Course Overview and Review of Probability

William W. Cohen
Machine Learning 10-601
OVERVIEW OF 601
SPRING 2016
SECTION B
• This is 10-601B (I’m William Cohen)
• Main information page: the class wiki
  – My home page → teaching
• Those pages
  – have all the lecture slides and homeworks
  – link to everything else we’ll be using
    • eg, Piazza, Autolab, MediaTech
What the course is about…

• Normal computer science:
  1. Human figures out how to do something
  2. Human programs the computer to do it

• Machine learning
  1. Somebody (maybe humans) does something … usually many times
  2. A computer learns how to do it from data

  – Classification learning – one kind of ML:
    • Task: label an object with one of a fixed set of labels
    • Examples: …
Classification learning

*classes*: 0, 1, 2, …, 9

5 \rightarrow \text{“5”}

*classes*: left, right, up, down, forward

f \rightarrow \text{“right”}

*classes*: noface, left, right, up, down, forward

“fwd”
Supervised Classification. Many other examples

- Weather prediction

- Medicine:
  - diagnose a disease
    - input: from symptoms, lab measurements, test results, DNA tests, …
    - output: one of set of possible diseases, or “none of the above”
    - examples: audiology, thyroid cancer, diabetes, …
      - or: response to chemo drug X
      - or: will patient be re-admitted soon?

- Computational Economics:
  - predict if a stock will rise or fall
  - predict if a user will click on an ad or not
    - in order to decide which ad to show
Classification learning

Classes: 1, 2, ..., 9

The function $f$ takes an input and produces an output.

$f$ (input) = output

Input: usually a vector

For example:

$= <p_{11}, p_{12}, ..., p_{1N}, p_{21}, ..., p_{2N}, ..., p_{NN}>$

$= <0, 0, 0, ..., 0, 1, 1, 0, 1, 1, 0, ..., 0, 0, 0>$
### Supervised Classification. Many other examples

- **Weather prediction**

![Weather Map Image]

#### Simple Training Data Set

<table>
<thead>
<tr>
<th>Day</th>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Wind</th>
<th>Play Tennis</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D2</td>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D3</td>
<td>Overcast</td>
<td>Hot</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D4</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D5</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D6</td>
<td>Rain</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>No</td>
</tr>
<tr>
<td>D7</td>
<td>Overcast</td>
<td>Cool</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D8</td>
<td>Sunny</td>
<td>Mild</td>
<td>High</td>
<td>Weak</td>
<td>No</td>
</tr>
<tr>
<td>D9</td>
<td>Sunny</td>
<td>Cool</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D10</td>
<td>Rain</td>
<td>Mild</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D11</td>
<td>Sunny</td>
<td>Mild</td>
<td>Normal</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D12</td>
<td>Overcast</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>Yes</td>
</tr>
<tr>
<td>D13</td>
<td>Overcast</td>
<td>Hot</td>
<td>Normal</td>
<td>Weak</td>
<td>Yes</td>
</tr>
<tr>
<td>D14</td>
<td>Rain</td>
<td>Mild</td>
<td>High</td>
<td>Strong</td>
<td>No</td>
</tr>
</tbody>
</table>

![Label Image]
Supervised Classification. Many other examples

- Weather prediction

\[ f: \langle \text{Outlook, Temperature, Humidity, Wind} \rangle \rightarrow \text{PlayTennis?} \]
What the course is about…

• Normal computer science:
  1. Human figures out how to do something
  2. Human programs the computer to do it

• Machine learning
  1. Somebody (maybe humans) does something … usually many times
  2. A computer learns how to do it from data

– Classification learning – one kind of ML:
  • Task: label an object with one of a fixed set of labels
  • Examples: learning to label feature vectors with a decision tree using TDIDT
What the course is about...

Crucial background
Supervised classification learning ... and regression (predicting numeric values)
Different learning algorithms, and different approaches and ways of analyzing performance

Why do these techniques generalize?

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 1/11</td>
<td>Course Overview</td>
<td>Nina</td>
</tr>
<tr>
<td>W 1/13</td>
<td>Intro to Probability</td>
<td>William</td>
</tr>
<tr>
<td>M 1/18</td>
<td>Martin Luther King Day</td>
<td></td>
</tr>
<tr>
<td>W 1/20</td>
<td>The Naive Bayes algorithm</td>
<td>William</td>
</tr>
<tr>
<td>M 1/25</td>
<td>Logistic Regression</td>
<td>William</td>
</tr>
<tr>
<td>W 1/27</td>
<td>Perceptrons and SVMs</td>
<td>William</td>
</tr>
<tr>
<td>M 2/1</td>
<td>Kernels</td>
<td>Nina</td>
</tr>
<tr>
<td>W 2/3</td>
<td>Linear Regression</td>
<td>Nina</td>
</tr>
<tr>
<td>M 2/8</td>
<td>Neural Networks and Backprop</td>
<td>Nina</td>
</tr>
<tr>
<td>W 2/10</td>
<td>Decision Trees and Rules</td>
<td>Nina</td>
</tr>
<tr>
<td>M 2/15</td>
<td>Boosting and Other Ensembles</td>
<td>Nina</td>
</tr>
<tr>
<td>W 2/17</td>
<td>Theory 1</td>
<td>Nina</td>
</tr>
<tr>
<td>M 2/22</td>
<td>Theory 2</td>
<td>Nina</td>
</tr>
<tr>
<td>W 2/24</td>
<td>Midterm Review</td>
<td>Nina</td>
</tr>
<tr>
<td>M 2/29</td>
<td>Midterm exam</td>
<td></td>
</tr>
</tbody>
</table>
What the course is about…

