Announcements

• HW3 due, Programming 3 out
• Quiz 3 Thursday
Part 3: Probability and Machine Learning

• Basic Probability and Statistics (Ch 13)
  – Note: this is meant to be a refresher rather than a mathematically rigorous introduction
  – Please review your old probability textbooks if too many cobwebs
Representing Uncertainty

• We have seen logical representations and how they can be used to solve problems

• But the real world is uncertain due to partial observability, cost of acquiring information, limits on computation etc

• How to represent/reason in this case?
Probability Theory

- A language that *augments* propositional logic with “degrees of belief,” and associated mechanics for reasoning in this augmented language

I think it is 60% likely that it will rain tomorrow.

\[ \text{RainTomorrow} = \text{true} \quad 60\% \]

(proposition) (degree of belief)
Random Variable (R.V.)

• A variable that refers to an uncertain fact
  – Analogous to proposition symbol
  – Has a domain that can be discrete or continuous
    • For this class, focus on discrete case

• For each value (or set of values), we can specify a *degree of belief* that shows how much we believe the stated fact---this is the *probability* associated with the fact
Example

• $\text{RainTomorrow} \in \{\text{True, False}\}$
  \[ \Pr(\text{RainTomorrow}=\text{True}) = 0.6 \]

• $\text{Current\_X\_Position} \in (-\infty, +\infty)$
  \[ \Pr(-1 \leq \text{Current\_X\_Position}) = 0.2 \]
Atomic Event

• If the state of the world is described by $n$ r.v.’s and we assign values to all of them, this defines an **atomic event**
  – (Analog of a row in a truth table)

• Example: suppose a footman is in a grid maze and is uncertain about an enemy archer’s $(x, y)$ location. Then $(x=2, y=3)$ could be an atomic event.
Events and the Sample Space

• Atomic events are **mutually exclusive and exhaustive**
  - At most one can be the true state of affairs
  - The true state of affairs must be one of them

• An “event” is a collection of atomic events
  - Example: the event \( \{x=2\} \) is the collection of atomic events \( \{(x=2, y=1), (x=2, y=2), (x=2, y=3), \ldots\} \)

• The “sample space” is the collection of all possible atomic events (\( \Omega \)) (“truth table”)

3/9/2015 Soumya Ray, Case Western Reserve U.
Joint Probability

• Just like we assign degrees of belief to single r.v.’s, we can do the same for groups of r.v.’s
  – \( \Pr(Rain\text{Tomorrow}=Yes, \ Cloudy\text{Tomorrow}=Yes) = 0.99 \)
  – \( \Pr(-1 \leq x, \ y \leq 1) = 0.2 \)
  – In particular, we can assign degrees of belief to atomic events

• This is called a “joint probability”
Axioms of Probability

• For any event $E$, $0 \leq \Pr(E) \leq 1$
• The probability of the sample space is 1
• For mutually disjoint events, the probability of the union is given by:

\[
\Pr(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} \Pr(E_i)
\]

In particular this must apply to atomic events.
Sample Space: Total “area” = 1

$Pr(E) = \text{area under } E$
Using the axioms

• Various other facts can be deduced from these axioms

• Suppose \( E \) is some event and \( \overline{E} \) is the event in \( \Omega \) that includes everything not in \( E \) (the “complement” of \( E \)). What is \( Pr(\overline{E}) \)?
Rationality and Probability Theory

• Could there be other ways of representing uncertainty?
  – Dempster-Shafer, “Fuzzy” logic, etc

• But probability theory has a major positive result: suppose someone’s degrees of belief for some set of events does NOT obey the axioms of probability. Then there is a way to bet against them such that they will always lose money (utility) over time. (Bruno de Finetti 1931)
Probability Density Functions

• Earlier we defined probabilities associated with r.v.’s: \( \Pr(\text{RainTomorrow}=\text{Yes})=0.8 \)

• A function that maps every value of an r.v. to a probability is called a probability density function (p.d.f.)

\[
\begin{align*}
P_{\text{RainTomorrow}}(x) &= \begin{cases} 
0.8 & \text{if } x = \text{Yes} \\
0.2 & \text{if } x = \text{No} 
\end{cases} \\
p_X(x) &= \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \ x \in \{-\infty, +\infty\}
\end{align*}
\]
PDFs must sum to 1

<table>
<thead>
<tr>
<th>Event $X_1 = x_{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event $X_1 = x_{12}$</td>
</tr>
<tr>
<td>$X_1$</td>
</tr>
<tr>
<td>$X_2$</td>
</tr>
<tr>
<td>Event $X_1 = x_{1n}$</td>
</tr>
</tbody>
</table>
Joint PDF

• Using joint probability, we can define joint density functions for collections of random variables

\[ p_{R,C}(R = x, C = y) = \begin{cases} 
0.5 & \text{if } x = \text{Yes}, y = \text{Yes} \\
0.2 & \text{if } x = \text{No}, y = \text{Yes} \\
0.2 & \text{if } x = \text{Yes}, y = \text{No} \\
0.1 & \text{if } x = \text{No}, y = \text{No} 
\end{cases} \]
Example

<table>
<thead>
<tr>
<th>CloudyTomorrow</th>
<th>RainTomorrow</th>
<th>WetGrass</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>0.4</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>0.01</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>0</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>0.01</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>0.15</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>0.02</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>0.01</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Joint Probability Density Function

Atomic Event

Event

Sample Space
First Order Logic

• Is it possible to augment FOL in the same way?
  – Hot topic of current research!
Terminology and Results
Conditional Probability

