \sum_{b \in B} P(A|b)P(b)$$

- Example: What is $P(\text{cloudy})$?

<table>
<thead>
<tr>
<th></th>
<th>sunny</th>
<th>rainy</th>
<th>cloudy</th>
<th>snow</th>
</tr>
</thead>
<tbody>
<tr>
<td>warning $= Y$</td>
<td>0.005</td>
<td>0.08</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>warning $= N$</td>
<td>0.415</td>
<td>0.12</td>
<td>0.31</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Independence

- **A and B are independent iff:**
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A)\ P(B)$
  - *Knowing B tells you nothing about A*

- **Examples**
  - Coin flip 1 and coin flip 2?
  - Weather and storm warning?
  - Weather and coin flip=$H$?
  - Weather and election?
Last time

- **KNN classifier**
  - Instance based learning
  - Learning: “memorizing” a dataset
  - Complexity of the hypothesis space: different values of K
- **Introduction to Statistics**
  - Basic Concepts: Random Variable, Events, Kolmogorov’s axioms
  - Probability distribution, joint and conditional

---

Conditional independence

- Two variables $A$ and $B$ are **conditionally** independent given $Z$ iff for all values of $A$, $B$, $Z$:
  \[ P(A, B | Z) = P(A | Z) P(B | Z) \]
- **Note**: independence does not imply conditional independence or vice versa
- A and $B$ are independent iff:
  - $P(A|B) = P(A)$ or $P(B|A) = P(B)$ or $P(A, B) = P(A) P(B)$
  - *How is this different?*
Example 1

**Conditional independence does not imply independence**

- Gender and lung cancer are not independent
  \[ P(C \mid G) \neq P(C) \]
- Gender and lung cancer are conditionally independent given smoking
  \[ P(C \mid G, S) = P(C \mid S) \]
- Why? Because gender indicates likelihood of smoking, and smoking causes cancer

Example 2

**Independence does not imply conditional independence**

- Sprinkler-on and raining are independent
  \[ P(S \mid R) = P(S) \]
- Sprinkler-on and raining are not conditionally independent given the grass is wet
  \[ P(S \mid R, W) \neq P(S \mid R) \]
- Why? Because once we know the grass is wet, if it’s not raining, then the explanation for the grass being wet has to be the sprinkler
Expectation

- Denotes the expected value or mean value of a random variable \(X\)
  \[
  E[X] = \sum_x x \cdot p(x)
  \]

- Discrete
  \[
  E[X] = \int_x x \cdot p(x) \, dx
  \]

- Continuous
  \[
  E[h(X)] = \sum_x h(x) \cdot p(x)
  \]

- Expectation of a function
  \[
  E[aX + b] = a \cdot E[X] + b
  \]
  \[
  E[X + Y] = E[X] + E[Y]
  \]

Example

- Let \(X\) be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
- \(X = \{0, 1, 2, 3\}\)
- \(P(X=0) = 1/8; P(X=1) = 3/8; P(X=2) = 3/8; P(X=3) = 1/8\)
- What is the expected value of \(X\), \(E[X]\)?
  \[
  E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})
  \]
  \[
  = \frac{3}{2}
  \]
Variance

- Denotes the expectation of the squared deviation of \( X \) from its mean
  \[
  \text{Var}(X) = E[(x - E[X])^2]
  \]
  \[
  = E[X^2] - (E[X])^2
  \]

- Standard deviation
  \[
  \sigma = \sqrt{\text{Var}(X)}
  \]

- Variance of a function
  \[
  \text{Var}(aX + b) = a^2 \cdot \text{Var}(X)
  \]

\[
\text{Var}(h(X)) = \sum_x (h(x) - E[h(x)])^2 \cdot p(x)
\]

Example

- Let \( X \) be a random variable that represents the number of heads which appear when a fair coin is tossed three times.
  \[
  E[X] = (0 \cdot \frac{1}{8}) + (1 \cdot \frac{3}{8}) + (2 \cdot \frac{3}{8}) + (3 \cdot \frac{1}{8})
  \]
  \[
  = \frac{3}{2}
  \]
- \( X = \{0, 1, 2, 3\} \)
- What is the variance of \( X \), \( \text{Var}(X) \)?

