Probabilistic Reasoning
AI Class 9 (Ch. 13)

Today’s Class

• Probability theory
• Probability notation
• Bayesian inference
  • From the joint distribution
  • Using independence / factoring
  • From sources of evidence

Probabilistic inference: finding posterior probability for a proposition, given observed evidence.

– R&N 490

Today’s Class

We don’t (can’t!) know everything about most problems.

• Most problems are not:
  • Deterministic
  • Fully observable
• Or, we can’t calculate everything.
  • Continuous problem spaces

Probability lets us understand, quantify, and work with this uncertainty.

Bayesian Reasoning

• Posteriors and priors
• What is inference?
• What is uncertainty?
• When/why use probabilistic reasoning?
• What is induction?
• What is the probability of two independent events?
• Frequentist/objectivist/subjectivist assumptions

Sources of Uncertainty

• Uncertain inputs
  • Missing data
  • Noisy data
• Uncertain knowledge
  • >1 cause => >1 effect
  • Incomplete knowledge of conditions or effects
  • Incomplete knowledge of causality
  • Probabilistic effects

• Uncertain outputs
  • Default reasoning (even deduction) is uncertain
  • Abduction & induction inherently uncertain
  • Incomplete deductive inference can be uncertain

Probabilistic reasoning only gives probabilistic results (summarizes uncertainty from various sources)

Decision Making with Uncertainty

• Rational behavior: for each possible action,
  • Identify possible outcomes
  • Compute probability of each outcome
  • Compute utility of each outcome
  • “goodness” or “desirability” per some formally specified definition
  • Compute probability-weighted (expected) utility of possible outcomes for each action
  • Select the action with the highest expected utility (principle of Maximum Expected Utility)

Also the definition of “rational” for deterministic decision-making!
Probability Theory

- **World**: The complete set of possible states
- **Random variables**: Problem aspects that take a value
  - “The number of blue squares we are holding,” \( B \)
  - “The combined value of two dice we rolled,” \( C \)
- **Event**: Something that happens
- **Sample Space**: All the things (outcomes) that could happen in some set of circumstances
  - Pull 2 squares from envelope A: what is the sample space?
  - How about envelope B?
- **World, redux**: A complete assignment of values to variables

Basic Probability

- Each \( P \) is a non-negative value in \([0,1]\)
  - \( P(1,1) = 1/36 \)
- Total probability of the sample space is 1
  - \( P(1,1) + P(1,2) + P(1,3) + \ldots + P(6,6) = 1 \)
- For mutually exclusive events, the probability for at least one of them is the sum of their individual probabilities
  - \( P(\text{sunny}) \lor P(\text{cloudy}) = P(\text{sunny}) + P(\text{cloudy}) \)

Experimental probability: Based on frequency of past events
Subjective probability: Based on expert assessment

Why Probabilities Anyway?

3 simple axioms \( \Rightarrow \) all rules of probability theory*

1. All probabilities are between 0 and 1.
   - \( 0 \leq P(a) \leq 1 \)
2. Valid propositions (tautologies) have probability 1, and unsatisfiable propositions have probability 0.
   - \( P(\text{true}) = 1 \)
   - \( P(\text{false}) = 0 \)
3. The probability of a disjunction is:
   - \( P(a \lor b) = P(a) + P(b) - P(a \land b) \)

Compound Probabilities

- Describe **independent** events
  - Do not affect each other in any way
  - **Joint** probability of two independent events \( A \) and \( B \)
    \[ P(A \land B) = P(A) \times P(B) \]
  - **Union** probability of two independent events \( A \) and \( B \)
    \[ P(A \lor B) = P(A) + P(B) - P(A \land B) = P(A) + P(B) - (P(A) \times P(B)) \]

Pull two squares from envelope A. What is the probability that they are BOTH red?

