Semi-Supervised Learning in Computers and Humans

Xiaojin Zhu

jerryzhu@cs.wisc.edu
Computer Sciences
University of Wisconsin–Madison

Outline

1. A Human Learning Experiment
2. A Machine Learning Explanation
3. Opportunities for Further Research

Outline

1 A Human Learning Experiment

2 A Machine Learning Explanation

3 Opportunities for Further Research

A camping story

A camping story

A camping story

badger

A camping story

A camping story

A camping story

raccoon

A camping story

?

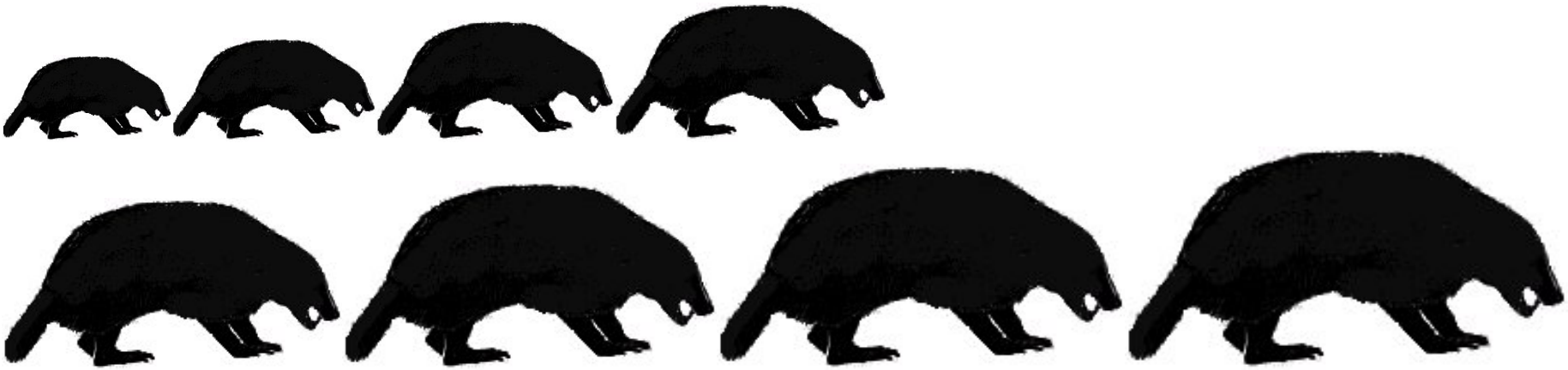
A camping story

raccoon

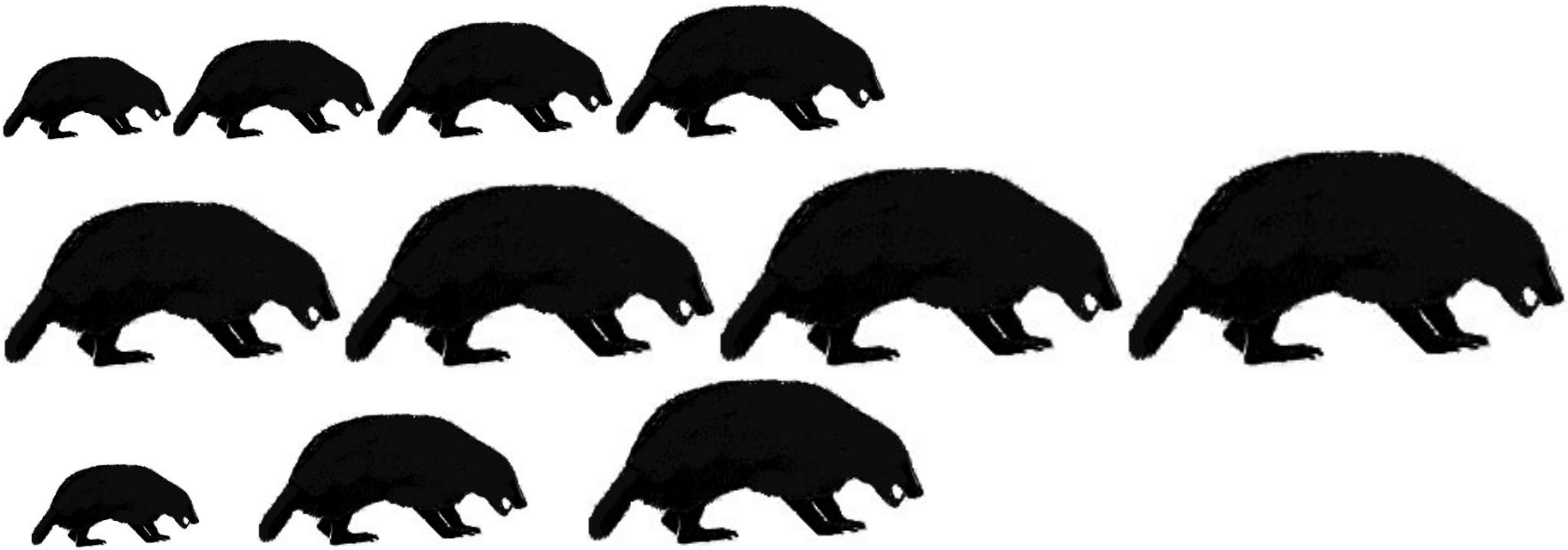
?

