Probability in Machine Learning

• The world is uncertain
  – Often want a probability distribution over possible outcomes

• Bayesian Methods in ML
  – Practical, widely used (e.g. Naïve Bayes)
  – Naturally incorporate prior knowledge
Random Variables (1 of 2)

• A random variable is a variable whose value is the result of an experiment
  – A short-hand for referring to attributes of events

• E.g., your grade in this course
  – Atomic Events $\Omega = $ set of possible scores on hmwks and final
  – Define a random variable $Grade$
    • Deterministic function from $\Omega$ to \{A, B, C\}
Random Variables (2 of 2)

• Random variables will be in uppercase
• When r.v. clear from context, abbreviate
  – e.g. P(A)
• Val(X) = set of values r.v. X can take
  – Val(Grade) = {A, B, C}
• Conjunction
  – Rather than write P((Grade = A) \cap (Age = 21)), we use P(Grade = A, Age = 21) or just P(A, 21).
Probability: Interpretations & Motivation

• Probabilities assign numbers to events

• Axioms
  – \( P(a) \geq 0 \)
  – \( P(\text{union of all possible events}) = 1 \)
  – When events \( \alpha \cap \beta = \emptyset \) then \( P(\alpha \cup \beta) = P(\alpha) + P(\beta) \)

• Interpretations
  – Frequentist
  – Bayesian/subjective

• Why use probability for subjective beliefs?
  – Beliefs that violate the axioms can lead to bad decisions \textit{regardless} of the outcome [de Finetti, 1931]
Continuous Random Variables

- For continuous r.v. $X$, specify a density $p(x)$, such that:

$$P(r \leq X \leq s) = \int_{x=r}^{s} p(x) \, dx$$

E.g.,

$$p(x) = \begin{cases} \frac{1}{a - b} & a \geq x \geq b \\ 0 & \text{otherwise} \end{cases}$$
Gaussian Density

- \( p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x - \mu)^2}{2\sigma^2}\right) \)
Distributions

\[ P(\text{Intelligence}) \]

- High: 0.2
- Low: 0.8

\[ P(\text{Grade}) \]

- A: 0.3
- B: 0.4
- C: 0.4
Joint Distribution

\[ P(\text{Intelligence}, \text{Grade}) \]
Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Joint Distribution specified with $2 \times 3 - 1 = 5$ values
Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>

$P(\text{Grade} = A, \text{Intelligence} = \text{Low}) = 0.07$
Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
<td>0.18</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
<td>0.03</td>
</tr>
</tbody>
</table>

\[ P(\text{Grade} = A) = 0.07 + 0.18 = 0.25 \]
### Joint Distribution

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>0.07</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>High</td>
<td>0.18</td>
<td>0.09</td>
</tr>
<tr>
<td>A</td>
<td></td>
<td>0.28</td>
<td>0.09</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>C</td>
<td></td>
<td>0.03</td>
<td></td>
</tr>
</tbody>
</table>

**P(Grade = A ∨ Intelligence = High)?**

\[
0.07 + 0.18 + 0.09 + 0.03 = 0.37
\]

⇒ Given the joint distribution, we can compute probabilities for any proposition by summing events.
Conditional Probability

• $P(\text{Grade} = \text{A} \mid \text{Intelligence} = \text{High}) = 0.6$
  
  – the probability of getting an A given only $\text{Intelligence} = \text{High}$, and nothing else.