Probability Distributions

- A distribution is the probabilities of all possible values of a random variable
- Ex: weather can be sunny, rainy, cloudy, or snowy
  - \( P(\text{Weather} = \text{sun}) = 0.6 \)
  - \( P(\text{Weather} = \text{rain}) = 0.1 \)
  - \( P(\text{Weather} = \text{cloud}) = 0.29 \)
  - \( P(\text{Weather} = \text{snow}) = 0.01 \)
  - \( P(\text{Weather}) = <0.6, 0.1, 0.29, 0.01> \)
- \( P(\text{Weather}) \): **probability distribution on Weather**
Probability Theory: Definitions

- **Conditional probability**: Probability of some effect given that we know cause(s)
  - Example: \( P(\text{alarm} | \text{burglary}) \)
  - (Technically, we only know \( b \) is true, not causal, but...)

- **Computing it**:
  - \( P(a | b) = \frac{P(a \land b)}{P(b)} \)

- **\( P(b) \)**: normalizing constant
  - (Later we’ll call this alpha)

---

Try It...

| P(A | B) | P(B | A) | P(B) |
|--------|--------|------|
| \( a \) | \( b \) | \( \alpha \) |

Example: Inference from the Joint

- **Product rule**:
  - \( P(a \land b) = P(a | b) P(b) \)

- **Marginalizing (summing out)**:
  - Finding distribution over one or a subset of variables
  - Marginal probability of \( B \) summed over all alarm states:
    - \( P(B) = \Sigma_{a} P(B | a) P(a) \)

- **Conditioning** over a subset of variables:
  - \( P(B | a) = \Sigma_{b} P(B | a, b) P(b) \)

---

Exercise: Inference from the Joint

- **Queries**:
  - What is the prior probability of \( B \)?
  - What is the conditional probability of \( B \), given \( A \)?
  - What is the conditional probability of \( A \), given \( B \)?

---

Example: Inference from the Joint

- **Queries**:
  - What is the prior probability of \( B \)?
  - What is the conditional probability of \( B \), given \( A \)?
  - What is the conditional probability of \( B \), given \( A \)?

---

Exercise: Inference from the Joint

- **Queries**:
  - What is the prior probability of \( B \)?
  - What is the conditional probability of \( B \), given \( A \)?
  - What is the conditional probability of \( A \), given \( B \)?

---

Exercise: Inference from the Joint

- **Queries**:
  - What is the prior probability of \( B \)?
  - What is the conditional probability of \( B \), given \( A \)?
  - What is the conditional probability of \( B \), given \( A \)?
Exercise: Inference from the joint

<table>
<thead>
<tr>
<th></th>
<th>P(smart &amp; study &amp; prep)</th>
<th>smart</th>
<th>~smart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>study &amp; ~study &amp; study</td>
<td>study</td>
<td>~study</td>
</tr>
<tr>
<td>prepared</td>
<td>.432</td>
<td>.16</td>
<td>.084</td>
</tr>
<tr>
<td>~prepared</td>
<td>.048</td>
<td>.16</td>
<td>.036</td>
</tr>
</tbody>
</table>

Queries:
- What is the prior probability of smart?
- What is the prior probability of study?
- What is the conditional probability of prepared, given study and smart?

\[ P(\text{study}) = .432 + .048 + .084 + .036 = 0.6 \]

Independence: \( \perp \)

- **Independent**: Two sets of propositions that do not affect each others’ probabilities
- Easy to calculate joint and conditional probability of independence:
  \( (A, B) \Rightarrow P(A \land B) = P(A) \cdot P(B) \text{ or } P(A \mid B) = P(A) \)
- Examples:
  - \( A = \text{alarm} \)
  - \( M = \text{moon phase} \)
  - \( B = \text{burglary} \)
  - \( L = \text{light level} \)
  - \( E = \text{earthquake} \)
  - \( A \perp B \perp E = f \)
  - \( M \perp L = f \)
  - \( A \perp M = t \)

