Signal Detection Theory and Bayesian Modeling

COGS 202: Computational Modeling of Cognition
Omar Shanta, Shuai Tang, Gautam Reddy,
Reina Mizrahi, Mehul Shah
Detection Theory and Psychophysics: A Review
Swets (1961)
“Over ensuing decades, the SD model, with only technical modifications to accommodate particular applications, has become almost universally accepted as a theoretical account of decision making in research on perceptual detection and recognition and in numerous extensions to applied domains”

- (Swets, 1988; Swets, Dawes, & Monahan, 2000)

This development may well be regarded as the most towering achievement of basic psychological research of the last half century.
When does Signal Detection Theory apply?

1. There are two true states of the world
2. An imperfect diagnostic procedure is used to make a decision

(Macmillian, and Creelman, 2005; Wixted et al., 2014)
Possible Responses

True State

Diagnostic Decision

<table>
<thead>
<tr>
<th>Present</th>
<th>Absent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit</td>
<td>Miss</td>
</tr>
<tr>
<td>False Alarm</td>
<td>Correct Rejection</td>
</tr>
</tbody>
</table>
Possible Responses

True State

Present

Absence

Diagnostic Decision

Present

Absent

Hit

Miss

False Alarm

Correct Rejection

False Alarm - Type I Error
Miss - Type II Error
What is signal detection theory?

- A combination of two theoretical structures:
  - Statistical decision theory
    - Used to treat the detection task as a decision process
    - Allows TSD to deal with the problem of *criterion for signal existence* that is employed by observer
  - Theory of ideal observer
    - Makes it possible to relate the level of detection performance attained by a real observer to the mathematically ideal detection performance
Why use signal detection theory?

- Allows us to establish a framework for experimental study or sensory systems
- The theory specifies the nature of the detection process
- Defines the experimental methods that are appropriate
- Deals with a key issue in psychophysics (and other related fields)

Most Importantly:

- *It allows us to define the criterion for making a positive response and their sensitivity to the presence or absence of a stimulus*
Why use signal detection theory?

- Allows us to establish a framework for experimental study or sensory systems
- The theory specifies the nature of the detection process
- Defines the experimental methods that are appropriate
- Deals with a key issue in psychophysics (and other related fields)

Most Importantly:

- *It allows us to define the **criterion** for making a positive response and their sensitivity to the presence or absence of a stimulus*
The Fundamental Detection Problem

“The observation interval contains either the noise alone, or a specified signal and the noise … and can be represented as a unidimensional variable.”
Likelihood Ratio Criterion

Can be thought of as similar to SNR

Likelihood Ratio: $L(y) = \frac{P(y|\text{present})}{P(y|\text{absent})}$, $y \in \{\text{present, absent}\}$

If $y = \text{present}$, the ratio is analogous to SNR

$L(\text{present}) = \frac{P(\text{present}|\text{present})}{P(\text{present}|\text{absent})}$ *When $L$ is large, the decision will be that signal is present.

$\text{SNR} = \frac{\text{Power}(\text{Signal})}{\text{Power}(\text{Noise})}$ *When SNR is large, the decision will be that signal is present.
Discriminability: $d'$

The degree to which the signals associated with target being present or absent are separated using a particular diagnostic procedure.

Good approximation to behavior of human observer

$$D' = \Sigma^{-1}(\mu_1 - \mu_2)$$
Discriminability: $d'$

The degree to which the signals associated with target being present or absent are separated using a particular diagnostic procedure.

Good approximation to behavior of human observer

$$D' = \Sigma^{-1}(\mu_1 - \mu_2)$$

Decision Boundary of Bayes Decision Rule in Equal Covariance Case!!
The Observer’s Criterion and the Optimal Criterion

Correct ID Rate = 0.98
False Alarm ID Rate = 0.50

Correct ID Rate = 0.84
False Alarm ID Rate = 0.16

Correct ID Rate = 0.50
False Alarm ID Rate = 0.02
The Observer’s Criterion and the Optimal Criterion

Correct ID Rate = 0.98
False Alarm ID Rate = 0.50

Correct ID Rate = 0.84
False Alarm ID Rate = 0.16

Correct ID Rate = 0.50
False Alarm ID Rate = 0.02
The Observer’s Criterion and the Optimal Criterion

Correct ID Rate = 0.98
False Alarm ID Rate = 0.50

Correct ID Rate = 0.84
False Alarm ID Rate = 0.16

Correct ID Rate = 0.50
False Alarm ID Rate = 0.02
The Observer’s Criterion and the Optimal Criterion

Correct ID Rate = 0.98
False Alarm ID Rate = 0.50

Correct ID Rate = 0.84
False Alarm ID Rate = 0.16

Correct ID Rate = 0.50
False Alarm ID Rate = 0.02

“Lax”

“Moderate”

“Strict”
The Observer’s Criterion and the Optimal Criterion

Correct ID Rate = 0.98
False Alarm ID Rate = 0.50

Correct ID Rate = 0.84
False Alarm ID Rate = 0.16

Correct ID Rate = 0.50
False Alarm ID Rate = 0.02
Receiver Operating Characteristic (ROC)
Implications for Psychophysical Methods

SDT provides a framework for the study of sensory systems with richer information.