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>M 3/7</td>
<td>Semi-Supervised Learning</td>
<td>Nina</td>
</tr>
<tr>
<td>W 3/9</td>
<td>Active Learning</td>
<td>Nina</td>
</tr>
<tr>
<td>M 3/14</td>
<td>Spring break</td>
<td></td>
</tr>
<tr>
<td>W 3/16</td>
<td>Spring break</td>
<td></td>
</tr>
<tr>
<td>M 3/21</td>
<td>Graphical Models 1</td>
<td>William</td>
</tr>
<tr>
<td>W 3/23</td>
<td>Graphical Models 2</td>
<td>William</td>
</tr>
<tr>
<td>M 3/28</td>
<td>Graphical Models for Sequential Data</td>
<td>William</td>
</tr>
<tr>
<td>M 4/4</td>
<td>Topic Models</td>
<td>William</td>
</tr>
<tr>
<td>W 4/6</td>
<td>PCA and dimension reduction</td>
<td>William</td>
</tr>
<tr>
<td>M 4/11</td>
<td>Matrix Factorization and collaborative filtering</td>
<td>William</td>
</tr>
<tr>
<td>W 4/13</td>
<td>Deep Learning 1</td>
<td>William</td>
</tr>
<tr>
<td>M 4/18</td>
<td>Deep Learning 2</td>
<td>William</td>
</tr>
<tr>
<td>W 4/20</td>
<td>Reinforcement Learning</td>
<td>Nina</td>
</tr>
</tbody>
</table>

Other models for learning: learning + unlabeled data; learning with q/a

Richer tools for describing probability distributions

More learning models and advanced topics
Why I’m starting with

A REVIEW OF PROBABILITY
WHY PROBABILITY IS IMPORTANT
What is machine learning good for?

• Tasks involving **uncertainty** and/or complexity.

• If there is uncertainty how should you
• ... represent it, formally?
• ... reason about it, precisely?
• ... generalize from data, justifiably?
A paradox of induction

• A black crow seems to support the hypothesis “all crows are black”.

• A pink highlighter supports the hypothesis “all non-black things are non-crows”.

• Thus, a pink highlighter supports the hypothesis “all crows are black”.

\[ \forall x \text{ CROW}(x) \Rightarrow \text{BLACK}(x) \]

or equivalently

\[ \forall x \neg \text{BLACK}(x) \Rightarrow \neg \text{CROW}(x) \]
Hume’s Problem of Induction

- David Hume (1711-1776): pointed out
  1. Empirically, induction seems to work
  2. Statement (1) is an application of induction.

- This stumped people for about 200 years (Karl Popper, 1902-1994)

1. Of the Different Species of Philosophy.
2. Of the Origin of Ideas
3. Of the Association of Ideas
4. Sceptical Doubts Concerning the Operations of the Understanding
5. Sceptical Solution of These Doubts
6. Of Probability
7. Of the Idea of Necessary Connexion
8. Of Liberty and Necessity
9. Of the Reason of Animals
10. Of Miracles
11. Of A Particular Providence and of A Future State
12. ....
If there is uncertainty how should you

... represent it, formally?

... reason about it, precisely?

Success stories:

- geometry (Euclid, 300 BCE): five axioms, proofs, and theorems
- logic and Boolean algebra (1854)
- probability and games of chance (1800’s, Laplace)
- axiomatic treatment of probability (1900’s, Kolmorogov)
Models

True about the world

- has 1 propeller (aft)
- ship number = 571
- 320’ long
- built in 1955

True about the model

- has 1 propeller (aft)
- ship number = 571
- 14” long
- built in 2013
Formal Model of Geometry

Axioms:
- two points $\rightarrow$ one line
- one line segment $\rightarrow$ one line
- one line segment, endpoint $\rightarrow$ one circle
- all right angles are congruent
- parallel lines don’t meet

True about the world

sum of degrees of angles in a triangle = 180

True about the model

sum of degrees of angles in a triangle = 180
Formal Model of Uncertainty

True about the world

True about the model

“?”

“?”
Experiments, Outcomes, Random Variables and Events

- A is a Boolean-valued random variable if
  - A denotes an event, a possible outcome of an “experiment”
  - there is uncertainty as to whether A occurs.
    the experiment is not deterministic

- Define P(A) as “the fraction of experiments in which A is true”
  - We’re assuming all possible outcomes are equiprobable

- Examples
  - You roll two 6-sided die (the experiment) and get doubles (A=doubles, the outcome)
  - I pick two students in the class (the experiment) and they have the same birthday (A=same birthday, the outcome)
Formal Model of Uncertainty

“Experiments”

“Outcomes”

Axioms of probability

- \( \Pr(A) \geq 0 \)
- \( \Pr(\text{True}) = 1 \)
- If \( A_1, A_2, \ldots \) are disjoint then

\[
\Pr(A_1 \text{ or } A_2 \text{ or } \cdots) = \Pr(A_1) + \Pr(A_2) + \cdots
\]
Some more examples of uncertainty

- A is a Boolean-valued random variable if
  - A denotes an event,
  - there is uncertainty as to whether A occurs.