badger

A camping story



A camping story



Classification

- input instance x (e.g., size)

Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)

Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)
- labeled data (=supervised experiences) $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$

Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)
- labeled data (=supervised experiences) $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- unlabeled data (=unsupervised experiences) $X_u = \{x_{l+1:n}\}$

Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)
- labeled data (=supervised experiences) $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- unlabeled data (=unsupervised experiences) $X_u = \{x_{l+1:n}\}$
- usually $l \ll n$

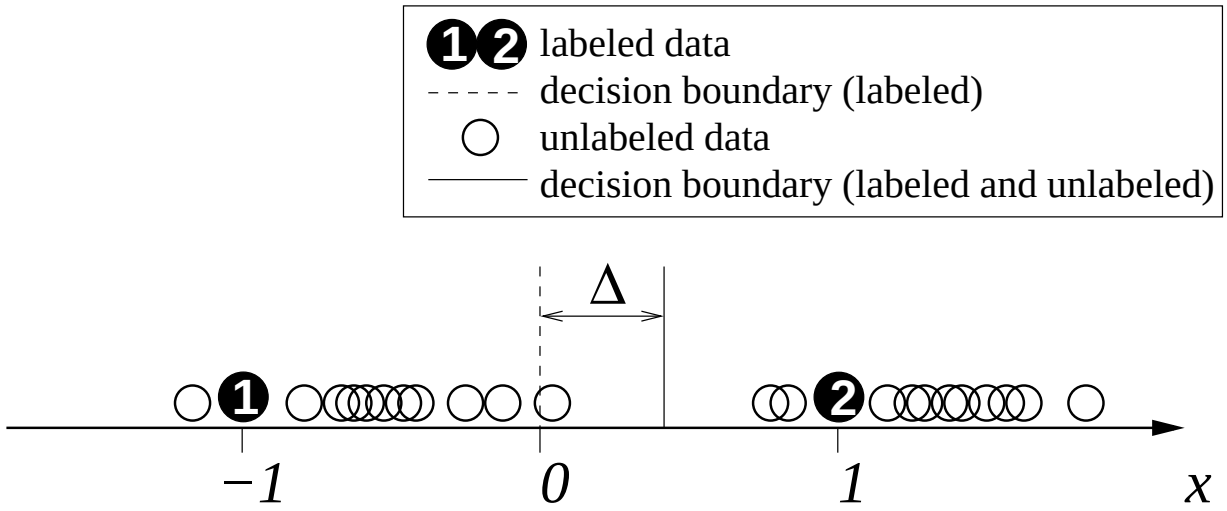
Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)
- labeled data (=supervised experiences) $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- unlabeled data (=unsupervised experiences) $X_u = \{x_{l+1:n}\}$
- usually $l \ll n$
- goal is to learn a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$