Independent Example

- \( \{\text{moon-phase, light-level}\} \perp \{\text{burglary, alarm, earthquake}\} \)
  - But maybe burglaries increase in low light
  - But, if we know the light level, moon-phase \( \perp \) burglary
  - Once we’re burglarized, light level doesn’t affect whether the alarm goes off; \( \{\text{light-level}\} \perp \{\text{alarm}\} \)
- We need:
  1. A more complex notion of independence
  2. Methods for reasoning about these kinds of (common) relationships

Exercise: Independence

<table>
<thead>
<tr>
<th></th>
<th>P(smart &amp; study &amp; prep)</th>
<th>smart</th>
<th>~smart</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>study &amp; ~study &amp; study</td>
<td>study</td>
<td>~study</td>
</tr>
<tr>
<td>prepared</td>
<td>.432</td>
<td>.16</td>
<td>.084</td>
</tr>
<tr>
<td>~prepared</td>
<td>.048</td>
<td>.16</td>
<td>.036</td>
</tr>
</tbody>
</table>

Queries:
- Is smart independent of study?
  - \( P(\text{smart} \mid \text{study}) = P(\text{smart}) \)
- Is prepared independent?
  - \( P(\text{prepared} \mid \text{study}) = P(\text{prepared}) \)

Conditional Probabilities

- Describes dependent events
  - Affect each other in some way
  - Typical in the real world
  - If we know some event has occurred, what does that tell us about the likelihood of another event?
Conditional Independence

- **moon-phase and burglary** are conditionally independent given light-level
- That is, \( M \perp B \) if we already know \( L \)
- Conditional independence is:
  - Weaker than absolute independence
  - Useful in decomposing full joint probability distributions

Absolute independence: \( A \perp B \), if:
- \( P(A \land B) = P(A)P(B) \)
- Equivalently, \( P(A | B) = P(A) \) and \( P(B | A) = P(B) \)
- \( A \) and \( B \) are conditionally independent given \( C \) if:
  - \( P(A \land B | C) = P(A | C)P(B | C) \)
- This lets us decompose the joint distribution:
  - \( P(A \land B \land C) = P(A | C)P(B | C)P(C) \)
- What does this mean?

Exercise: Conditional Independence

<table>
<thead>
<tr>
<th>( P(\text{smart} \land \text{study} \land \text{prep}) )</th>
<th>smart study</th>
<th>~smart study</th>
</tr>
</thead>
<tbody>
<tr>
<td>prepared</td>
<td>.432</td>
<td>.16</td>
</tr>
<tr>
<td>~prepared</td>
<td>.048</td>
<td>.16</td>
</tr>
</tbody>
</table>

- Queries:
  - Is \( \text{smart} \) conditionally independent of \( \text{prepared} \), given \( \text{study} \)?
  - Is \( \text{study} \) conditionally independent of \( \text{prepared} \), given \( \text{smart} \)?

Naïve Bayes Algorithm

- Estimate the probability of each class:
  - Compute the posterior probability (Bayes rule)
    \[
    P(C_i | D) = \frac{P(C_i)P(D | C_i)}{P(D)}
    \]
  - Choose the class with the highest probability
  - Assumption of attribute independency (Naïve assumption): Naïve Bayes assumes that all of the attributes are independent.

Bayesian Inference

- In the setting of diagnostic/evidential reasoning
  - Know: prior probability of hypothesis \( H \)
  - Want to compute the posterior probability \( P(H | E_i) \)
  - Bayes' theorem (formula 1):
    \[
    P(H | E_i) = \frac{P(H)P(E_i | H)}{P(E_i)}
    \]
Simple Bayesian Diagnostic Reasoning

• Knowledge base:
  - Evidence / manifestations: \( E_1, \ldots, E_m \)
  - Hypotheses / disorders: \( H_1, \ldots, H_n \)
  - \( E_j \) and \( H_i \) are binary; hypotheses are mutually exclusive (non-overlapping) and exhaustive (cover all possible cases)
  - Conditional probabilities: \( P(E_j | H_i), i = 1, \ldots, n; j = 1, \ldots, m \)
• Cases (evidence for a particular instance): \( E_1, \ldots, E_m \)
• Goal: Find the hypothesis \( H_i \) with the highest posterior
  - \( \max_i P(H_i | E_1, \ldots, E_m) \)