- More examples
  - A = You wake up tomorrow with a headache
  - A = The US president in 2023 will be male
  - A = there is intelligent life elsewhere in our galaxy
  - A = the 1,000,000,000,000,000,000,000,000th digit of \( \pi \) is 7
  - A = I woke up today with a headache

- Define \( P(A) \) as “the fraction of possible worlds in which A is true”
  - … seems a little awkward …
  - what if we just define \( P(A) \) as an arbitrary measure of belief?
Formal Model of Uncertainty

“Axioms of probability

- \( \Pr(A) \geq 0 \)
- \( \Pr(\text{True}) = 1 \)
- If \( A_1, A_2, \ldots \) are disjoint then

\[
\Pr(A_1 \text{ or } A_2 \text{ or } \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots
\]

“Beliefs”

“Observations”

True about the world

True about the model
These Axioms are Not to be Trifled With
- Andrew Moore

• There have been many many other approaches to understanding “uncertainty”:
  • Fuzzy Logic, three-valued logic, Dempster-Shafer, non-monotonic reasoning, …

• 40 years ago people in AI argued about these; now they mostly don’t
  – Any scheme for combining uncertain information, uncertain “beliefs”, etc,… really should obey these axioms to be internally consistent (from Jayne, 1958; Cox 1930’s)
  – If you gamble based on “uncertain beliefs”, then [you can be exploited by an opponent] ⇐ [your uncertainty formalism violates the axioms] - di Finetti 1931 (the “Dutch book argument”)

[Image of Andrew Moore]
I guess this is Andrey Kolmorogov
EXAMPLES OF REASONING WITH AXIOMS OF PROBABILITY
Andrew’s Axioms of Probability

A1,A2,A3,A4
1. 0 <= P(A) <= 1
2. P(True) = 1
3. P(False) = 0
4. P(A or B) = P(A) + P(B) - P(A and B)

Are these the same as before?
Andrew’s Axioms of Probability

A1,…,A4
1. \(0 \leq P(A) \leq 1\) ✔
2. \(P(\text{True}) = 1\) ✔
3. \(P(\text{False}) = 0\)
4. \(P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)\)

Monotonicity: if A is a subset of B, then \(P(A) \leq P(B)\)

Proof:
- A subset of B \(\Rightarrow B = A + C\) for \(C = B - A\)
- A and C are disjoint \(\Rightarrow P(B) = P(A \text{ or } C) = P(A) + P(C)\)
- \(P(C) \geq 0\)
- So \(P(B) \geq P(A)\)

Monotonicity and K2 \(\Rightarrow A1\)

K1,…,K3
1. \(\Pr(A) \geq 0\)
2. \(\Pr(\text{True}) = 1\)
3. If A1, A2, …. are disjoint then

\[ \Pr(A1 \text{ or } A2 \text{ or } \ldots) = \Pr(A1) + \Pr(A2) + \ldots \]
Andrew’s Axioms of Probability

1. \(0 \leq P(A) \leq 1\)
2. \(P(\text{True}) = 1\)
3. \(P(\text{False}) = 0\)
4. \(P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)\)

Theorem: \(P(\neg A) = 1 - P(A)\)

Proof:
- \(P(A \text{ or } \neg A) = P(\text{True}) = 1\) \(\text{K2}\)
- \(A \text{ and } \neg A \text{ are disjoint} \Rightarrow P(A) + P(\neg A) = P(A \text{ or } \neg A)\) \(\text{K3}\)
- \(P(A) + P(\neg A) = 1\)

....then solve for \(P(\neg A)\)
Andrew’s Axioms of Probability

1. $0 \leq P(A) \leq 1$ ✔
2. $P(\text{True}) = 1$ ✔
3. $P(\text{False}) = 0$ ✔
4. $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ ✔

Theorem: $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Proof:
- $E_1 = A \text{ and } \sim(A \text{ and } B)$
- $E_2 = (A \text{ and } B)$
- $E_3 = B \text{ and } \sim(A \text{ and } B)$
- $E_1 \text{ or } E_2 \text{ or } E_3 = A \text{ or } B$ and $E_1, E_2, E_3$ disjoint ➔
  - $P(A \text{ or } B) = P(E_1) + P(E_2) + P(E_3)$
- further $P(A) = P(E_1) + P(E_2)$ and $P(B) = P(E_3) + P(E_2)$
- ...

1. $\Pr(A) \geq 0$
2. $\Pr(\text{True}) = 1$
3. If $A_1, A_2, \ldots$ are disjoint then
   \[ \Pr(A_1 \text{ or } A_2 \text{ or } \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots \]
Conclusion…

• T/F: Do you want to be famous for developing a new theory of uncertainty?

• Answer: no. It would either violate one of the axioms, or just be a syntactic variant of probability
KEY CONCEPTS IN PROBABILITY
Probability - what you need to really, really know

• Probabilities are cool
• Random variables and events
• The axioms of probability
  – These define a formal model for uncertainty
• Independence
Independent Events

- Definition: two events A and B are independent if $\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B)$.
- Intuition: outcome of A has no effect on the outcome of B (and vice versa).
  - We need to assume the different rolls are independent to solve the problem.
  - You frequently need to assume the independence of something to solve any learning problem.
Some practical problems

Pearl Purple Cheaters Dice

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- You’re the DM in a D&D game.
- Joe brings his own d20 and throws 4 critical hits in a row to start off
  - DM=dungeon master
  - D20 = 20-sided die
  - “Critical hit” = 19 or 20
- What are the odds of that happening with a fair die?
- Ci=critical hit on trial i, i=1,2,3,4
- \( P(C_1 \text{ and } C_2 \ldots \text{ and } C_4) = P(C_1) \times \ldots \times P(C_4) = (1/10)^4 \)
- To get there we assumed the rolls were independent
Multivalued Discrete Random Variables

