Outline

• Limits of Bayesian classification
• Bayesian concept learning
• Probabilistic models for unsupervised and semi-supervised category learning
Limitations

• Is categorization just discrimination among mutually exclusive classes?
  – Overlapping concepts? Hierarchies? “None of the above”? Can we learn a single new concept?

• Are most categories Gaussian, or any simple parametric shape?
  – What about superordinate categories?
  – What about learning rule-based categories?

• How do we learn concepts from just a few positive examples?
  – Learning with high certainty from little data.
  – Generalization from one example.
Feldman (1997)

Here is a blicket:

Please draw six more blickets.
Feldman (1997)

Image removed due to copyright considerations.
Feldman (1997)

Image removed due to copyright considerations.
Limitations

• Is prototypicality = degree of membership?
  – Armstrong et al.: No, for classical rule-based categories
  – Not for complex real-world categories either: “Christmas eve”, “Hollywood actress”, “Californian”, “Professor”
  – For natural kinds, huge variability in prototypicality independent of membership.

• Richer concepts?
  – Meaningful stimuli, background knowledge, theories?
  – Role of causal reasoning? “Essentialism”?

• Difference between “perceptual” and “cognitive” categories?
Outline

• Limits of Bayesian classification
• Bayesian concept learning
• Probabilistic models for unsupervised and semi-supervised category learning
Concepts and categories

• A category is a set of objects that are treated equivalently for some purpose.
• A concept is a mental representation of the category.
• Functions for concepts:
  – Categorization/classification
  – Prediction
  – Inductive generalization
  – Explanation
  – Reference in communication and thought
Everyday concept learning

- Learning words from examples

Image removed due to copyright considerations.
Everyday concept learning

- Learning words from examples
- Inductive generalization

Squirrels have biotinic acid in their blood.
Gorillas have biotinic acid in their blood.  

---

Horses have biotinic acid in their blood.
Tenenbaum (2000)

• Takes reference and generalization as primary.
• Concept is a pointer to a set of things in the world.
  – Learner constructs a hypothesis space of possible sets of entities (as in the classical view).
  – You may not know what that set is (unlike in the classical view).
  – Through learning you acquire a probability distribution over possible sets.
The number game

- Program input: number between 1 and 100
- Program output: “yes” or “no”
The number game

• Learning task:
  – Observe one or more positive ("yes") examples.
  – Judge whether other numbers are "yes" or "no".

Image removed due to copyright considerations.
The number game

<table>
<thead>
<tr>
<th>Examples of “yes” numbers</th>
<th>Generalization judgments ($N = 20$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Diffuse similarity</td>
</tr>
</tbody>
</table>

Image removed due to copyright considerations.
### The number game

<table>
<thead>
<tr>
<th>Examples of “yes” numbers</th>
<th>Generalization judgments ($n = 20$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>Diffuse similarity</td>
</tr>
<tr>
<td>60 80 10 30</td>
<td>Images removed due to copyright considerations.</td>
</tr>
</tbody>
</table>

**Rule:**
“multiples of 10”
# The number game

<table>
<thead>
<tr>
<th>Examples of “yes” numbers</th>
<th>Generalization judgments (N = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
</tr>
<tr>
<td>60 80 10 30</td>
<td>Diffuse similarity</td>
</tr>
<tr>
<td>60 52 57 55</td>
<td>Rule: “multiples of 10”</td>
</tr>
<tr>
<td></td>
<td>Focused similarity: numbers near 50-60</td>
</tr>
</tbody>
</table>

Images removed due to copyright considerations.
The number game

<table>
<thead>
<tr>
<th>Examples of “yes” numbers</th>
<th>Generalization judgments ($N = 20$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>Diffuse similarity</td>
</tr>
<tr>
<td>16 8 2 64</td>
<td>Rule: “powers of 2”</td>
</tr>
<tr>
<td>16 23 19 20</td>
<td>Focused similarity: numbers near 20</td>
</tr>
</tbody>
</table>
The number game

Main phenomena to explain:

– Generalization can appear either similarity-based (graded) or rule-based (all-or-none).
– Learning from just a few positive examples.

60

60 80 10 30

Diffuse similarity

Images removed due to copyright considerations.

60 52 57 55

Rule:
“multiples of 10”

Focused similarity:
numbers near 50-60
Divisions into “rule” and “similarity” subsystems?