1) The most salient implication is the ability to perform and analyze catch trials (trials containing noise alone) and the treatment of positive responses
   a) **Not Adequate**: Simply remind observer to avoid false-positives when one is made after a few catch trials
      i) Drives criterion for decisions up to a point where it cannot be measured
   b) **Adequate**: Use enough catch trials to obtain good estimate of true response criterion

2) SDT Requirements:
   a) Measurable Noise / Measurable Response Criterion (LR)
   b) Enough catch trials
   c) Analysis yielding measure of sensitivity independent of Response Criterion
Theory of Ideal Observers

“Makes it possible to relate the level of detection performance attained by a real observer to the mathematically ideal detection performance”

Maximum $d’$ of Ideal Observer

$$d’=\sqrt{2E/N_0}$$

$E$: Signal Energy

$N_0$: Noise Power

$d’$ of Real Observer

$$d’_{obs}=\eta*\sqrt{2E/N_0}$$

*Relation through $\eta$: Varying $\eta$ can determine range over which humans adjust the parameters of their sensory system (likelihood criterion)
Theory of Ideal Observers

New Question: What is the nature of the discrepancy between real and ideal observer?

1) Humans perform worse for case of “signal specified exactly”
   a) Human Observer has “noisy decision process” (Variable Criterion)
   b) Noise in human sensory systems (senses)
   c) Faulty Memory in humans
      i) Forgetting signal characteristics introduces uncertainty into the model and it will have suboptimal performance
      ii) Introducing a memory aid increases performance of human observer
Examples and Extended Applications of SDT

Potential uses include problems that:

1) Lengthen Observation Interval
2) Number of Observations before a decision increases
3) Number of signals in an interval increases
4) Number of observers concentrating on same signal increases

SDT has been applied to:

1) Recognition/Identification instead of Detection
2) Deferred Decision: Observer decides to make decision now or get another sample
3) Speech Communication
Bayesian Modeling
Probabilistic Models of Cognition: Foundation

- How is Cognitive Science related to Probability?
- What applications does this relationship have?
- What scientific works helped further the field?
Probabilistic Models of Cognition: Foundation

- ‘Sophisticated’ probabilistic models applied to graphs and grammars
- Probability theory is two-fold
  - Normative - reasoning events (mathematics)
  - Descriptive - analyze actions and situations (psychology)
The Role of Mathematics

● Math usually seen as formal, disciplined, focused on patterns

● Provides framework for cognitive theories in many fields
  ○ NLP
  ○ Vision/Motor control
  ○ Learning
  ○ Etc.
Cognition and Probability

- Cognition involves inferring new information from prior input
- Probability is “calculus” for this type of uncertain inference
  - Subjective Interpretation of Probability
- Connectionist network vs. symbolic rule-based processing
Subjective Probability

- ‘Frequentist’ interpretation - coin flips, dice rolls
- ‘Subjective’ interpretation - ‘degrees of belief’
  - My degree of belief for a coin under a table \(\sim \frac{1}{2}\)
  - May increase to 1 as I look underneath the table
  - My friend may have different ‘subjective’ probability of the same event
Subjective Probability

- Justification for this interpretation
  - Cox’s Axioms - common sense, consistency, divisibility/comparability
  - Dutch Book - rational people have subjective probabilities
- Applied towards conditional probabilities: Bayes’ Theorem

\[
P(h|d) = \frac{P(d|h) \cdot P(h)}{P(d)}
\]
What is a Sophisticated Probability Model?