• Suppose \( A \) can take on more than 2 values

• \( A \) is a random variable with arity \( k \) if it can take on exactly one value out of \( \{v_1, v_2, \ldots, v_k\} \)
  – Example: \( V = \{1, 2, 3, \ldots, 20\} \): good for 20-sided dice games

• Notation: let's write the event \( A \text{ Has Value Of } v \) as \( "A = v" \)

• To get the right behavior: define as axioms:

\[
P(A = v_i \land A = v_j) = 0 \text{ if } i \neq j
\]

\[
P(A = v_1 \lor A = v_2 \lor \ldots \lor A = v_k) = 1
\]
....or in pictures

\[ \sum_{j=1}^{k} P(A = v_j) = 1 \]
More about Multivalued Random Variables

Another example: modeling English text

\[ V = \{ \text{aaliyah, aardvark, \ldots, zymurge, zynga} \} \]

very useful for modeling text (but hard to keep track of unless you have a computer):

\[ A_1 = \text{another} \quad A_2 = \text{example} \quad A_3 = v \quad A_4 = \text{“} = \text{“} \quad \ldots \quad A_{31} = \text{computer} \quad A_{32} = \text{“} ) \text{“} \quad A_{33} = \text{“} : \text{“} \]
Continuous Random Variables

• The discrete case: sum over all values of A is 1

\[ \sum_{j=1}^{k} P(A = v_j) = 1 \]

• The continuous case: infinitely many values for A and the integral is 1

\[ \int_{-\infty}^{\infty} f_P(x) \, dx = 1 \]

also….

\[ \forall x, f_p(x) \geq 0 \]

1. \( \Pr(A) \geq 0 \)
2. \( \Pr(\text{True}) = 1 \)
3. If \( A_1, A_2, \ldots \) are disjoint then

\[ \Pr(A_1 \text{ or } A_2 \text{ or } \ldots) = \Pr(A_1) + \Pr(A_2) + \ldots \]

\( f(x) \) is a probability density function (pdf)
Continuous Random Variables

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}. \]
KEY CONCEPTS IN PROBABILITY:

CONDITIONAL PROBABILITY AND THE CHAIN RULE
Probability - what you need to really, really know

• Probabilities are cool
• Random variables and events
• The axioms of probability
• Independence, binomials, multinomials, continuous distributions, pdf’s
• Conditional probabilities
A practical problem

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have two d20 die, one loaded, one standard
- Loaded die will give a 19/20 (“critical hit”) half the time.
- In the game, someone hands me a random die, which is fair (A) with P(A)=0.5. Then I roll, and either get a critical hit (B) or not (~B).
- What is P(B)?

\[
P(B) = P(B \text{ and } A) + P(B \text{ and } \sim A) = 0.1 \times 0.5 + 0.5 \times (0.5) = 0.3
\]
A practical problem

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

• I have lots of standard d20 die, lots of loaded die, all identical.
• Loaded die will give a 19/20 (“critical hit”) half the time.
• In the game, someone hands me a random die, which is fair (A) or loaded (~A), with P(A) depending on how I mix the die. Then I roll, and either get a critical hit (B) or not (~B)
• Can I mix the dice together so that P(B) is anything I want - say, p(B) = 0.137?

\[ P(B) = P(B \text{ and } A) + P(B \text{ and } \neg A) = 0.1\lambda + 0.5(1-\lambda) = 0.137 \]

“mixture model”

\[ \lambda = (0.5 - 0.137)/0.4 = 0.9075 \]

Mixtures let us build more complex models from simpler ones.
It’s more convenient to say
• “if you’ve picked a fair die then …” i.e. Pr(critical hit | fair die)=0.1
• “if you’ve picked the loaded die then…..” Pr(critical hit | loaded die)=0.5

Conditional probability:
Pr(B | A) = P(B^A)/P(A)
Definition of Conditional Probability

\[ P(A \cap B) = P(A \mid B) \times P(B) \]

Corollary: The Chain Rule

\[ P(A \cap B) = P(A \mid B) \times P(B) \]
Some practical problems

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have 3 standard d20 dice, 1 loaded die.

- Experiment: (1) pick a d20 uniformly at random then (2) roll it. Let \( A = \) d20 picked is fair and \( B = \) roll 19 or 20 with that die. What is \( P(B) \)?

\[
P(B) = P(B \mid A) \ P(A) + P(B \mid \sim A) \ P(\sim A) = 0.1 \times 0.75 + 0.5 \times 0.25 = 0.2
\]
\[ P(B) = P(B | A)P(A) + P(B | \sim A)P(\sim A) \]
NOTATION
Very widely used shortcut

**Shortcut**
\[ \Pr(A \land B) = \Pr(A \mid B) \Pr(B) \]

**Long Form**
\[ \Pr(A = 1 \land B = 1) = \Pr(A = 1 \mid B = 1) \Pr(B = 1) \]
\[ \Pr(A = 0 \land B = 0) = \Pr(A = 0 \mid B = 0) \Pr(B = 0) \]
\[ \Pr(A = 1 \land B = 0) = \Pr(A = 1 \mid B = 0) \Pr(B = 0) \]

\[ A \in \{a_1, \ldots, a_{K_a}\}, B \in \{b_1, \ldots, b_{K_b}\} \]
\[ \Pr(A = a_i \land B = b_j) = \Pr(A = a_i \mid B = b_j) \Pr(B = b_j) \]
for every \(i\) and \(j\): \(1 \leq i \leq K_a, 1 \leq j \leq K_b\)

Put another way: the chain rule holds for any events \(A\) and \(B\). For multivalued discrete variables, there are many possible "\(A\) events" (events I could denote by \(A\)) and many possible "\(B\) events".