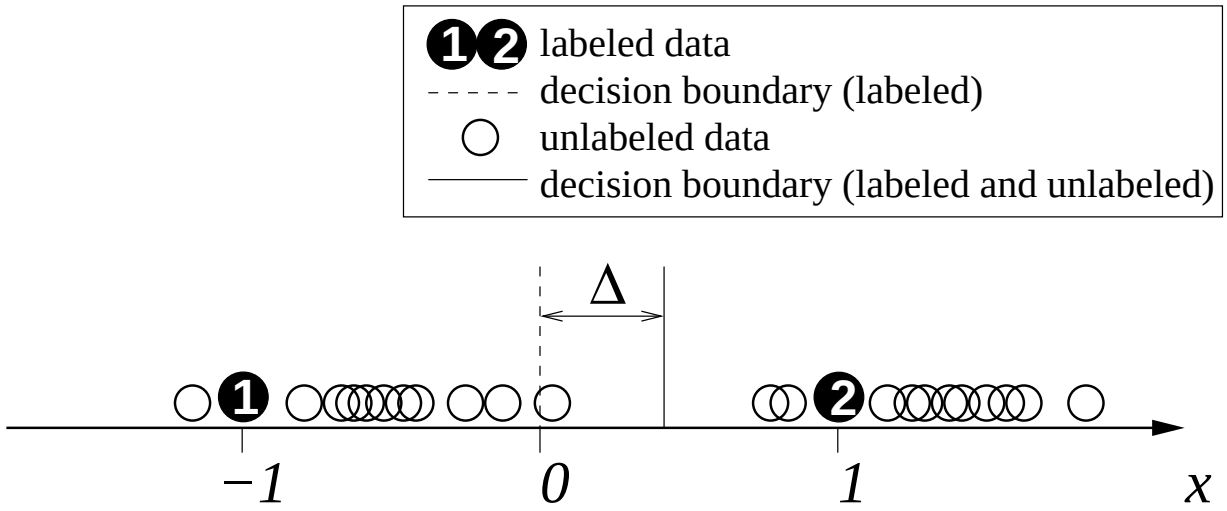
Classification

- input instance x (e.g., size)
- label y (two choices, e.g., raccoon vs. badger)
- labeled data (=supervised experiences) $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- unlabeled data (=unsupervised experiences) $X_u = \{x_{l+1:n}\}$
- usually $l \ll n$
- goal is to learn a classifier $f : \mathcal{X} \mapsto \mathcal{Y}$
  - from (X_l, Y_l) alone: supervised learning
  - from (X_l, Y_l) and X_u: semi-supervised learning

The intuition

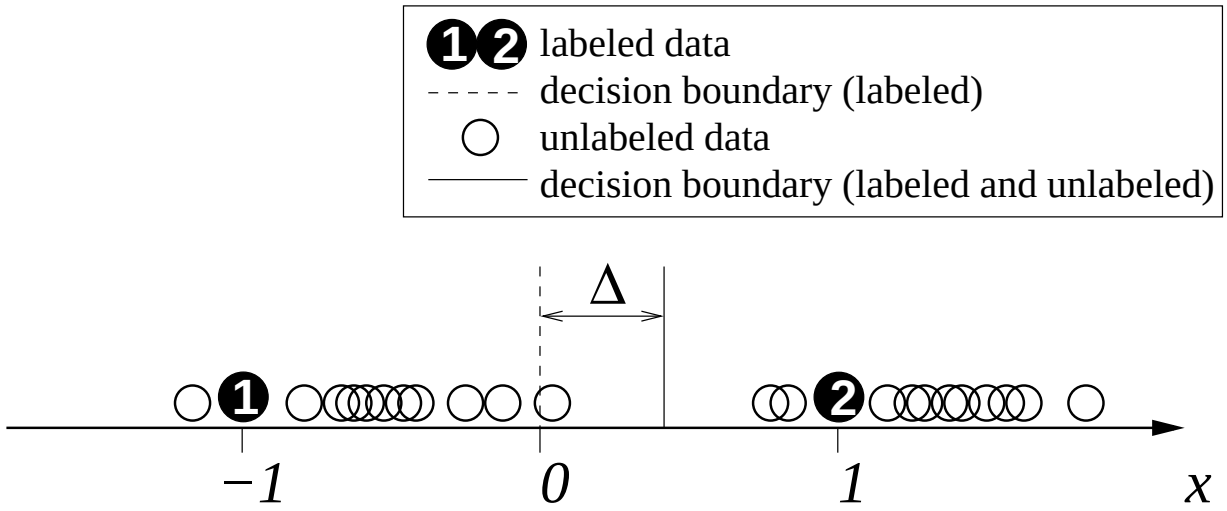


The intuition



- Assumption: one cluster (e.g., Gaussian) per class

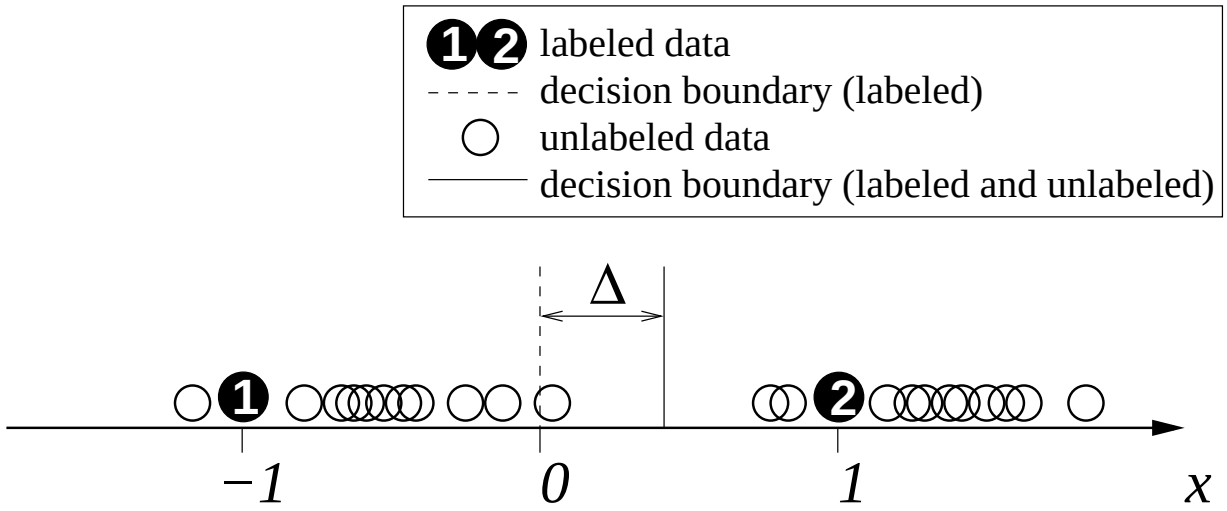
The intuition



- Assumption: one cluster (e.g., Gaussian) per class
- Decision boundaries differ in supervised vs. semi-supervised learning (more formally later...)