- Cognitive agents modeled through probability distributions
- Reasoning is modeled with techniques from statistical learning
- Meant for systems like intelligent machines, computer vision, machine learning
Sophisticated Prob Models - Early Examples

<table>
<thead>
<tr>
<th>Ulf Grenander - Brown U. Statistician and Applied Mathematics</th>
<th>Created vocabulary to help machines recognize patterns in the world</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judea Pearl - UCLA CS and COGS</td>
<td>Championed probabilistic approach to AI. Bayesian Networks - Probabilistic model through directed acyclic graphs</td>
</tr>
</tbody>
</table>
Applied Sophisticated Probabilistic Models

- Applying towards human cognition is not straightforward
- Probabilistic mind often has poor judgement
  - Fall victim to various probabilistic fallacies
- Ideal for well-optimized cognitive processes (vision, motor control)
Probabilistic Models of Vision

- Most advanced models in cognitive science come in vision-related applications
  - Ideal observer to Bayesian Theory

![Image](chart.png)

- Proximity
  (We organize nearby objects together. Thus, you should see columns because the dots in columns are closer than the dots in rows.)

- Similarity
  (We organize together objects that are similar in shape. Thus, you should see rows instead of columns.)

- Good Continuation
  (We organize lines to minimize abrupt changes. Thus, you should see the curved line as one, with a straight diagonal line cutting through it.)

- Closure
  (We organize lines to create whole figures when possible. Thus, you should organize this figure as a square in spite of the gap.)
Probabilistic Models of Vision

- ‘Grammatical’ models of vision
- Stochastic grammars for image parsing
- Probabilistic models of language processing, psycholinguistics
- Analysis-By-Synthesis
  - Use Bayes Decision Rule (Bayes Theorem)
## Impact on Causal Learning

<table>
<thead>
<tr>
<th>Model</th>
<th>Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>P.W. Cheng’s Causal Power</td>
<td>How variables affect other variables Parameter estimation in bayesian networks</td>
</tr>
<tr>
<td>Tenenbaum &amp; Griffith’s Causal Structure Judgements</td>
<td>Bayesian model selection with networks</td>
</tr>
<tr>
<td>Alison Gopnik’s Causal Learning Model</td>
<td>Model children’s causal learning with Bayesian networks</td>
</tr>
</tbody>
</table>
Human Causal Learning

- Rescorla-Wagner associative learning model
  \[ \Delta V_{x}^{n+1} = \alpha_x \beta (\lambda - V_{tot}) \]
  \[ V_{x}^{n+1} = V_{x}^{n} + \Delta V_{x}^{n+1} \]
- Peter Dayan and colleagues - Animal learning behavior
- Explain structure of computational and neural mechanisms of the brain
Marr’s Levels of Probabilistic Explanation

<table>
<thead>
<tr>
<th>Computational theory</th>
<th>Representation and algorithm</th>
<th>Hardware implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is the goal of the computation, why is it appropriate, and what is the logic of the strategy by which it can be carried out?</td>
<td>How can this computational theory be implemented? In particular, what is the representation for the input and output, and what is the algorithm for the transformation?</td>
<td>How can the representation and algorithm be realized physically?</td>
</tr>
</tbody>
</table>

*Figure 1–4.* The three levels at which any machine carrying out an information-processing task must be understood.
Computational Theory

- Cognitive Processes are naturally subjective and uncertain
- Focused on nature of problem being solved
- Does not matter how the system solves the problem
- Belief that cognitive systems operate via ‘heuristic tricks’
  - Learn through past experiences
Representation and Algorithms

- Human cognition is very flexible
- Some problems can be solved through explicit probabilistic methods
  - Stochastic grammars for language or vision
  - Bayesian Networks
- Bayesian computations are difficult when scaled up
- Sophisticated Probabilistic Models in various applications
  - Hypotheses for algorithms that make up probabilistic cognition
Hardware Implementation

- Naturally maps onto computational architectures
  - Distributed, autonomous, running in parallel
  - Qualitative features of neural architecture
- The brain and nervous system may act using probabilistic models
  - Research done in computational neuroscience
  - Further supports probabilistic cognition
Bayesian Modeling

What is Bayesian Inference?

----The idea that we can infer the nature of the source of data using tools from probability theory.

Allows us to -

Compare hypotheses using Bayes rule:

----A proper method for comparing hypotheses of varying complexity.

Impose structure to probabilistic models:

----Graphical models and efficient algorithms to perform inference and sampling.
Fundamentals of Bayesian Inference

- Basic Bayes
- Comparing two simple hypotheses
- Comparing infinitely many hypotheses
- Comparing hypotheses that differ in complexity
Notation:

$H$ is the set of all hypotheses considered by agent

$h$ is one hypothesis in $H$

$d$ is the data observed by agent
This formulation of Bayes’ rule makes it apparent that the posterior probability of \( h \) is proportional to the product of the prior probability and the likelihood.
2/ Comparing two simple hypotheses

- $d$ is the observation that you got. For example: $HHTHTHTTHT$
- $h_0$: probability of heads is 0.5
- $h_1$: probability of heads is 0.9
- $\theta$ denotes the probability of heads, either 0.5 or 0.9.