Consequence: estimating \(\Pr(A \mid B)\) might mean estimating many numbers....
Definition of Conditional Probability

\[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} \]

Corollary: The Chain Rule

\[ P(A \cap B) = P(A \mid B) \cdot P(B) \]
KEY CONCEPTS IN PROBABILITY:

SMOOTHING
Some practical problems

I bought a loaded d20 on EBay…but it didn’t come with any useful specs. How can I find out how it behaves?

<table>
<thead>
<tr>
<th>Face Shown</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

1. Collect some data (20 rolls)
2. Estimate $\Pr(i) = \frac{C(\text{rolls of } i)}{C(\text{any roll})}$
One solution

I bought a loaded d20 on EBay…but it didn’t come with any specs. How can I find out how it behaves?

Face Shown

P(1)=0
P(2)=0
P(3)=0
P(4)=0.1
...
P(19)=0.25
P(20)=0.2

MLE = maximum likelihood estimate

But: Do I really think it’s \textit{impossible} to roll a 1,2 or 3?
A better solution

I bought a loaded d20 on EBay… but it didn’t come with any specs. How can I find out how it behaves?

0. Imagine some data (20 rolls, each face shows up 1x)
1. Collect some data (20 rolls)
2. Estimate $\Pr(i) = \frac{\text{C(rolls of } i)}{\text{C(any roll)}}$
A better solution

I bought a loaded d20 on eBay…but it didn’t come with any specs. How can I find out how it behaves?

\[ \hat{\text{Pr}}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)} \]

\[
\begin{align*}
\text{P}(1) &= \frac{1}{40} \\
\text{P}(2) &= \frac{1}{40} \\
\text{P}(3) &= \frac{1}{40} \\
\text{P}(4) &= \frac{2+1}{40} \\
\cdots \\
\text{P}(19) &= \frac{5+1}{40} \\
\text{P}(20) &= \frac{4+1}{40} = \frac{1}{8}
\end{align*}
\]

0.25 vs. 0.125 – really different! Maybe I should “imagine” less data?
A better solution?

Pearl Purple Cheaters Dice

![Dice Image]

\[
\hat{Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}
\]

\[
P(1) = \frac{1}{40}
\]

\[
P(2) = \frac{1}{40}
\]

\[
P(3) = \frac{1}{40}
\]

\[
P(4) = \frac{2+1}{40}
\]

…

\[
P(19) = \frac{5+1}{40}
\]

\[
P(20) = \frac{4+1}{40} = \frac{1}{8}
\]

0.25 vs. 0.125 – really different! Maybe I should “imagine” less data?
A better solution?

Q: What if I used $m$ rolls with a probability of $q=1/20$ of rolling any $i$?

$$
\hat{Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}
$$

$$
\hat{Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}
$$

I can use this formula with $m>20$, or even with $m<20$ … say with $m=1$
A better solution

Q: What if I used $m$ rolls with a probability of $q=1/20$ of rolling any $i$?

$$\hat{\Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}$$

$$\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}$$

If $m\gg C(ANY)$ then your imagination $q$ rules.
If $m<<C(ANY)$ then your data rules BUT you never ever ever end up with $\Pr(i)=0$. 
Terminology – more later

This is called a **uniform Dirichlet prior**

C(i), C(ANY) are **sufficient statistics**

\[
\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}
\]

**MLE** = maximum likelihood estimate

**MAP** = maximum a posteriori estimate
RECAP:
PROBABILITIES ARE AWESOME
B = black
C = crow

collect statistics for P(B | C)
KEY CONCEPTS IN PROBABILITY:

BAYES RULE
Some practical problems

- I have 3 standard d20 dice, 1 loaded die.
- Experiment: (1) pick a d20 uniformly at random then (2) roll it. Let $A=$d20 picked is fair and $B=$roll 19 or 20 with that die.
- Suppose B happens (e.g., I roll a 20). What is the chance the die I rolled is fair? i.e. what is $P(A \mid B)$?
\[ P(A \mid B) = ? \]

\[
P(A \text{ and } B) = P(A \mid B) \times P(B) \\
P(A \text{ and } B) = P(B \mid A) \times P(A) \\
P(A \mid B) \times P(B) = P(B \mid A) \times P(A)
\]

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of analogical or inductive reasoning….
Probability - what you need to really, really know

• Probabilities are cool
• Random variables and events
• The Axioms of Probability
• Independence, binomials, multinomials, …
• Conditional probabilities
• Bayes Rule
Some practical problems

- Joe throws 4 critical hits in a row, is Joe cheating?
- \( A = \text{Joe using cheater’s die} \)
- \( C = \text{roll 19 or 20}; \ P(C | A) = 0.5, \ P(C | \sim A) = 0.1 \)
- \( B = C_1 \text{ and } C_2 \text{ and } C_3 \text{ and } C_4 \)
- \( \Pr(B | A) = 0.0625 \quad \Pr(B | \sim A) = 0.0001 \)

\[
P(A | B) = \frac{P(B | A)P(A)}{P(B | A)P(A) + P(B | \sim A)P(\sim A)}
\]

\[
P(A | B) = \frac{0.0625 \times P(A)}{0.0625 \times P(A) + 0.0001 \times (1 - P(A))}
\]
What’s the experiment and outcome here?

• Outcome A: Joe is cheating
• Experiment:
  – Joe picked a die uniformly at random from a bag containing 10,000 fair die and one bad one.
  – Joe is a D&D player picked uniformly at random from set of 1,000,000 people and $n$ of them cheat with probability $p>0$.
  – I have no idea, but I don’t like his looks. Call it $P(A)=0.1$
Remember: Don’t Mess with The Axioms

• A subjective belief can be treated, mathematically, like a probability
  – Use those axioms!