• Category learning
  – Nosofsky, Palmeri et al.: RULEX
  – Erickson & Kruschke: ATRIUM

• Language processing
  – Pinker, Marcus et al.: Past tense morphology

• Reasoning
  – Sloman
  – Rips
  – Nisbett, Smith et al.
Bayesian model

- **H**: Hypothesis space of possible concepts:
  - $h_1 = \{2, 4, 6, 8, 10, 12, \ldots, 96, 98, 100\}$ ("even numbers")
  - $h_2 = \{10, 20, 30, 40, \ldots, 90, 100\}$ ("multiples of 10")
  - $h_3 = \{2, 4, 8, 16, 32, 64\}$ ("powers of 2")
  - $h_4 = \{50, 51, 52, \ldots, 59, 60\}$ ("numbers between 50 and 60")
  - \ldots

Representational interpretations for *H*:

- Candidate rules
- Features for similarity
- “Consequential subsets” (Shepard, 1987)
Where do the hypotheses come from?

Additive clustering (Shepard & Arabie, 1977):

\[ s_{ij} = \sum_k w_k f_{ik} f_{jk} \]

- \( s_{ij} \): similarity of stimuli \( i,j \)
- \( w_k \): weight of cluster \( k \)
- \( f_{ik} \): membership of stimulus \( i \) in cluster \( k \)
  (1 if stimulus \( i \) in cluster \( k \), 0 otherwise)

Equivalent to similarity as a weighted sum of common features (Tversky, 1977).
Additive clustering for the integers 0-9:

\[ s_{ij} = \sum_k w_k f_{ik} f_{jk} \]

<table>
<thead>
<tr>
<th>Rank</th>
<th>Weight</th>
<th>Stimuli in cluster</th>
<th>Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.444</td>
<td>*</td>
<td>powers of two</td>
</tr>
<tr>
<td>1</td>
<td>0.345</td>
<td>* *</td>
<td>small numbers</td>
</tr>
<tr>
<td>2</td>
<td>0.331</td>
<td>* * *</td>
<td>multiples of three</td>
</tr>
<tr>
<td>3</td>
<td>0.291</td>
<td>* * * *</td>
<td>large numbers</td>
</tr>
<tr>
<td>4</td>
<td>0.255</td>
<td>* * * * *</td>
<td>middle numbers</td>
</tr>
<tr>
<td>5</td>
<td>0.216</td>
<td>* * * * *</td>
<td>odd numbers</td>
</tr>
<tr>
<td>6</td>
<td>0.214</td>
<td>* * * *</td>
<td>smallish numbers</td>
</tr>
<tr>
<td>7</td>
<td>0.172</td>
<td>* * * * *</td>
<td>largish numbers</td>
</tr>
</tbody>
</table>
Three hypothesis subspaces for number concepts

• Mathematical properties (24 hypotheses):
  – Odd, even, square, cube, prime numbers
  – Multiples of small integers
  – Powers of small integers

• Raw magnitude (5050 hypotheses):
  – All intervals of integers with endpoints between 1 and 100.

• Approximate magnitude (10 hypotheses):
  – Decades (1-10, 10-20, 20-30, …)
Bayesian model

- **$H$:** Hypothesis space of possible concepts:
  - Mathematical properties: even, odd, square, prime, . . . .
  - Approximate magnitude: {1-10}, {10-20}, {20-30}, . . . .
  - Raw magnitude: all intervals between 1 and 100.

- **$X = \{x_1, \ldots, x_n\}$:** $n$ examples of a concept $C$.

- Evaluate hypotheses given data:
  \[
p(h \mid X) = \frac{p(X \mid h)p(h)}{p(X)}
  \]
  - $p(h)$ [“prior”]: domain knowledge, pre-existing biases
  - $p(X|h)$ [“likelihood”]: statistical information in examples.
  - $p(h|X)$ [“posterior”]: degree of belief that $h$ is the true extension of $C$. 
Bayesian model

- **$H$:** Hypothesis space of possible concepts:
  - Mathematical properties: even, odd, square, prime, . . .
  - Approximate magnitude: \{1-10\}, \{10-20\}, \{20-30\}, . . .
  - Raw magnitude: all intervals between 1 and 100.

