Anderson’s Rational Analysis Framework

1. Specify what the system is trying to optimize (accuracy, caloric input, detection, correct classification)

2. Make assumptions about the structure of the environment to which the system is thought to be adapted.

3. Specify nature of costs (tradeoffs in terms of effort, work, costs for being wrong, etc...)

4. Derive optimal solution using some extension of Bayesian decision theory

5. Look at empirical results and see if predictions are correct (basically the same as any model)

6. If wrong, recast 1-3 and recompute 4 with the new assumptions
Bayes Rule

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

\[ p(d) = \sum_{h \in H} p(d|h)p(h) \]
Bayes Rule

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

**Posterior**

**Likelihood** \( p(d|h)p(h) \)

**Prior** \( p(h) \)

**Normalizing Constant (aka evidence)** \( p(d) \)

\[ p(d) = \sum_{h \in H} p(d|h)p(h) \]
Bayes Rule

\[ p(h|d) = \frac{p(h \& d)}{p(d)} \]

\[ p(h|d)p(d) = p(h \& d) \]

\[ p(d|h)p(h) = p(h \& d) \]

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]
Coin Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

50%  90%
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

HHHHHHHHHHHHH
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[ h_0 \]

\[ h_1 \]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one a random. Flip it ten times. Which coin did you pick up?

50% 90%

HHHHHHHHHHHHHH

\[ h_0 \quad p(h_0) = p(h_1) = 0.5 \quad h_1 \]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

Prior

\[ p(h_0) = p(h_1) = 0.5 \]

\[ h_0 \]

\[ h_1 \]

Likelihood

\[ p(d|\theta) = \theta^N_H (1 - \theta)^N_T \]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
p(h_0 \mid H H H H H H H H H H) = \frac{p(d \mid 0.5)p(h_0)}{p(d)}
\]

\[
p(h_1 \mid H H H H H H H H H H) = \frac{p(d \mid 0.9)p(h_1)}{p(d)}
\]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
\frac{p(h_1 | HHHHHHHHHHH)}{p(h_0 | HHHHHHHHHHH)} = \frac{P(d|h_1) P(h_1)}{P(d|h_0) P(h_0)}
\]

Posterior Odds
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
p(\theta = 0.9 \mid H H H H H H H H H H) = \frac{P(d \mid \theta = 0.9) \cdot 0.5}{P(d \mid \theta = 0.5) \cdot 0.5}
\]

\[
h_0 \quad 50\% \quad H H H H H H H H H H \quad 90\% \quad h_1
\]

Monday, March 2, 2009
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
\begin{align*}
&h_0 \quad 50\% \\
&H H H H H H H H H H H H H H H H \\
&h_1 \quad 90\% \\
\end{align*}
\]

\[
\frac{p(\theta = 0.9 | H H H H H H H H H H H H H H H H)}{p(\theta = 0.5 | H H H H H H H H H H H H H H H H)} = \frac{0.9^{10} (1 - 0.9)^0}{0.5^{10} (1 - 0.5)^0} = 0.5
\]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one a random. Flip it ten times. Which coin did you pick up?

\[ p(\theta = 0.9 \mid H H H H H H H H H H) = \frac{p(\theta = 0.9 \mid H H H H H H H H H H)}{p(\theta = 0.5 \mid H H H H H H H H H H)} = 357.05 \]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
p(\theta = 0.9 \mid HHTHTHTHTT) = \frac{0.9^5 (1 - 0.9)^5}{0.5^5 (1 - 0.5)^5} \approx 0.5
\]

\[
p(\theta = 0.5 \mid HHTHTHTHTT) = \frac{0.5^5 (1 - 0.5)^5}{0.5^5 (1 - 0.5)^5} = 0.5
\]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one at random. Flip it ten times. Which coin did you pick up?