• There have been many many other approaches to understanding “uncertainty”:
  • Fuzzy Logic, three-valued logic, Dempster-Shafer, non-monotonic reasoning, …

• 25 years ago people in AI argued about these; now they mostly don’t
  – Any scheme for combining uncertain information, uncertain “beliefs”, etc,... really should obey these axioms
  – If you gamble based on “uncertain beliefs”, then [you can be exploited by an opponent] ⇔ [your uncertainty formalism violates the axioms] - di Finetti 1931 (the “Dutch book argument”)
Some practical problems

\[
P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}
\]

\[
\frac{P(A \mid B)}{P(\neg A \mid B)} = \frac{P(B \mid A)P(A) / P(B)}{P(B \mid \neg A)P(\neg A) / P(B)} = \frac{P(B \mid A)}{P(B \mid \neg A)} \times \frac{P(A)}{P(\neg A)}
\]

\[= \frac{0.0625}{0.0001} \times \frac{P(A)}{P(\neg A)} = 6,250 \times \frac{P(A)}{P(\neg A)}
\]

- Joe throws 4 critical hits in a row, is Joe cheating?
- A = Joe using cheater’s die
- C = roll 19 or 20; P(C \mid A)=0.5, P(C \mid \neg A)=0.1
- B = C1 and C2 and C3 and C4
- Pr(B \mid A) = 0.0625 \quad P(B \mid \neg A)=0.0001

Moral: with enough evidence the prior P(A) doesn’t really matter.
Stopped here Wed 11/13
KEY CONCEPTS IN PROBABILITY:
SMOOTHING, MLE, AND MAP
Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE’s, smoothing, and MAPs
Some practical problems

I bought a loaded d20 on EBay...but it didn't come with any specs. How can I find out how it behaves?

1. Collect some data (20 rolls)
2. Estimate \( \Pr(i) = \frac{C(\text{rolls of } i)}{C(\text{any roll})} \)
I bought a loaded d20 on EBay…but it didn’t come with any specs. How can I find out how it behaves?

**MLE = maximum likelihood estimate**

But: Do I really think it’s *impossible* to roll a 1,2 or 3?
A better solution

I bought a loaded d20 on EBay…but it didn’t come with any specs. How can I find out how it behaves?

0. Imagine some data (20 rolls, each i shows up 1x)
1. Collect some data (20 rolls)
2. Estimate $\Pr(i) = \frac{C(\text{rolls of } i)}{C(\text{any roll})}$
A better solution

I bought a loaded d20 on EBay…but it didn’t come with any specs. How can I find out how it behaves?

\[
\hat{P}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}
\]

\[
P(1) = \frac{1}{40}
\]
\[
P(2) = \frac{1}{40}
\]
\[
P(3) = \frac{1}{40}
\]
\[
P(4) = \frac{2+1}{40}
\]
\[
\vdots
\]
\[
P(19) = \frac{5+1}{40}
\]
\[
P(20) = \frac{4+1}{40} = \frac{1}{8}
\]

0.25 vs. 0.125 – really different! Maybe I should “imagine” less data?
A better solution?

Pearl Purple Cheaters Dice

\[
\hat{Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}
\]

\[
P(1) = \frac{1}{40}
\]
\[
P(2) = \frac{1}{40}
\]
\[
P(3) = \frac{1}{40}
\]
\[
P(4) = \frac{2+1}{40}
\]
\[
\ldots
\]
\[
P(19) = \frac{5+1}{40}
\]
\[
P(20) = \frac{4+1}{40} = \frac{1}{8}
\]

0.25 vs. 0.125 – really different! Maybe I should “imagine” less data?
A better solution?

Q: What if I used $m$ rolls with a probability of $q = 1/20$ of rolling any $i$?

$$\hat{\Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}$$

$$\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}$$

I can use this formula with $m > 20$, or even with $m < 20$ … say with $m = 1$
Q: What if I used \( m \) rolls with a probability of \( q=1/20 \) of rolling any \( i \)?

\[
\hat{\Pr}(i) = \frac{C(i) + 1}{C(ANY) + C(IMAGINED)}
\]

\[
\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}
\]

If \( m >> C(ANY) \) then your imagination \( q \) rules.

If \( m << C(ANY) \) then your data rules BUT you never ever end up with \( \Pr(i)=0 \)
Terminology – more later

This is called a **uniform Dirichlet** prior

C(i), C(ANY) are **sufficient statistics**

\[
\hat{\Pr}(i) = \frac{C(i) + mq}{C(ANY) + m}
\]

**MLE** = maximum likelihood estimate

**MAP** = maximum a posteriori estimate
Why we call this a MAP

• Simple case: replace the die with a coin
  – Now there’s one parameter: \( q = P(H) \)
  – I start with a prior over \( q \), \( P(q) \)
  – I get some data: \( D = \{D1 = H, D2 = T, \ldots\} \)
  – I compute maximum of posterior of \( q \)

\[
P(q \mid D) = \frac{P(D \mid q)P(q)}{P(D)} = \frac{1}{\int_0^1 P(D \mid r)P(r)dr} P(D \mid q)P(q)
\]
Why we call this a MAP

• Simple case: replace the die with a coin
  – Now there’s one parameter: $q=P(H)$
  – I start with a prior over $q$, $P(q)$
  – I get some data: $D=\{D1=H, D2=T, \ldots\}$
  – I compute the posterior of $q$