- **$X = \{x_1, \ldots, x_n\}$:** $n$ examples of a concept $C$.

- Evaluate hypotheses given data:

  \[
  p(h \mid X) = \frac{p(X \mid h) p(h)}{\sum_{h' \in H} p(X \mid h') p(h')}
  \]

  - $p(h)$ [“prior”]: domain knowledge, pre-existing biases
  - $p(X \mid h)$ [“likelihood”]: statistical information in examples.
  - $p(h \mid X)$ [“posterior”]: degree of belief that $h$ is the true extension of $C$. 

Likelihood: $p(X|h)$

- **Size principle**: Smaller hypotheses receive greater likelihood, and exponentially more so as $n$ increases.

$$p(X | h) = \left[ \frac{1}{\text{size}(h)} \right]^n \quad \text{if } x_1, \ldots, x_n \in h$$

$$= 0 \quad \text{if any } x_i \not\in h$$

- Follows from assumption of randomly sampled examples.
- Captures the intuition of a representative sample.
Illustrating the size principle

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>24</td>
<td>26</td>
<td>28</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>34</td>
<td>36</td>
<td>38</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>44</td>
<td>46</td>
<td>48</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>54</td>
<td>56</td>
<td>58</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>64</td>
<td>66</td>
<td>68</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>74</td>
<td>76</td>
<td>78</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>84</td>
<td>86</td>
<td>88</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>94</td>
<td>96</td>
<td>98</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>
Illustrating the size principle

<table>
<thead>
<tr>
<th></th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>14</td>
<td>16</td>
<td>18</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>24</td>
<td>26</td>
<td>28</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>34</td>
<td>36</td>
<td>38</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>44</td>
<td>46</td>
<td>48</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>52</td>
<td>54</td>
<td>56</td>
<td>58</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>62</td>
<td>64</td>
<td>66</td>
<td>68</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>72</td>
<td>74</td>
<td>76</td>
<td>78</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>82</td>
<td>84</td>
<td>86</td>
<td>88</td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>94</td>
<td>96</td>
<td>98</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

Data slightly more of a coincidence under $h_1$
Illustrating the size principle

Data much more of a coincidence under $h_1$
Relation to the “subset principle”

• Asymptotically equivalent
  – Subset principle = maximum likelihood
  – Size principle more useful when learning from just a few examples.

• Size principle is graded, while subset principle is all-or-none.

• Bayesian formulation allows the size principle to trade off against the prior.
Prior: \( p(h) \)

- Choice of hypothesis space embodies a strong prior: effectively, \( p(h) \sim 0 \) for many logically possible but conceptually unnatural hypotheses.

- Prevents overfitting by highly specific but unnatural hypotheses, e.g. “multiples of 10 except 50 and 70”.
Constructing more flexible priors

• Start with a base set of regularities $R$ and combination operators $C$.

• Hypothesis space = closure of $R$ under $C$.
  – $C = \{\text{and, or}\}$: $H =$ unions and intersections of regularities in $R$ (e.g., “multiples of 10 between 30 and 70”).
  – $C = \{\text{and-not}\}$: $H =$ regularities in $R$ with exceptions (e.g., “multiples of 10 except 50 and 70”).

• Two qualitatively similar priors:
  – Description length: number of combinations in $C$ needed to generate hypothesis from $R$.
  – Bayesian Occam’s Razor, with model classes defined by number of combinations: more combinations $\rightarrow$ more hypotheses $\rightarrow$ lower prior
Prior: $p(h)$

- Choice of hypothesis space embodies a strong prior: effectively, $p(h) \sim 0$ for many logically possible but conceptually unnatural hypotheses.

- Prevents overfitting by highly specific but unnatural hypotheses, e.g. “multiples of 10 except 50 and 70”.