\[
\text{Var}(X) = \left( \left[ 0 - \frac{3}{2} \right]^2 \cdot \frac{1}{8} \right) + \left( \left[ 1 - \frac{3}{2} \right]^2 \cdot \frac{3}{8} \right) + \left( \left[ 2 - \frac{3}{2} \right]^2 \cdot \frac{3}{8} \right) + \left( \left[ 3 - \frac{3}{2} \right]^2 \cdot \frac{1}{8} \right)
\]

\[
= \left( \frac{9}{4} \cdot \frac{1}{8} \right) + \left( \frac{1}{4} \cdot \frac{3}{8} \right) + \left( \frac{1}{4} \cdot \frac{3}{8} \right) + \left( \frac{9}{4} \cdot \frac{1}{8} \right)
\]

\[
= \frac{3}{4}
\]
Example

- You flip a fair coin twice
  1. The first flip is heads
  2. The second flip is tails
  3. The two flips are not the same
- Are (1) and (2): independent? Conditionally independent? Neither?
- Good news!
- You will get a chance to think more about these concepts
  – See assignment 2

Common distributions

- Bernoulli
- Binomial
- Multinomial
- Poisson
- Normal
Bernoulli

- Binary variable (0/1) that takes the value of 1 with probability $p$
  - E.g., Outcome of a fair coin toss is Bernoulli with $p=0.5$
    $$P(x) = p^x(1-p)^{1-x}$$
    $$E[X] = 1(p) + 0(1 - p) = p$$
    $$Var(X) = E[X]^2 - (E[X])^2$$
    $$= 1^2(p) + 0^2(1 - p) - p^2$$
    $$= p(1 - p)$$

Binomial

- Describes the number of successful outcomes in $n$ independent Bernoulli($p$) trials
  - E.g., Number of heads in a sequence of 10 tosses of a fair coin is Binomial with $n=10$ and $p=0.5$
    $$P(x) = \binom{n}{x} p^x(1-p)^{n-x}$$
    $$E[X] = np$$
    $$Var[X] = np(1-p)$$
Multinomial

- Generalization of binomial to $k$ possible outcomes; outcome $i$ has probability $p_i$ of occurring
  - E.g., Number of {outs, singles, doubles, triples, homeruns} in a sequence of 10 times at bat is Multinomial
- Let $X_i$ denote the number of times the $i$-th outcome occurs in $n$ trials:
  \[
P(x_1, \ldots, x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}
  \]
  \[
  E[X_i] = np_i
  \]
  \[
  Var(X_i) = np_i(1 - p_i)
  \]

Poisson

- Describes the number of successful outcomes occurring in a fixed interval of time (or space) if the “successes” occur independently with a known average rate
  - E.g., Number of emergency calls to a service center per hour, when the average rate per hour is $\lambda=10$
  \[
P(x) = \frac{\lambda^x e^{-\lambda}}{x!}
  \]
  \[
  E[X] = \lambda
  \]
  \[
  Var[X] = \lambda
  \]
Normal (Gaussian)

- Important distribution that gives well-known bell shape
- Central limit theorem:
  - Distribution of the mean of \( n \) samples becomes normally distributed as \( n \uparrow \), regardless of the distribution of the underlying population

\[
P(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2}
\]

\[
E[X] = \mu
\]

\[
Var(X) = \sigma^2
\]

Multivariate RV

- A multivariate random variable \( \mathbf{X} \) is a set \( X_1, X_2, ..., X_p \) of random variables
- \textbf{Joint} density function: \( P(\mathbf{x})=P(x_1, x_2, ..., x_p) \)
- \textbf{Marginal} density function: the density of any subset of the complete set of variables, e.g.:

\[
P(x_1) = \sum_{x_2, x_3} p(x_1, x_2, x_3)
\]

- \textbf{Conditional} density function: the density of a subset conditioned on particular values of the others, e.g.:

\[
P(x_1|x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)}
\]