\[
p(\theta = 0.9 \mid HHTHTHTHTTHT) = \frac{0.9^5 (1 - 0.9)^5}{0.5^5 (1 - 0.5)^5} \approx 0.5
\]

\[
p(\theta = 0.5 \mid HHTHTHTHTTHT) = \frac{0.5^5 (1 - 0.5)^5}{0.5^5 (1 - 0.5)^5} = 0.5
\]
Example

From Griffiths and Yuille reading:

Box with two coins. One is fair, one is biased (90% heads). Pick one a random. Flip it ten times. Which coin did you pick up?

\[
p(\theta = 0.9 | HHTHTHTTTHT) = 5.90e - 06
\]

\[
p(\theta = 0.5 | HHTHTHTTTHT) = \frac{1}{2}
\]
Bayes Rule for continuous distributions

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

\[ p(d) = \sum_{h \in H} p(d|h)p(h) \]
Bayes Rule for continuous distributions

\[ p(h) \]

\[ p(d|h) \]

\[ p(h|d) \]

from Kruschke, forthcoming
Bayes Rule for continuous distributions

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

\[ p(d) = \int p(d|h)p(h)\,dh \]

Allows us to simultaneously consider many, many hypotheses at once.

Doesn’t have to be a single number. Can be an entire distribution of values.
Bayes Rule for continuous distributions

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

\[ p(d) = \int p(d|h)p(h) \, dh \]

Allows us to simultaneously consider many, many hypotheses at once.

This can be hard to calculate!
Modern approaches

1. **Mathematical integration and multiplication of distributions** often depends on *conjugate priors* for mathematical simplification (basically you choose your likelihood or model, then choose a distribution for expressing the prior that makes evaluating the integral in the last slide easier because the forms is known)

2. **Grid based methods** (allow arbitrary priors and likelihood models but get complex with large number of parameters)

3. **Stochastic approximation methods** (MCMC, Gibbs sampler, particle filters, importance sampling: discussed next time)
Subjective Probabilities

1. In the Bayesian approach, probabilities reflect subjective “degrees of belief” associated with particular outcomes.

2. This is different than the frequentist view of probabilities where probability means the relative frequency of an event as the number of trials goes to infinity:

   \[ p(x) = \lim_{n_t \to \infty} \frac{n_x}{n_t} \]
Bayes rule: where prior belief and empirical data meet

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

Basically an answer to the question, how should my beliefs change given some data?

Major criticism to keep in mind: where do the priors come from? What do they mean? Is there any reasonable procedure for determining the priors in any situation?
Three formal goals for Bayesian data analysis (Kruschke, forthcoming Bayesian data analysis book)

1. **Estimation** of parameter values (like last example), given some data, try to infer the parameters that determine the process.

2. **Prediction** we want to use our current beliefs and uncertainty to predict future outcomes in the world.

3. **Model comparison** we want to be able to tell which models or hypothesis provides a better account of the data.
Many ways to do this, but our final estimate of $p(h|d)$ gives use estimates of various h’s. In the coin example each h corresponds to a different theta.

Thus, the **maximum likelihood** hypothesis corresponds to our best estimate of the value of theta (turns out it is in simple the relative frequency of heads vs. tails in the case we considered)

Other approaches include more information about the distribution of $p(h|d)$ (mean, 95% confidence interval, etc...)
Prediction

To predict a new value \( y \):

\[
p(y) = \int p(y|h)p(h)\,dh
\]

In terms of the coin example, we want to predict if the next thing to come up is going to be heads or not. The basic idea is to compute a weighted average of the probability of getting a heads on a single coin flip \( p(y|\theta) \) times the probability that we currently believe the true bias of the coin is \( \theta \) summed over all the situations.

\[
p(H e a d s) = \sum p(H e a d s|\theta)p(\theta)
\]

G89.3405 Spr. 2009
Model Comparison

We’ve covered this in the past. Basically we can compare the likelihood of one model to another using a likelihood ratio.

A key aspect of Bayesian model selection is that it naturally takes into account model complexity (unlike AIC).
Ok, but what about incremental learning?