- $p(h)$ encodes relative plausibility of alternative theories:
  - Mathematical properties: $p(h) \sim 1$
  - Approximate magnitude: $p(h) \sim 1/10$
  - Raw magnitude: $p(h) \sim 1/50$ (on average)

- Also degrees of plausibility within a theory, e.g., for magnitude intervals of size $s$:

  \[ p(s) = \frac{s}{\gamma} e^{-s/\gamma}, \quad \gamma = 10 \]
Hierarchical priors

- Higher-order hypothesis: is this coin fair or unfair?
- Example probabilities:
  - \( P(\text{fair}) = 0.99 \)
  - \( P(\theta|\text{fair}) \) is Beta(1000,1000)
  - \( P(\theta|\text{unfair}) \) is Beta(1,1)
- 25 heads in a row propagates up, affecting \( \theta \) and then \( P(\text{fair}|D) \)

\[
\frac{P(\text{fair}|25 \text{ heads})}{P(\text{unfair}|25 \text{ heads})} = \frac{P(25 \text{ heads}|\text{fair})}{P(25 \text{ heads}|\text{unfair})} \cdot \frac{P(\text{fair})}{P(\text{unfair})} = 9 \times 10^{-5}
\]
Hierarchical priors

- Higher-order hypothesis: is this concept mathematical or magnitude-based?
- Example probabilities:
  - $P(\text{magnitude}) = 0.99$
  - $P(h|\text{magnitude})$ ...
  - $P(h|\text{mathematical})$ ...
- Observing 8, 4, 64, 2, 16, … could quickly overwhelm this prior.
Posterior: \[ p(h \mid X) = \frac{p(X \mid h)p(h)}{\sum_{h' \in H} p(X \mid h')p(h')} \]

- \( X = \{60, 80, 10, 30\} \)
- Why prefer “multiples of 10” over “even numbers”? \( p(X \mid h) \).
- Why prefer “multiples of 10” over “multiples of 10 except 50 and 20”? \( p(h) \).
- Why does a good generalization need both high prior and high likelihood? \( p(h \mid X) \sim p(X \mid h) p(h) \)
Bayesian Occam’s Razor

Probabilities provide a common currency for balancing model complexity with fit to the data.

Figure by MIT OCW.
Generalizing to new objects

Given $p(h|X)$, how do we compute $p(y \in C | X)$, the probability that $C$ applies to some new stimulus $y$?
Generalizing to new objects

Hypothesis averaging:

Compute the probability that $C$ applies to some new object $y$ by averaging the predictions of all hypotheses $h$, weighted by $p(h|X)$:

$$p(y \in C \mid X) = \sum_{h \in H} \left( \sum_{y \in C \mid h} p(y \in C \mid h) \right) p(h \mid X)$$

$$= \sum_{h \supset \{y, X\}} p(h \mid X)$$
Examples:
16

Image removed due to copyright considerations.
Examples:

16
8
2
64

Image removed due to copyright considerations.
Examples:

16
23
19
20

Image removed due to copyright considerations.
<table>
<thead>
<tr>
<th>+ Examples</th>
<th>Human generalization</th>
<th>Bayesian Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 80 10 30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>60 52 57 55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 8 2 64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16 23 19 20</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Images removed due to copyright considerations.
Summary of the Bayesian model

- How do the statistics of the examples interact with prior knowledge to guide generalization?
  \[ \text{posterior} \propto \text{likelihood} \times \text{prior} \]

- Why does generalization appear rule-based or similarity-based?
  \[ \text{hypothesis averaging + size principle} \]
  \[ \downarrow \]
  - broad \( p(h|X) \): similarity gradient
  - narrow \( p(h|X) \): all-or-none rule
Summary of the Bayesian model

• How do the statistics of the examples interact with prior knowledge to guide generalization?
  \[
  \text{posterior} \propto \text{likelihood} \times \text{prior}
  \]

• Why does generalization appear rule-based or similarity-based?
  \[
  \text{hypothesis averaging} + \text{size principle}
  \]
  \[
  \text{Many } h \text{ of similar size: broad } p(h|X)
  \]
  \[
  \text{One } h \text{ much smaller: narrow } p(h|X)
  \]
Discussion points

• Relation to “Bayesian classification”?
  – Causal attribution versus referential inference.
  – Which is more suited to natural concept learning?

• Relation to debate between rules / logic / symbols and similarity / connections / statistics?

• Where do the hypothesis space and prior probability distribution come from?

• What about learning “completely novel concepts”, where you don’t already have a hypothesis space?
Hierarchical priors

Latent structure captures what is common to all coins, and also their individual variability.
Hierarchical priors

- Latent structure captures what is common to all concepts, and also their individual variability
- Is this all we need?
• Hypothesis space is not just an arbitrary collection of hypotheses, but a principled system.
• Far more structured than our experience with specific number concepts.