[Castelli & Cover 96; Ratsaby & Venkatesh 95; Nigam et al. 00]

The intuition



- Assumption: one cluster (e.g., Gaussian) per class
- Decision boundaries differ in supervised vs. semi-supervised learning (more formally later...)

  [Castelli & Cover 96; Ratsaby & Venkatesh 95; Nigam et al. 00]
- Do humans learn from both labeled and unlabeled data?

Human learning: a behavioral experiment

The plan

Detect human decision boundaries for:

- labeled only vs. labeled and unlabeled data
- same labeled data, different unlabeled data

Participants and materials

- 22 University of Wisconsin students

Participants and materials

- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)

Participants and materials

- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)
- Stimuli parameterized in 1D

Participants and materials

- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)
- Stimuli parameterized in 1D
- Cover story: microscopic pollen from two flowers

Participants and materials

- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)
- Stimuli parameterized in 1D
- Cover story: microscopic pollen from two flowers
- Press B or N to classify

Participants and materials

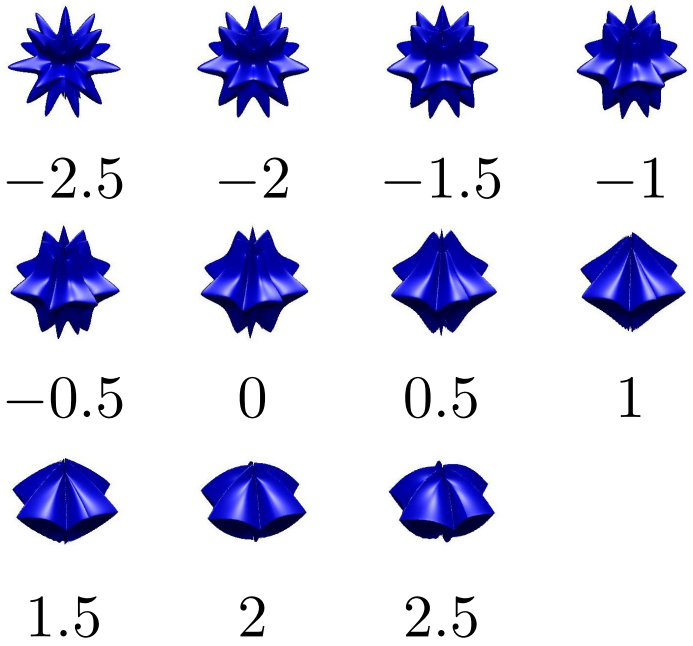
- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)
- Stimuli parameterized in 1D
- Cover story: microscopic pollen from two flowers
- Press B or N to classify
- Labels: audio feedback

Participants and materials

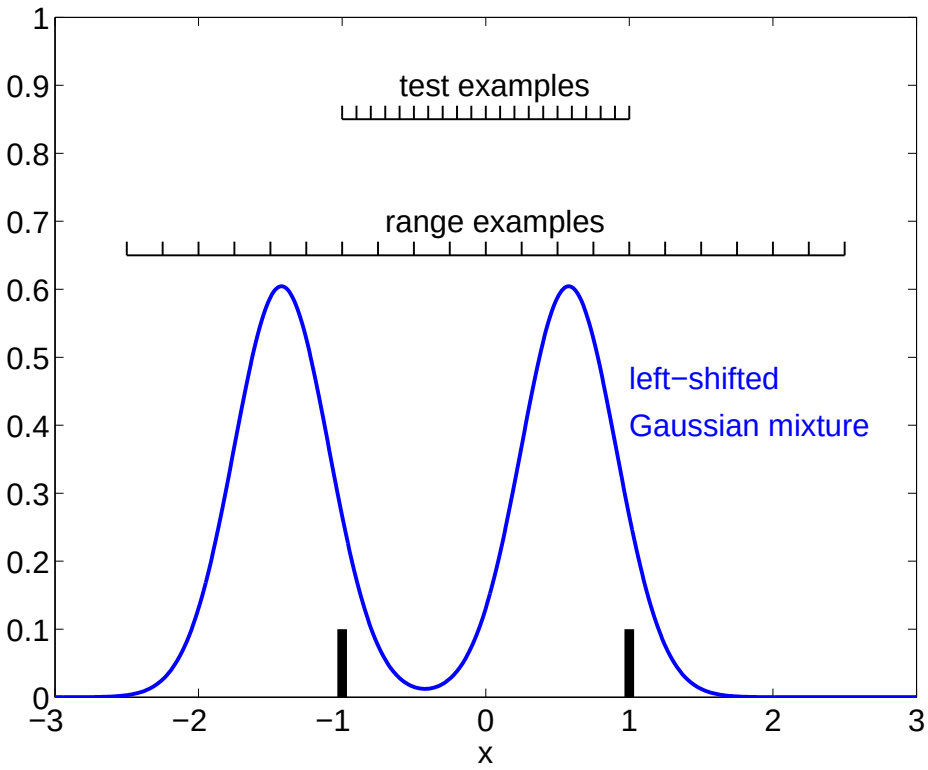
- 22 University of Wisconsin students
- Visual stimuli displayed one at a time (nothing stays on screen)
- Stimuli parameterized in 1D
- Cover story: microscopic pollen from two flowers
- Press B or N to classify
- Labels: audio feedback
- No audio feedback for unlabeled data