• The math works if the pdf $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$

• $\alpha, \beta$ are imaginary pos/neg examples
Why we call this a MAP

- The math works if the pdf $f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$

- $\alpha, \beta$ are imaginary pos/neg examples
Why we call this a MAP

- This is called a *beta* distribution
- The generalization to multinomials is called a *Dirichlet* distribution
- Parameters are $\alpha_1, \ldots, \alpha_K$

$$f(x_1, \ldots, x_K) = \frac{1}{\text{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i-1}$$
SOME MORE TERMS
Probability Density Function

- Discrete distributions

\[ \sum_{i} P(X = x_i) = 1 \]

- Continuous: Probability density function (pdf) vs Cumulative Density Function (CDF):

\[ P(x \leq a) = \int_{-\infty}^{a} f(\tau) d\tau \]
Cumulative Density Functions

• Total probability

\[ P(\Omega) = \int_{-\infty}^{\infty} f(x) \, dx = 1 \]

• Probability Density Function (PDF)

\[ \frac{d}{dx} F(x) = f(x) \]

• Properties:

\[ P(a \leq x \leq b) = \int_{b}^{a} f(x) \, dx = F(b) - F(a) \]

\[ \lim_{x \to -\infty} F(x) = 0 \]

\[ \lim_{x \to \infty} F(x) = 1 \]

\[ F(a) \geq F(b) \ \forall a \geq b \]
Expectations

• Mean/Expected Value:

\[ E[x] = \bar{x} = \int x f(x) \, dx \]

• Variance:

\[ Var(x) = E[(x - \bar{x})^2] = E[x^2] - (\bar{x})^2 \]

• More examples:

\[ E[x^2] = \int x^2 f(x) \, dx \]

\[ E[g(x)] = \int g(x) f(x) \, dx \]
Multivariate versions....

• Joint for \((x,y)\)

\[
P ((x, y) \in A) = \int \int_A f(x, y) dx dy
\]

• Marginal:

\[
f(x) = \int f(x, y) dy
\]

• Conditionals:

\[
f(x|y) = \frac{f(x, y)}{f(y)}
\]

• Chain rule:

\[
f(x, y) = f(x|y)f(y) = f(y|x)f(x)
\]
Bayes Rule

• Standard form:

\[ f(x|y) = \frac{f(y|x)f(x)}{f(y)} \]

• Replacing the bottom:

\[ f(x|y) = \frac{f(y|x)f(x)}{\int f(y|x)f(x)\,dx} \]
Binomial

• Distribution:

\[ x \sim Binomial(p, n) \]

\[ P(x = k) = \binom{n}{k} p^k (1 - p)^{n-k} \]

• Mean/Var:

\[ E[x] = np \]

\[ Var(x) = np(1 - p) \]
Uniform

- Anything is equally likely in the region \([a, b]\)

- Distribution:

\[ x \sim U(a, b) \]

- Mean/Var

\[
f(x) = \begin{cases} 
\frac{1}{b-a} & \text{if } a \leq x \leq b \\
0 & \text{otherwise}
\end{cases}
\]

\[
E[x] = \frac{a + b}{2}
\]

\[
Var(x) = \frac{a^2 + ab + b^2}{3}
\]
Gaussian (Normal)

- If I look at the height of women in country xx, it will look approximately Gaussian.
- Small random noise errors, look Gaussian/Normal.

Distribution:

\[ x \sim N(\mu, \sigma^2) \]

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

- Mean/Var:

\[ E[x] = \mu \]

\[ Var(x) = \sigma^2 \]
Why Do People Use Gaussians

- Central Limit Theorem: (loosely)
  - Sum of a large number of IID random variables is approximately Gaussian
Sum of Gaussians

- The sum of two Gaussians is a Gaussian:

\[ x \sim N(\mu, \sigma^2) \quad y \sim N(\mu_y, \sigma_y^2) \]

\[ ax + b \sim N(a\mu + b, (a\sigma)^2) \]

\[ x + y \sim N(\mu + \mu_y, \sigma^2 + \sigma_y^2) \]
Multivariate Gaussians

- Distribution for vector $x$

$$x = (x_1, \ldots, x_N)^T, \quad x \sim N(\mu, \Sigma)$$

- PDF:

$$f(x) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$$E[x] = \mu = (E[x_1], \ldots, E[x_N])^T$$

$$\text{Var}(x) \rightarrow \Sigma = \begin{pmatrix}
\text{Var}(x_1) & \text{Cov}(x_1, x_2) & \cdots & \text{Cov}(x_1, x_N) \\
\text{Cov}(x_2, x_1) & \text{Var}(x_2) & \cdots & \text{Cov}(x_2, x_N) \\
\vdots & \vdots & \ddots & \vdots \\
\text{Cov}(x_N, x_1) & \text{Cov}(x_N, x_2) & \cdots & \text{Var}(x_N)
\end{pmatrix}$$
Multivariate Gaussians

\[ f(x) = \frac{1}{(2\pi)^{N/2} |\Sigma|^{1/2}} e^{-\frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu)} \]

\[ E[x] = \mu = (E[x_1], \ldots, E[x_N])^T \]

\[ \text{Var}(x) \rightarrow \Sigma = \begin{pmatrix} \text{Var}(x_1) & \text{Cov}(x_1, x_2) & \ldots & \text{Cov}(x_1, x_N) \\ \text{Cov}(x_2, x_1) & \text{Var}(x_2) & \ldots & \text{Cov}(x_2, x_N) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(x_N, x_1) & \text{Cov}(x_N, x_2) & \ldots & \text{Var}(x_N) \end{pmatrix} \]

\[ \text{Cov}(x_1, x_2) = \frac{1}{n} \sum_{i=1}^{n} (x_{1,i} - \mu_1)(x_{2,i} - \mu_2) \]
Covariance examples

Anti-correlated
Covariance: -9.2

Independent (almost)
Covariance: 0.6

Correlated
Covariance: 18.33
How much do grad students sleep?