  • If we know $\text{Motivation} = \text{High}$ or $\text{OtherInterests} = \text{Many}$, the probability of an A changes even given high $\text{Intelligence}$

• Formal Definition:

  – $P(\alpha \mid \beta) = P(\alpha, \beta) / P(\beta)$

  • When $P(\beta) > 0$
Conditional Probability

• Also:
  – \( P(A \mid B, C) = \frac{P(A, B, C)}{P(B, C)} \)

• More generally:
  – \( P(A \mid B) = \frac{P(A, B)}{P(B)} \)
    – (Boldface indicates vectors of variables)

• \( P(Grade = A \mid Grade = A, Intelligence = \text{high}) \) ?
• \( P(CuriousGeorge \mid MonkeyWithVacuum, Cape) \)?
Conditional Probability

### Conditional Probability Table

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>A</td>
<td>0.07</td>
</tr>
<tr>
<td>B</td>
<td>0.28</td>
</tr>
<tr>
<td>C</td>
<td>0.35</td>
</tr>
</tbody>
</table>

P(Grade = A | Intelligence = High) ?

\[
P(\text{Grade} = \text{A}, \text{Intelligence} = \text{High}) = 0.18
\]

\[
P(\text{Intelligence} = \text{High}) = 0.18 + 0.09 + 0.03 = 0.30
\]

\[
\Rightarrow P(\text{Grade} = \text{A} \mid \text{Intelligence} = \text{High}) = \frac{0.18}{0.30} = 0.6
\]
Conditional Probability

| Grade | Intelligence | P(Intelligence | Grade = A) |
|-------|--------------|---------------|
| A     | Low          | 0.07          |
|       | High         | 0.18          |
| B     | Low          | 0.28          |
|       | High         | 0.09          |
| C     | Low          | 0.35          |
|       | High         | 0.03          |

P(Intelligence | Grade = A)?
### Conditional Probability

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>A</td>
<td>0.28</td>
<td>0.72</td>
</tr>
<tr>
<td>B</td>
<td>0.76</td>
<td>0.24</td>
</tr>
<tr>
<td>C</td>
<td>0.92</td>
<td>0.08</td>
</tr>
</tbody>
</table>

\[
P(\text{Intelligence} \mid \text{Grade})?
\]

Actually three separate distributions, one for each Grade value
(has three independent parameters total)
Chain Rule

\[ P(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i \mid X_{i-1} = x_{i-1}, \ldots, X_1 = x_1) \]

• E.g., \( P(\text{Grade}=B, \text{Int.} = \text{High}) \)
  \[ = P(\text{Grade}=B \mid \text{Int.} = \text{High})P(\text{Int.} = \text{High}) \]

• Can be used for distributions...
  – \( P(A, B) = P(A \mid B)P(B) \)
Handy Rules for Conditional Probability

- $P(A \mid B = b)$ is a single distribution, like $P(A)$
- $P(A \mid B)$ is not a single distribution
  - a set of $\mid \text{Val}(B)\mid$ distributions
- Any statement true for arbitrary distributions is also true if you condition on a new r.v.
  - $P(A, B) = P(A \mid B)P(B)? \text{ (chain rule)}$
    Then also $P(A, B \mid C) = P(A \mid B, C) P(B \mid C)$
- Likewise, any statement true for arbitrary distributions is also true if you replace an r.v. with two/more new r.v.s
  - $P(A \mid B) = P(A, B) / P(B) ? \text{ (def. of cond. Prob)}$
  - $P(A \mid C, D) = P(A, C, D) / P(C, D) \text{ or } P(A \mid B) = P(A, B) / P(B)$
Queries

• Given subsets of random variables $Y$ and $E$, and assignments $e$ to $E$
  – Find $P(Y | E = e)$

• Answering queries = inference
  – The whole point of probabilistic models, more or less
    – $P(Disease | Symptoms)$
    – $P(StockMarketCrash | RecentPriceActivity)$
    – $P(CodingRegion | DNASequence)$
    – ...(the other key task is learning)
### Answering Queries: Summing Out

<table>
<thead>
<tr>
<th>Grade</th>
<th>Intelligence = Low</th>
<th>Intelligence=High</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time=Lots</td>
<td>Time=Little</td>
</tr>
<tr>
<td>A</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>B</td>
<td>0.14</td>
<td>0.14</td>
</tr>
<tr>
<td>C</td>
<td>0.10</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[
P(\text{Grade} \mid \text{Time} = \text{Lots})? \]
\[
\sum_{v \in \text{Val}(\text{Intelligence})} P(\text{Grade}, \text{Intelligence} = v \mid \text{Time} = \text{Lots})
\]
Answering Queries: Solved?