Visual stimuli

Stimuli parameterized by a continuous variable x. Some examples:



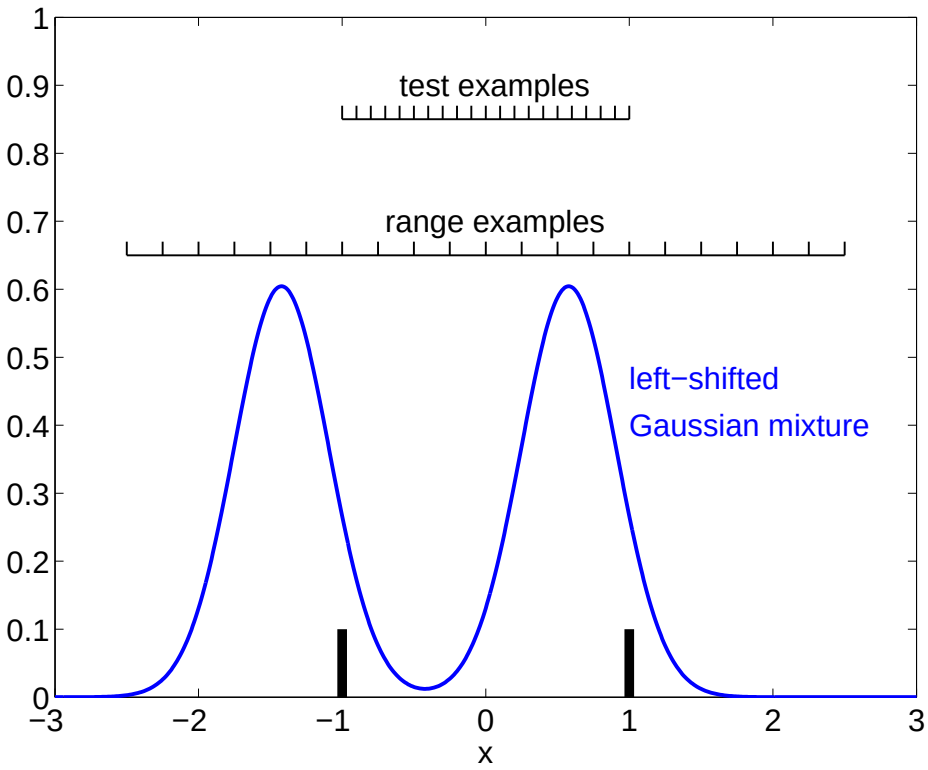
Experiment procedure



stimuli

1. (labeled) 10 $(x = 1, B)$, 10 $(x = -1, N)$. The only labeled block.

