Basics of Probability

Lecture 1

Doug Downey, Northwestern EECS 395/495
Winter 2010
Events

• **Event space** $\Omega$
  - E.g. for dice, $\Omega = \{1, 2, 3, 4, 5, 6\}$

• **Set of measurable events** $S \subseteq 2^\Omega$
  - E.g.,
    
    $\alpha = \text{event we roll an even number} = \{2, 4, 6\} \in S$
  - $S$ must:
    
    • Contain the empty event $\emptyset$ and the trivial event $\Omega$
    • Be closed under union & complement
      
      $\alpha, \beta \in S \rightarrow \alpha \cup \beta \in S$ and $\alpha \in S \rightarrow \Omega - \alpha \in S$
Probability Distributions

A probability distribution $P$ over $(\Omega, S)$ is a mapping from $S$ to real values such that:

$P(\alpha) \geq 0$

$P(\Omega) = 1$

$\alpha, \beta \in S \land \alpha \cap \beta = \emptyset \rightarrow P(\alpha \cup \beta) = P(\alpha) + P(\beta)$
Probability Distributions

Can visualize probability as fraction of area
Probability: Interpretations & Motivation

• Interpretations
  – Frequentist
  – Bayesian/subjective

• Why use probability for subjective beliefs?
  – Beliefs that violate the axioms can lead to bad decisions regardless of the outcome [de Finetti, 1931]
  – Example: $P(A) = 0.6$, $P(\text{not } A) = 0.8$ ?
  – Example: $P(A) > P(B)$ and $P(B) > P(A)$ ?
Random Variables (1 of 2)

- A **random variable** is a function from $\Omega$ to a value
  - A short-hand for referring to *attributes* of events.
- E.g., your grade in this course
  - Let $\Omega =$ set of possible scores on hmwks and final
  - Cumbersome to have separate events GradeA, GradeB, GradeC
  - So instead define a random variable *Grade*
    - Deterministic function from $\Omega$ to \{A, B, C\}
Random Variables (2 of 2)

• Denote $P(\text{Grade} = A)$ as $P(\text{Grade} = A)$
  – Random variables will be in uppercase
  – When r.v. clear from context, abbreviate (e.g. $P(A)$)

• $\text{Val}(X)$ = set of values r.v. $X$ can take
  – $\text{Val}(\text{Grade}) = \{A, B, C\}$

• Conjunction
  – Rather than write $P((\text{Grade} = A) \cap (\text{Age} = 21))$, we use $P(\text{Grade} = A, \text{Age} = 21)$ or just $P(A, 21)$. 
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.,

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

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

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

• Called “marginal” because they apply to only one r.v.
Joint Distribution

$P(\text{Intelligence, Grade})$

- Grade=A
  - Intelligence=Low: 0.07
  - Intelligence=High: 0.18

- Grade=B
  - Intelligence=Low: 0.28
  - Intelligence=High: 0.09

- Grade=C
  - Intelligence=Low: 0.35
  - Intelligence=High: 0.03
# Joint Distribution

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

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

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

• $P(\text{Grade} = A \mid \text{Intelligence} = \text{High}) = 0.6$
  – 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:
  – $P(\alpha \mid \beta) = \frac{P(\alpha, \beta)}{P(\beta)}$
    • 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)

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

<table>
<thead>
<tr>
<th>Grade</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<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(Grade = A | Intelligence = High) ?

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

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

\[
=> P(Grade = A | Intelligence = High) = \frac{0.18}{0.30} = 0.6
\]
### 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>0.07</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>

$P(\text{Intelligence} \mid \text{Grade} = A)$?

<table>
<thead>
<tr>
<th>Intelligence</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>0.28</td>
</tr>
<tr>
<td>High</td>
<td>0.72</td>
</tr>
</tbody>
</table>
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 $|\text{Val}(B)|$ distributions
- Any statement true for arbitrary distributions is also true if you condition on a new r.v.
  - $P(A, B) = P(A \mid B)P(B)$? (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
  - $P(A \mid B) = P(A, B) / P(B)$? (def. of cond. Prob)
  - $P(A \mid C, D) = P(A, C, D) / P(C, D)$ 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 $P(Y \mid E = e)$

• Answering queries = inference
  – The whole point of probabilistic models, more or less
  – $P(Disease \mid Symptoms)$
  – $P(StockMarketCrash \mid RecentPriceActivity)$
  – $P(CodingRegion \mid DNASequence)$
  – $P(PlayTennis \mid Weather)$
  – ...(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>

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

• Given subsets of random variables $Y$ and $E$, and assignments $e$ to $E$
  – Find $\text{MAP}(Y \mid e) = \arg \max_y P(y \mid e)$

• MAP stands for “maximum a posteriori”
  – (more later)
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

Note: Book calls this “marginal independence” when applied to r.v.s, but just “independence” when applied to events
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.,
    - $(\text{GetIntoNU} \perp \text{GetIntoStanford} \mid \text{Application})$
    - but **NOT** $(\text{GetIntoNU} \perp \text{GetIntoStanford})$
Conditional Independence (3 of 3)

• How does Conditional Independence save the day?

\[
P(NU, Stanford, App) = P(NU \mid Stanford, App) \cdot P(Stanford \mid App) \cdot P(App)
\]

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

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

\[
P(NU, Stanford, App) = P(NU \mid App) \cdot P(Stanford \mid App) \cdot P(App)
\]

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

All we need is 4+4+1=9 numbers

(vs. \(3 \times 3 \times 2 - 1 = 17\) for the full joint)

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

Conditional independence eliminates this!
Properties of Conditional Independence

• Decomposition
  – \((X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z)\)

• Weak Union
  – \((X \perp Y, W \mid Z) \Rightarrow (X \perp Y \mid Z, W)\)

• Contraction
  – \((X \perp W \mid Z, Y) \land (X \perp Y \mid Z) \Rightarrow (X \perp Y, W \mid Z)\)
Bayes’ Rule

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

• Example:
Bayes’ Rule

• $P(A \mid B) = P(B \mid A) \frac{P(A)}{P(B)}$

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

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

- $P(\text{Model} \mid \text{Data}) = P(\text{Data} \mid \text{Model}) \cdot P(\text{Model}) / P(\text{Data})$

- $P(\text{Model})$ : Prior

- $P(\text{Data} \mid \text{Model})$ : Likelihood

- $P(\text{Model} \mid \text{Data})$ : Posterior
What have we learned?

• Probability – a calculus for dealing with uncertainty
  – Built from small set of axioms (ignore at your peril)

• Joint Distribution $P(A, B, C, \ldots)$
  – 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