Sequence of data D1, D2, D3, D4.... How should we update our beliefs on a trial-by-trial basis?

\[
p(h|d_1) = \frac{p(d_1|h)p(h)}{p(d_1)}
\]

\[
p(h|d_1, d_2) = \frac{p(d_2|h, d_1)p(h|d_1)}{p(d_2|d_1)} = \frac{p(d_2|h, d_1)}{p(d_2|d_1)}
\]

\[
p(d_2|h, d_1) = p(d_2|h)
\]

\[
p(d_2|d_1) = p(d_2)
\]
Ok, but what about incremental learning?

Sequence of data D1, D2, D3, D4…. How should we update our beliefs on a trial-by-trial basis?

\[
p(h|d_1) = \frac{p(d_1|h)p(h)}{p(d_1)}
\]

\[
p(h|d_1, d_2) = \frac{p(d_2|h, d_1)p(h|d_1)}{p(d_2|d_1)}
\]

\[
p(h|d_1, d_2) = \frac{p(d_2|h)p(d_1|h)p(h)}{p(d_1)p(d_2)}
\]

\[
p(d_2|h, d_1) = p(d_2|h)
\]

\[
p(d_2|d_1) = p(d_2)
\]

Monday, March 2, 2009
Ok, but what about incremental learning?

Sequence of data D1, D2, D3, D4…. How should we update our beliefs on a trial-by-trial basis?

\[
p(h|d_1, d_2) = \frac{p(d_2|h) \frac{p(d_1|h)p(h)}{p(d_1)}}{p(d_2)}
\]

\[
p(h|d_1, d_2) = \frac{p(d_2|h) \frac{p(d_1|h)}{p(d_2)}}{p(d_1)} p(h)
\]

\[
p(h|d_1, d_2) = \frac{p(d_1|h) \frac{p(d_2|h)}{p(d_2)}}{p(d_1)} p(h)
\]

Exchangability!!

Order doesn’t matter!!
Incremental Learning

\[ p(h \mid d) = \frac{p(d \mid h)p(h)}{p(d)} \]

Iterate:

Defines an “ideal learner”
Incremental Learning

\[ p(h|d) = \frac{p(d|h)p(h)}{p(d)} \]

Iterate:

- No (explicit) Learning Rate!
- No Momentum!

Defines an “ideal learner”
A simple example

Rectangle game
A SIMPLE “INFORMATION SEARCH” GAME: BATTLESHIP

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

computation and cognition lab // new york university
A SIMPLE “INFORMATION SEARCH” GAME: BATTLESHIP

- 3 rectangular shaped object “hidden” in a grid
- The rectangles don’t overlap
- On each trial you can turn one rectangle over and reveal if it is a “miss” or a “hit” (of a particular ship)
- You loose a point for each observation you make
- Goal is to learn the exact size and location of each rectangle in as few observations as possible
Tenenbaum (1999) presented a simple Bayesian learner that updates beliefs about the location of a unknown rectangle hidden in a grid.

Each “hypothesis” is represented by a quadruple \((l_1, l_2, s_1, s_2)\) representing the coordinates of the upper left corner of each rectangle \((l_1, l_2)\) and the height and width \((s_1, s_2)\).

The likelihood function \(p(X|h)\) is simple: If data item \(X\) is in the rectangle, the likelihood is assigned a 1.0 otherwise 0.0.

\[
p(X|h) = \begin{cases} 1 & \text{if } X \in h \\ 0 & \text{otherwise} \end{cases}
\]
\[ p(h|X) = \frac{p(X|h)p(h)}{p(X)} \]

where
\[ p(X) = \sum_{h \in H} p(X|h)p(h) \]

Iterate:

\[ p(h|X) = \frac{p(X|h)p(h)}{p(X)} \]

Defines an “ideal learner”
A simple measure to summarize our uncertainty about something

Luce (2003) has a nice paper titled “Whatever happened to information theory in psychology?”

\[
H(X) = - \sum_x p(x) \cdot \log(p(x))
\]

**Dirac function**

\[
H(X) = 0
\]

**Uniform distribution**

\[
H(X) = 1.0
\]
HIGH ENTROPY

Lots of uncertainty about the true value
LOW ENTROPY (PINK)

Less of uncertainty about the true value
Intuition is that one measure of the “value” of information is how much changes our uncertainty.