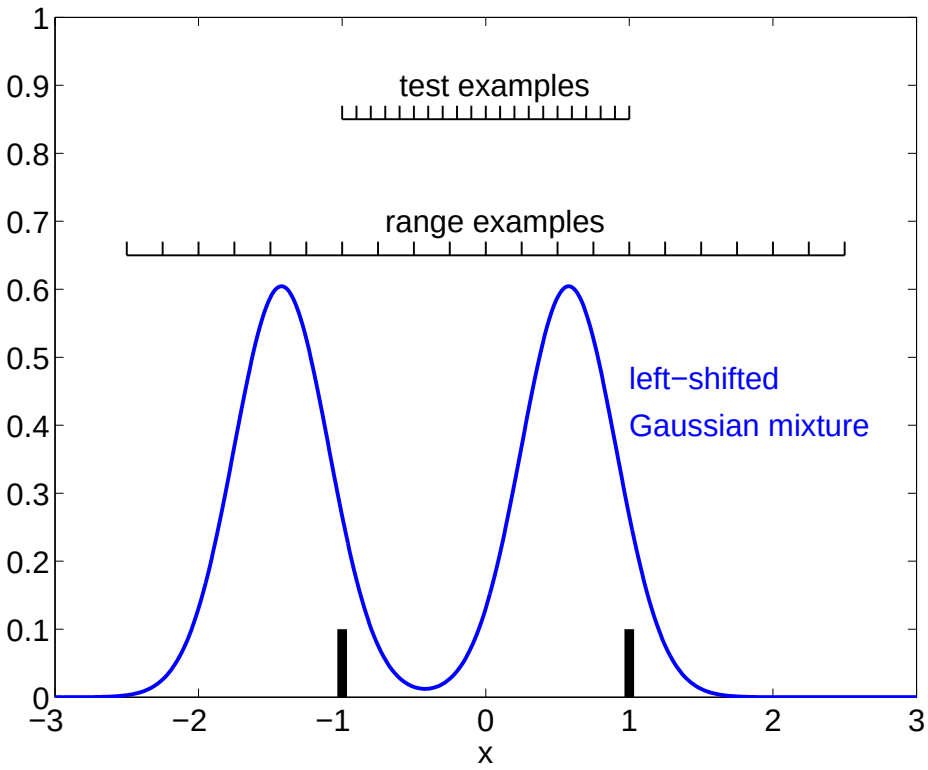
Experiment procedure



stimuli

1. (labeled) 10 $(x = 1, B)$, 10 $(x = -1, N)$. The only labeled block.
2. (test-1) $x = -1, -0.9, \ldots, 0.9, 1$

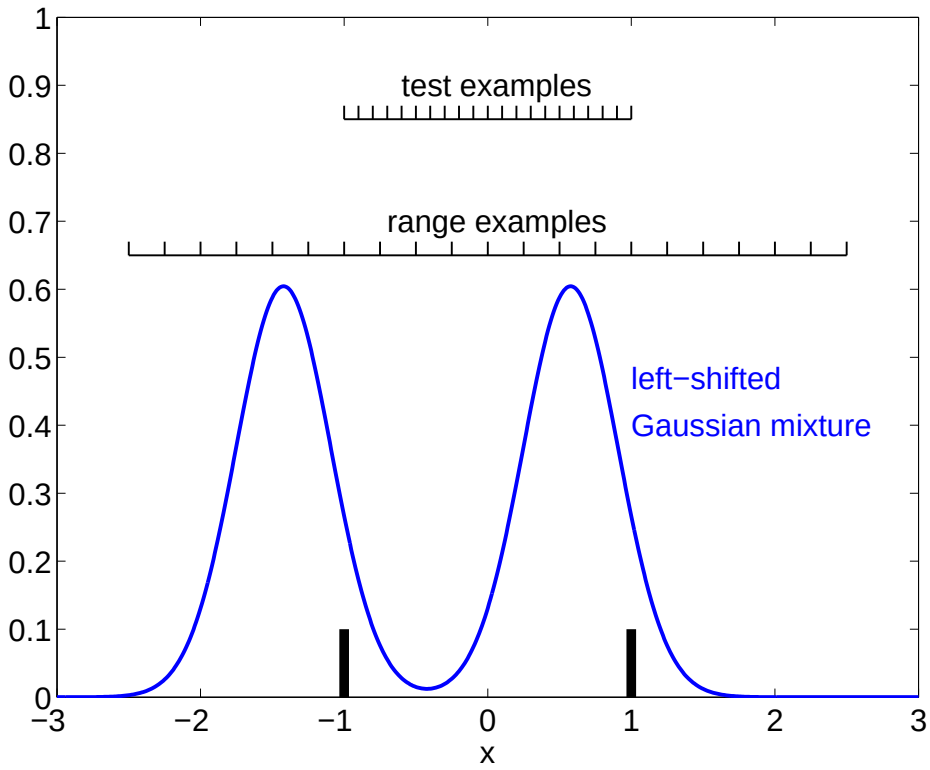
Experiment procedure



stimuli

1. (labeled) 10 $(x = 1, B)$, 10 $(x = -1, N)$. The only labeled block.
2. (test-1) $x = -1, -0.9, \ldots, 0.9, 1$
3. (unlabeled-1) 230 stimuli $\sim$ offset 2 Gaussian, left- or right-shifted. 21 range stimuli evenly in $[-2.5, 2.5]$