• Given the joint distribution, we can answer any query by summing

• ...but, joint distribution of 500 Boolean variables has $2^{500} - 1$ parameters (about $10^{150}$)

• For non-trivial problems (~25 boolean r.v.s or more), using the joint distribution requires
  – Way too much **computation** to compute the sum
  – Way too many **observations** to learn the parameters
  – Way too much **space** to store the joint distribution
Conditional Independence (1 of 3)

• Independence
  – \( P(A, B) = P(A) \times P(B) \), denoted \( A \perp B \)
  – E.g. consecutive dice rolls
    • Gambler’s fallacy
  – Rare in (real) applications
Conditional Independence (2 of 3)

• Conditional Independence
  – $P(A, B \mid C) = P(A \mid C) P(B \mid C)$, denoted $(A \perp B \mid C)$
  – Much more common
  – E.g.,
    $(GetIntoNU \perp GetIntoStanford \mid Application)$, but **NOT** $(GetIntoNU \perp GetIntoStanford)$
Conditional Independence (3 of 3)

- How does Conditional Independence save the day?

\[ P(NU, Stanford, App) = P(NU|Stanford, App) \times P(Stanford|App) \times P(App) \]

Now, \((A \perp B | C)\) means \(P(A | B, C) = P(A | C)\)

So since \((NU \perp Stanford | App)\), we have

\[ P(NU, Stanford, App) = P(NU | App) \times P(Stanford | App) \times P(App) \]

Say \(App \in \{\text{Good, Bad}\}\) and \(School \in \{\text{Yes, No, Wait}\}\)

All we need is \(4+4+1=9\) numbers

(vs. \(3\times3\times2-1=17\) for the full joint)

Full joint has size \textbf{exponential} in # of r.v.s

Conditional independence eliminates this!
Bayes’ Rule

- \( P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)} \)

- Example:

  \[
P(\text{symptom} \mid \text{disease}) = 0.95, \quad P(\text{symptom} \mid \neg \text{disease}) = 0.05
  \]

  \[
P(\text{disease} = 0.0001)
  \]

  \[
P(\text{disease} \mid \text{symptom}) = \frac{P(\text{symptom} \mid \text{disease}) \cdot P(\text{disease})}{P(\text{symptom})}
  \]

  \[
  = \frac{0.95 \cdot 0.0001}{0.95 \cdot 0.0001 + 0.05 \cdot 0.9999} = 0.002
  \]
Bayes’ Rule

• \( P(A \mid B) = P(B \mid A) \, P(A) / P(B) \)

• Also:
  - \( P(A \mid B, C) = P(B \mid A, C) \, P(A \mid C) / P(B \mid C) \)

• More generally:
  - \( P(\mathbf{A} \mid \mathbf{B}) = P(\mathbf{B} \mid \mathbf{A}) \, P(\mathbf{A}) / P(\mathbf{B}) \)
  - (Boldface indicates vectors of variables)
Terms for Bayes

\[ P(Model \mid Data) = \frac{P(Data \mid Model) \cdot P(Model)}{P(Data)} \]

- \( P(Model) \) : Prior
- \( P(Data \mid Model) \) : Likelihood
- \( P(Model \mid Data) \) : Posterior
What have we learned?

• Probability – a calculus for dealing with uncertainty
  – Built from small set of axioms (ignore at your peril)
• Joint Distribution $P(A, B, C, ...)$
  – Specifies probability of all combinations of r.v.s
  – Intractable to compute exhaustively for non-trivial problems
• Conditional Probability $P(A \mid B)$
  – Specifies probability of $A$ given $B$
• Conditional Independence
  – Can radically reduce number of variable combinations we must assign unique probabilities to.
• Bayes’ Rule