Priors

• Four values total here:
  - \( P(H | E) = (P(E | H) \cdot P(H)) / P(E) \)
  - \( P(H | E) \) — what we want to compute
  - Three we already know, called the priors
    - \( P(E | H) \)
    - \( P(H) \)
    - \( P(E) \)
  - (In ML we use the training set to estimate the priors)

Bayesian Diagnostic Reasoning II

• Bayes’ rule says that
  - \( P(H_i | E_1, \ldots, E_m) = P(E_1, \ldots, E_m | H_i) \cdot P(H_i) / P(E_1, \ldots, E_m) \)
• Assume each piece of evidence \( E_i \) is conditionally independent of the others, given a hypothesis \( H_i \), then:
  - \( P(E_1, \ldots, E_m | H_i) = \prod_j P(E_j | H_i) \)
• If we only care about relative probabilities for the \( H_i \), then we have:
  - \( P(H_i | E_1, \ldots, E_m) = \alpha \cdot P(H_i) \prod_j P(E_j | H_i) \)

Bayes Example: Diagnosing Meningitis

• Your patient comes in with a stiff neck.
• Is it meningitis?
• Suppose we know that
  - Stiff neck is a symptom in 50% of meningitis cases
  - Meningitis (m) occurs in 1/50,000 patients
  - Stiff neck (s) occurs in 1/20 patients
• So probably not. But specifically?

• Stiff neck is a symptom in 50% of meningitis cases
• Meningitis (m) occurs in 1/50,000 patients
• Stiff neck (s) occurs in 1/20 patients
• Then
  - \( P(s | m) = 0.5, P(m) = 1/50000, P(s) = 1/20 \)
  - \( P(m | s) = (P(s | m) \cdot P(m)) / P(s) \)
    = \( (0.5 \times 1/50000) / 1/20 = .0002 \)
• So we expect that one in 5000 patients with a stiff neck to have meningitis.

Analysis of Naïve Bayes Algorithm

• Advantages:
  - Sound theoretical basis
  - Works well on numeric and textual data
  - Easy implementation and computation
  - Has been effective in practice (e.g., typical spam filter)
Limitations of Simple Bayesian Inference

- Cannot easily handle multi-fault situations, nor cases where intermediate (hidden) causes exist:
  - Disease D causes syndrome S, which causes correlated manifestations M_1 and M_2
- Consider a composite hypothesis H_1 \land H_2, where H_1 and H_2 are independent. What is the relative posterior?
  - P(H_1 \land H_2 | E_1, …, E_m) = \alpha P(E_1, …, E_m | H_1 \land H_2) P(H_1 \land H_2)
  - = \alpha \prod_{l=1}^{m} P(E_m | H_1 \land H_2) P(H_1) P(H_2)
- How do we compute P(E_j | H_1 \land H_2)?

Limitations of Simple Bayesian Inference II

- Assume H1 and H2 are independent, given E1, …, Ej?
- P(H_1 \land H_2 | E_1, …, E_j) = P(H_1 | E_1, …, E_j) P(H_2 | E_1, …, E_j)
- This is a very unreasonable assumption:
  - Earthquake and Burglar are independent, but not given Alarm:
    - P(Burglar | alarm, earthquake) < P(Burglar | alarm)
- Simple application of Bayes’ rule doesn’t handle causal chaining:
  - A: this year’s weather; B: cotton production; C: next year’s cotton price
  - A influences C indirectly: A \rightarrow B \rightarrow C
  - P(C | B, A) = P(C | B)
- Need a richer representation to model interacting hypotheses, conditional independence, and causal chaining
- Next time: conditional independence and Bayesian networks!