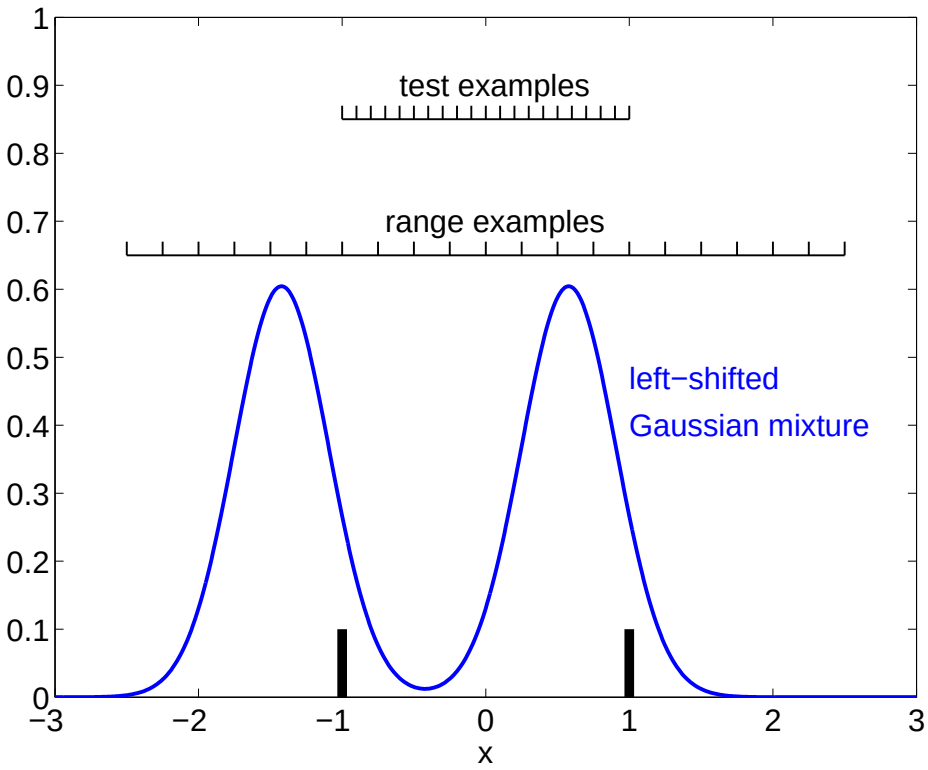
Experiment procedure



stimuli

1. (labeled) 10 $(x = 1, B)$, 10 $(x = -1, N)$. The only labeled block.
2. (test-1) $x = -1, -0.9, \ldots, 0.9, 1$
3. (unlabeled-1) 230 stimuli $\sim$ offset 2 Gaussian, left- or right-shifted. 21 range stimuli evenly in $[-2.5, 2.5]$
4. (unlabeled-2) similar to block 3
5. (unlabeled-3) similar to block 3

Experiment procedure



stimuli

1. (labeled) 10 $(x = 1, B)$, 10 $(x = -1, N)$. The only labeled block.
2. (test-1) $x = -1, -0.9, \ldots, 0.9, 1$
3. (unlabeled-1) 230 stimuli $\sim$ offset 2 Gaussian, left- or right-shifted. 21 range stimuli evenly in $[-2.5, 2.5]$
4. (unlabeled-2) similar to block 3
5. (unlabeled-3) similar to block 3
6. (test-2) $x = -1, -0.9, \ldots, 0.9, 1$