• The conditional probability of $X$ given $Y$ is:

$$p_{X|Y}(X = x | Y = y) = \frac{p_{X,Y}(X = x, Y = y)}{p_Y(Y = y)}$$

$X=x$, $Y=y$ (",” means AND)
Product Rule

• From the definition of conditional probability:

\[ p_{X,Y}(X = x, Y = y) = p_Y(Y = y) p_{X|Y}(X = x | Y = y) \]
Marginalization

• For any two random variables $X$ and $Y$:

$$p_X(X = x) = \sum_y p_{X,Y}(X = x, Y = y)$$

<table>
<thead>
<tr>
<th>$Y = y_1$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y = y_2$</td>
<td></td>
</tr>
<tr>
<td>$Y = y_3$</td>
<td></td>
</tr>
<tr>
<td>$X = x$</td>
<td></td>
</tr>
</tbody>
</table>
Conditioning

\[ p(X = x) = \sum_y p(X = x, Y = y) \]

= \sum_y p(X = x | Y = y) p(Y = y) \hspace{1cm} \text{Product Rule}

\hspace{1cm} \text{Marginalization}
Bayes’ Rule
(Rev. Thomas Bayes 1763)

\[ p(C = c \mid E = e) = \frac{p(C = c, E = e)}{p(E = e)} \]

\[ = \frac{p(E = e \mid C = c)p(C = c)}{p(E = e)} \]

\[ = \frac{\sum_{c'} p(E = e \mid C = c')p(C = c')}{\sum_{c'} p(E = e \mid C = c')p(C = c')} \]

Def. of Conditional Prob.
Product Rule
Conditioning
The importance of Bayes Rule

• Let $C$ be a random variable with values that are possible “causes”
• Let $E$ denote a random variable with values that are possible effects of each cause
• It is often easy to specify $p(E=e|C=c)$, much harder to specify $p(C=c|E=e)$
• Bayes Rule therefore allows us to reason backwards over uncertain events---fundamental to learning
Example

- Lung cancer can be caused by smoking or by a genetic defect. 5% of the population are smokers. 2 in 3 who smoke and 1 in 100 who don’t get the disease.

- Suppose X has lung cancer. What is the probability X smokes?
Example

\[ P(S) = 0.05, \ P(LC \mid S) = 0.67, \ P(LC \mid \bar{S}) = 0.01 \]

\[ P(S \mid LC) = \frac{P(LC \mid S)P(S)}{P(LC \mid S)P(S) + P(LC \mid \bar{S})P(\bar{S})} \]

\[ = \frac{0.67 \times 0.05}{0.67 \times 0.05 + 0.01 \times 0.95} = 0.78 \]
Statistical Independence

• Two r.v.'s $X$ and $Y$ are statistically independent if

$$p_{X,Y}(X = x, Y = y) = p_X(X = x)p_Y(Y = y)$$

• If so, we can reason separately about $x$ and $y$ and then combine results---key factor in gaining efficiency (later)
Consequence

\[ p_{X|Y}(X = x \mid Y = y) = \frac{p_{X,Y}(X = x, Y = y)}{p_Y(Y = y)} \]

\[ = \frac{p_X(X = x)p_Y(Y = y)}{p_Y(Y = y)} \]

\[ = p_X(X = x) \]
Summarizing a PDF

• A PDF is a large table of numbers

• But generally, we don’t need to know the entire thing; often the “highlights” are enough
  – *Expectation* and *Variance*
  – (statistics)
Expectation

• The expectation of r.v. $X$ is defined as:

$$E(X) = \sum_{x} x p_X(x)$$

• The “average value” of $X$ under $p_X(x)$
Expectation example

• A coin has 0.99 probability of showing heads. You get $0 if the coin shows heads, and $10 else. How much do you expect to get if I toss the coin?

\[
E(X) = \sum_{x} xp_{X}(x) = (0 \times 0.99 + 10 \times 0.01) = $0.1
\]
Variance

• The variance of r.v. $X$ is defined as:

$$V(X) = E([X - E(X)]^2)$$

$$= \sum_x (x - E(X))^2 p_X(x)$$

• The “average spread” of values of a r.v. around the average of the r.v.
Variance example

• A coin has 0.99 probability of showing heads. You get $0 if the coin shows heads, and $10 else. What is the variance of your takings?

\[
E(X) = \sum_{x} xp_{X}(x) = (0 \times 0.99 + 10 \times 0.01) = \$0.1
\]

\[
V(X) = E([X - E(X)]^2)
\]

\[
= (0 - 0.1)^2 \times 0.99 + (10 - 0.1)^2 \times 0.01
\]

\[
= 0.99
\]
Variance example

• A coin has 0.99 probability of showing heads. You get $10 if the coin shows heads, and $0 else. What is the variance of your takings?

\[
E(X) = \sum_x x p_x(x) = (10 \times 0.99 + 0 \times 0.01) = $9.9
\]

\[
V(X) = E(\left[ X - E(X) \right]^2)
\]

\[
= (10 - 9.9)^2 \times 0.99 + (0 - 9.9)^2 \times 0.01
\]

\[
= 0.99
\]
Variance example 3

• A coin has 0.5 probability of showing heads. You get $0 if the coin shows heads, and $10 else. What is the variance of your takings?

\[
E(X) = \sum_{x} x p_X(x) = (0 \cdot 0.5 + 10 \cdot 0.5) = $5
\]

\[
V(X) = E([X - E(X)]^2)
\]

\[
= (0 - 5)^2 \cdot 0.5 + (10 - 5)^2 \cdot 0.5
\]

\[
= 25
\]