Mutual Information Gain measures the change in entropy as we receive a new piece of information (e.g. how does our confidence in our beliefs change from seeing a new piece of data?)

\[ I(X; Y) = H(X) - H(X|Y) \]

Read as “how much does our uncertainty about X change given that we know Y”.

If \( H(X) = H(X|Y) \), the Information gain is 0.0

If Y gives reduces any uncertainty about X, then \( I(X;Y) \) will be positive.
Following each sample (and corresponding feedback) we compute the expected information gain from each remaining point in the grid.

\[
E[I(y)] = \sum_C [p(x \in C) \cdot [H(C) - H(C|y \in C)] + (1 - p(x \in C)) \cdot [H(C) - H(C|y \notin C)]]
\]

where

\[
p(y \in C) = \sum_{h \in C} p(y|h)p(h)
\]
NO CHANGE IN ENTROPY

However, seems like new knowledge was acquired
THUS, THIS IS JUST ONE MEASURE OF THE QUALITY OF INFORMATION

- Recently, Nelson (2007) has examined a number of information sampling norms including maximum expected information gain, max p, absolute impact, etc...

- Generally all these ways of characterizing the “value” of information are roughly the same in practice, although there are places where they can be distinguished (to date this has been done with probabilistic reasoning tasks rather than in information-search contexts)

- Right now we’ve focused on expected information gain although plan to consider others
<table>
<thead>
<tr>
<th>Hidden Board (current sample in blue)</th>
<th>Cat A Posterior</th>
<th>Cat B Posterior</th>
<th>Cat C Posterior</th>
<th>Expected Information Gain</th>
</tr>
</thead>
</table>

Monday, March 2, 2009
INFORMAVORE IN ACTION

Hidden Board (current sample in blue)  Cat A Posterior  Cat B Posterior  Cat C Posterior  Expected Information Gain

computation and cognition lab // new york university

Monday, March 2, 2009
<table>
<thead>
<tr>
<th>Hidden Board (current sample in blue)</th>
<th>Cat A Posterior</th>
<th>Cat B Posterior</th>
<th>Cat C Posterior</th>
<th>Expected Information Gain</th>
</tr>
</thead>
</table>

---

Monday, March 2, 2009
INFORMAVORE IN ACTION

Hidden Board (current sample in blue)

Cat A Posterior

Cat B Posterior

Cat C Posterior

Expected Information Gain

computation and cognition lab // new york university

Monday, March 2, 2009
<table>
<thead>
<tr>
<th>Hidden Board (current sample in blue)</th>
<th>Cat A Posterior</th>
<th>Cat B Posterior</th>
<th>Cat C Posterior</th>
<th>Expected Information Gain</th>
</tr>
</thead>
</table>

Monday, March 2, 2009
INFORMAVORE IN ACTION

Hidden Board (current sample in blue)

Cat A Posterior

Cat B Posterior

Cat C Posterior

Expected Information Gain

Monday, March 2, 2009
<table>
<thead>
<tr>
<th>Hidden Board (current sample in blue)</th>
<th>Cat A Posterior</th>
<th>Cat B Posterior</th>
<th>Cat C Posterior</th>
<th>Expected Information Gain</th>
</tr>
</thead>
</table>

Monday, March 2, 2009
INFORMAVORE IN ACTION

Hidden Board (current sample in blue)
Cat A Posterior
Cat B Posterior
Cat C Posterior
Expected Information Gain

computation and cognition lab // new york university

Monday, March 2, 2009