With no priori reason to favor one hypothesis over the other, we can take

$$P(h_0) = P(h_1) = 0.5$$
2/ Comparing two simple hypotheses

According the Bernoulli distribution, we have

\[ P(d|\theta) = \theta^{NH} (1 - \theta)^{NT} \]

Then, we want to compute \( P(h_0|d) \) and \( P(h_1|d) \)

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

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

Posterior odds:

\[
\frac{P(h_1|d)}{P(h_0|d)} = \frac{P(d|h_1)}{P(d|h_0)} \frac{P(h_1)}{P(h_0)}
\]
2/ Comparing two simple hypotheses

\[
\frac{P(h_1|d)}{P(h_0|d)} = 357 : 1
\]

\[
\frac{P(h_0|d)}{P(h_1|d)} = 165 : 1
\]
3/ Comparing infinitely many hypotheses

How can we infer $\theta$ from a sequence like $HHHHHHHHHH$?

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

\[
P(d) = \int_0^1 P(d|\theta)p(\theta)d\theta
\]
3/ Comparing infinitely many hypotheses

Assume a prior: suppose $p(\theta)$ follows a uniform distribution.

\[
P(d|\theta) = \theta^{N_H} (1 - \theta)^{N_T}
\]

\[
p(\theta|d) = \frac{P(d|\theta)p(\theta)}{P(d)}\quad P(d) = \int_0^1 P(d|\theta)p(\theta)d\theta
\]

\[
p(\theta|d) = \frac{(N_H + N_T + 1)!}{N_H! N_T!} \theta^{N_H} (1 - \theta)^{N_T}
\]
3/ Comparing infinitely many hypotheses

Assumption 1: $p(\theta)$ follows uniform distribution.

There are two methods to obtain a point estimate from a posterior distribution.

1. The maximum a posteriori (MAP) for $\theta$ is $\frac{N_H}{N_H+N_T}$, which is also the maximum-likelihood estimation.
2. The posterior mean is $\frac{(N_H+1)}{(N_H+N_T+2)}$.

Example: HHHHHHHHHH?
3/ Comparing infinitely many hypotheses

Assumption1: $p(\theta)$ follows uniform distribution.

The posterior mean is sensitive to the fact that we might not want to put as much weight in a single head as a sequence of ten heads in a row.
3/ Comparing infinitely many hypotheses

Assumption 2. $\theta \sim \text{Beta}(V_H+1, V_T+1)$, $V_H = V_T = 1000$

$$p(\theta|d) = \frac{(N_H + N_T + V_H + V_T + 1)!}{(N_H + V_H)! (N_T + V_T)!} \theta^{N_H + V_H} (1 - \theta)^{N_T + V_T}$$

MAP for $\theta$: $(N_H + V_H)/(N_H + N_T + V_H + V_T)$

Posterior mean for $\theta$: $(N_H + V_H + 1)/(N_H + N_T + V_H + V_T + 2)$
Comparing hypotheses that differ in complexity

$h_0$ is the hypothesis that $\theta = 0.5$

$h_1$ is the hypothesis that $\theta$ takes a value drawn from a uniform distribution on $[0,1]$

With no priori reason to favor one hypothesis over the other, we can take

$$P(h_0) = P(h_1) = 0.5$$
4/ Comparing hypotheses that differ in complexity

First, compute the likelihood under each hypothesis.

\[ P(d \mid h_0) = 0.5^{N_H + N_T} \]

\[
P(d|h_1) = \int_0^1 P(d \mid \theta, h_1)p(\theta \mid h_1)d\theta
\]

\[
= \frac{(N_H + N_T + 1)!}{N_H!N_T!} = \frac{1}{B(N_H + 1, N_T + 1)}
\]
4/ Comparing hypotheses that differ in complexity

Second, we plot the posterior odds in favor of $h_1$. 

![Graph showing log posterior odds in favor of $h_1$ vs. number of heads ($N_{H}$).]
4/ Comparing hypotheses that differ in complexity

Third, we plot the posterior probability v.s. $\theta$
4/ Comparing hypotheses that differ in complexity

Conclusion:

(1) Complex hypotheses have more degrees of freedom that can be adapted to the data, and can thus always be made to fit better than simple hypotheses.

(2) But this flexibility also makes a worse fit possible.