- Lets try to estimate the distribution of the time students spend sleeping (outside class).
Possible statistics

• \( X \)
  Sleep time

• Mean of \( X \):
  \( E\{X\} \)
  7.03

• Variance of \( X \):
  \( Var\{X\} = E\{(X-E\{X\})^2\} \)
  3.05
Covariance: Sleep vs. GPA

• Co-Variance of $X_1$, $X_2$:

\[ \text{Covariance}\{X_1,X_2\} = \mathbb{E}\{(X_1-\mathbb{E}\{X_1\})(X_2-\mathbb{E}\{X_2\})\} \]

\[ = 0.88 \]
KEY CONCEPTS IN PROBABILITY:

THE JOINT DISTRIBUTION
Probability - what you need to really, really know

- Probabilities are cool
- Random variables and events
- The Axioms of Probability
- Independence, binomials, multinomials
- Conditional probabilities
- Bayes Rule
- MLE’s, smoothing, and MAPs
- The joint distribution
Some practical problems

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

• I have 1 standard d6 die, 2 loaded d6 die.
• Loaded high: \( P(X=6)=0.50 \)  
  Loaded low: \( P(X=1)=0.50 \)
• Experiment: pick one d6 uniformly at random (A) and roll it. What is more likely – rolling a seven or rolling doubles?

Three combinations: HL, HF, FL

\[
P(D) = P(D \cap A=HL) + P(D \cap A=HF) + P(D \cap A=FL) \\
= P(D \mid A=HL)P(A=HL) + P(D\mid A=HF)P(A=HF) + P(A\mid A=FL)P(A=FL)
\]
Some practical problems

This is a set of polyhedral dice that will roll high/low numbers. The set consists of 3 six-sided dice (1 high, 1 low, 1 standard), 2 ten-sided dice (1 high, 1 standard), and 2 twenty-sided dice (1 high, 1 standard).

- I have 1 standard d6 die, 2 loaded d6 die.
- Loaded high: \( P(X=6)=0.50 \)   Loaded low: \( P(X=1)=0.50 \)
- Experiment: pick one d6 uniformly at random (A) and roll it. Repeat a second time. What is more likely – rolling a seven or rolling doubles?

Three combinations: HL, HF, FL

<table>
<thead>
<tr>
<th>Roll 1</th>
<th>Roll 2</th>
</tr>
</thead>
</table>
| \( \begin{array}{cccccc}
1 & D & & & & 7 \\
2 & D & & & & 7 \\
3 & D & 7 & & & \\
4 & & 7 & D & & \\
5 & & & & D & \\
6 & & & & D & \\
\end{array} \) |
A brute-force solution

<table>
<thead>
<tr>
<th>A</th>
<th>Roll 1</th>
<th>Roll 2</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>FL</td>
<td>1</td>
<td>1</td>
<td>$1/3 \times 1/6 \times 1/2$</td>
</tr>
<tr>
<td>FL</td>
<td>1</td>
<td>2</td>
<td>$1/3 \times 1/6 \times 1/10$</td>
</tr>
<tr>
<td>FL</td>
<td>1</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>FL</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>FL</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>FL</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>HF</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td></td>
</tr>
</tbody>
</table>

A joint probability table shows $P(X_1=x_1 \text{ and } \ldots \text{ and } X_k=x_k)$ for every possible combination of values $x_1,x_2,\ldots,x_k$.

With this you can compute any $P(A)$ where $A$ is any boolean combination of the primitive events $(X_i=X_k)$, e.g.

- $P(\text{doubles})$
- $P(\text{seven or eleven})$
- $P(\text{total is higher than 5})$
- $\ldots$
Recipe for making a joint distribution of $M$ variables:

Example: Boolean variables $A, B, C$
The Joint Distribution

Recipe for making a joint distribution of $M$ variables:

1. Make a truth table listing all combinations of values of your variables (if there are $M$ Boolean variables then the table will have $2^M$ rows).

Example: Boolean variables $A, B, C$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The Joint Distribution

Recipe for making a joint distribution of $M$ variables:

1. Make a truth table listing all combinations of values of your variables (if there are $M$ Boolean variables then the table will have $2^M$ rows).

2. For each combination of values, say how probable it is.

Example: Boolean variables $A$, $B$, $C$

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>
The Joint Distribution

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have $2^M$ rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Recipe for making a joint distribution of $M$ variables:

1. Make a truth table listing all combinations of values of your variables (if there are $M$ Boolean variables then the table will have $2^M$ rows).
2. For each combination of values, estimate how probable it is from data.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>Prob</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.30</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0.10</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.05</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0.10</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0.25</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Density Estimation

• Our Joint Distribution learner is our first example of something called Density Estimation

• A Density Estimator learns a mapping from a set of attributes values to a Probability
Density Estimation – looking ahead

- Compare it against the two other major kinds of models:

  - **Classifier**
    - Input Attributes
    - Prediction of categorical output or class
      - One of a few discrete values
  
  - **Density Estimator**
    - Input Attributes
    - Probability
  
  - **Regressor**
    - Input Attributes
    - Prediction of real-valued output