Advice:

A more complex hypothesis will be favored only if its greater complexity consistently provides a better account of data.
Representing structured probability distributions

Directed graphical models (Bayesian networks)

Undirected graphical models (Markov Random Fields)

Uses of graphical models
1/ Directed graphical models

Suppose we have \( n \) binary random variables, the joint distribution

\[
P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)
\]

involves \( O(2^n) \) numbers for random variables.
1/ Directed graphical models

\[ P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) \]

\[ = \prod_{i=1}^{n} P(X_i = x_i | X_1 = x_1, X_2 = x_2, \ldots, X_{i-1} = x_{i-1}) \]

\[ = P(X_1 = x_1) \cdot P(X_2 = x_2 | X_1 = x_1) \cdot P(X_3 = x_3 | X_2 = x_2, X_1 = x_1) \]

\[ \cdots P(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_2 = x_2, X_1 = x_1) \]

The numbers for binary random variables = \(2^0 + 2^1 + 2^2 + \ldots + 2^{n-1} = 2^n - 1\)

(1) Can we have more compact representations?
(2) Can we have more efficient algorithms?
1/ Directed graphical models

Guidelines:

(1) Bayes nets use acyclic graphs, which means no node can be its own ancestor.
(2) Conditional independence is a basic tool to do inference in Bayes nets.

\[ P(x_1, x_2, \ldots, x_m) = \prod_i P(x_i | Pa(X_i)) \] where \( Pa(X_i) \) is the set of parents of \( X_i \)
1/ Directed graphical models

Simple example from Prof. Saul.

B=burglary happens

E=earthquake happens

A=alarm rings

J=John calls

M=Mary calls
1/ Directed graphical models

(1) Nodes represent random variables
(2) Edges represent conditional dependencies
1/ Directed graphical models

\[ P(M, J, A, B, E) \]
\[ = P(M)P(J|M)P(A|J, M)P(B|A)P(E|A, B) \]

The total numbers used for representing joint probability is 13 now, rather than 31.

31 = \(2^5 - 1\)

13 = 1+2+4+2+4
1/ Directed graphical models

Causal Graphical Models

An edge is assumed to indicate a direct causal relationship.

(1) Representing the probabilities of events that one might observe
(2) Representing the probabilities of events that one can produce through intervening on a system
2/ Undirected graphical models/Markov random fields

(1) Nodes represent random variables
(2) Undirected edges define a neighborhood structure on the graph
2/ Undirected graphical models

(1) Used to model image data in vision
(2) Used as generative models for textures
(3) Used for image segmentation
3/ Uses of graphical models

- Intuitive representation of causal relationships
- Efficient algorithms for learning and inference
- Applications:
  - Language modeling
  - Mixture models
  - Artificial neural networks (Hopfield nets, Boltzmann machines)
Algorithms for approximate inference

(1) Exact inference is computationally expensive, particularly with loopy graphs

(2) Popular algorithms:
   (a) Markov Chain Monte Carlo (MCMC) methods
   (b) The Expectation-Maximization (EM) Algorithm
Approximate inference

(1) Can write down joint pdf easily as a product of marginal distributions involving nodes and their parents.

(2) Draw samples from the joint pdf and estimate other marginal distributions.

\[ P(M = m| B = b) \sim N(M = m, B = b)/N(B=b) \]

This is too slow, there are smarter methods.
1/ Markov Chain Monte Carlo (MCMC) methods

Estimating marginal distribution, probability of the query given the evidence (data)

\[ P(Q = q| E = e) \]

- Instantiate evidence vectors.
- Initialize other variables
- Sample from conditional distribution of an RV over its Markov blanket.
- Repeat step 3 \( N \) times and we have

\[ P(Q = q| E = e) \sim \frac{N(q)}{N} \]
Gaussian mixture models/latent variables
2/ The Expectation-Maximization (EM) Algorithm

- Used to learn parameters using maximum likelihood estimation.
- Finds parameters that maximizes likelihood of observed data

**E step:** evaluating the expectation of the ‘complete log-likelihood’ $\log P(x,z|\theta)$ with respect to $P(z|x,\theta)$

**M step:** maximizing the resulting quantity with respect to $\theta$

\[ \rightarrow Z \rightarrow X \]
2/ The Expectation-Maximization (EM) Algorithm

EM algorithm applies to a Gaussian mixture model with two clusters.

1) Use current estimate of $\theta$ to estimate $Z$ in the form of $P(z|x,\theta)$
2) Use estimate of $Z$ to update $\theta$. 
Conclusion

- Probabilistic models are powerful tools to model the complexity of human cognition.
- They allow us to develop intuitive models of human rationality.
Discriminability: The degree to which the signals associated with target being present or absent are separated using a particular diagnostic procedure.