Experiment procedure

- Half L-subjects, half R-subjects

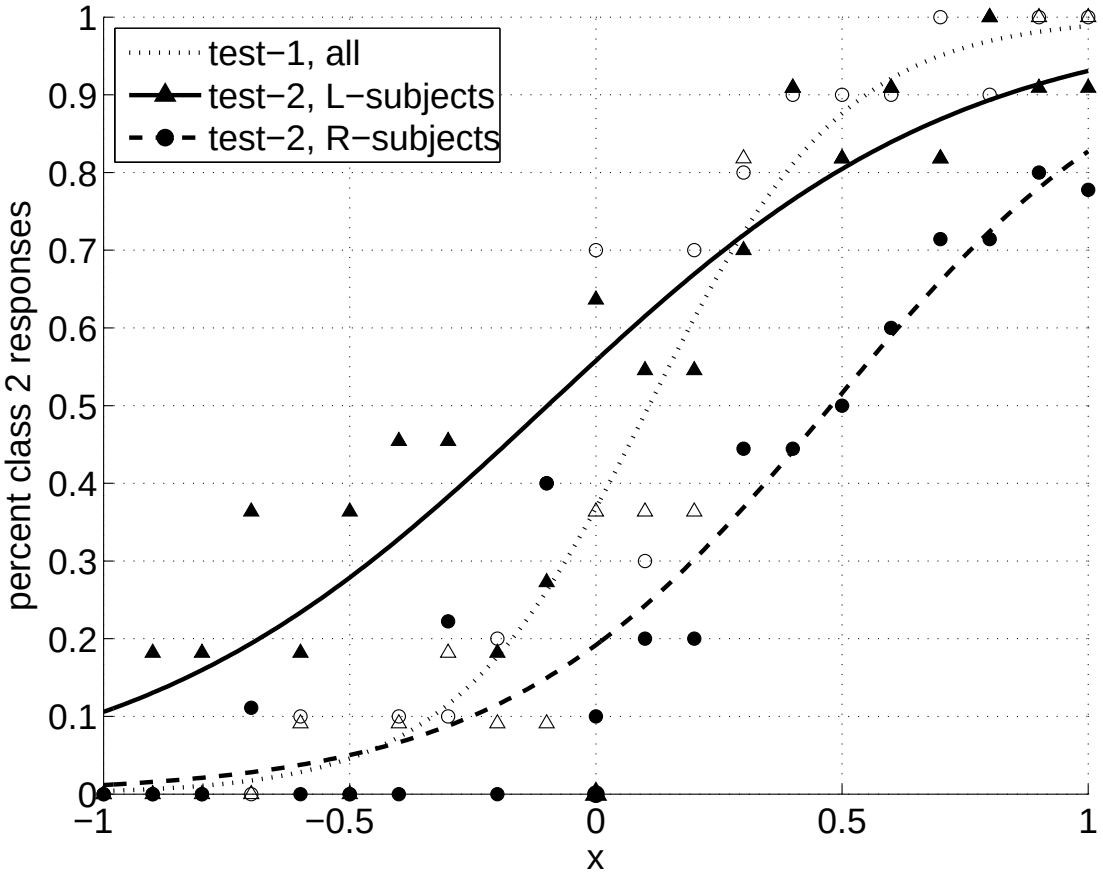
Experiment procedure

- Half L-subjects, half R-subjects
- Order within each block randomized

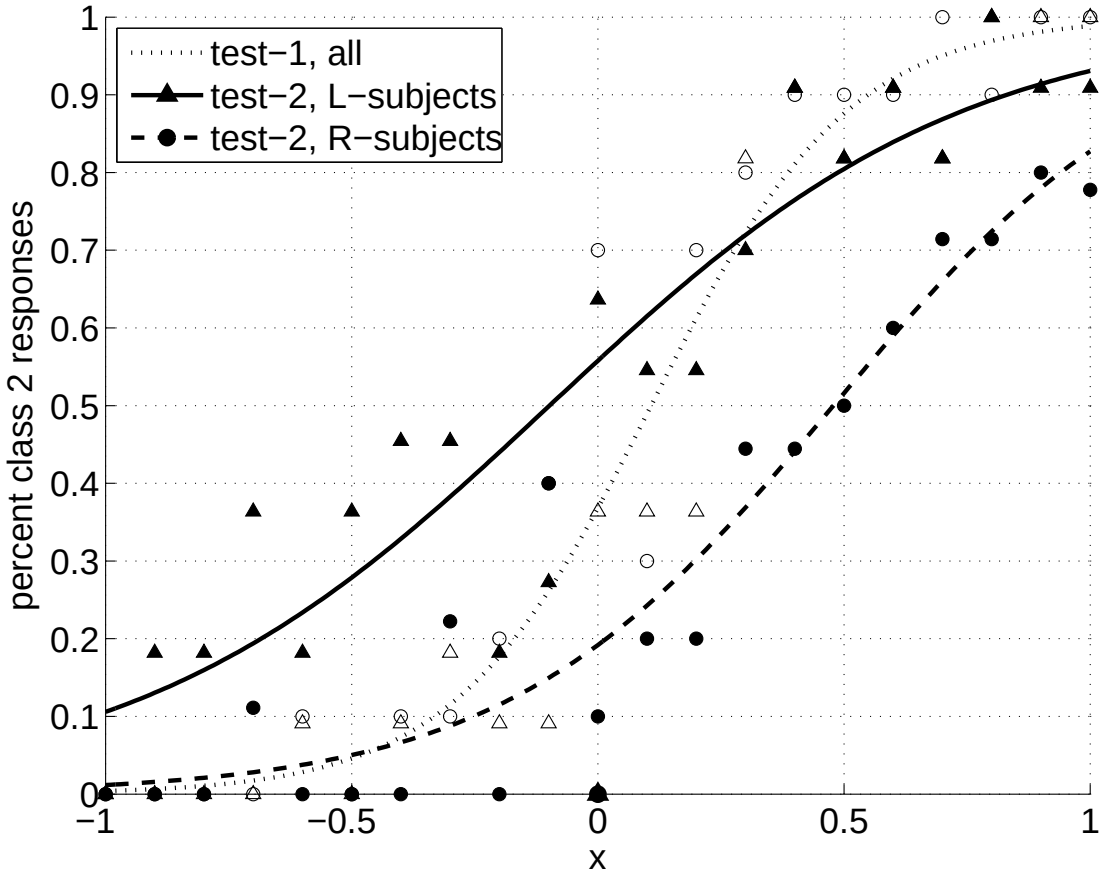
Experiment procedure

- Half L-subjects, half R-subjects
- Order within each block randomized
- Record their decisions and response times

Observation 1: Unlabeled data affects decision boundary



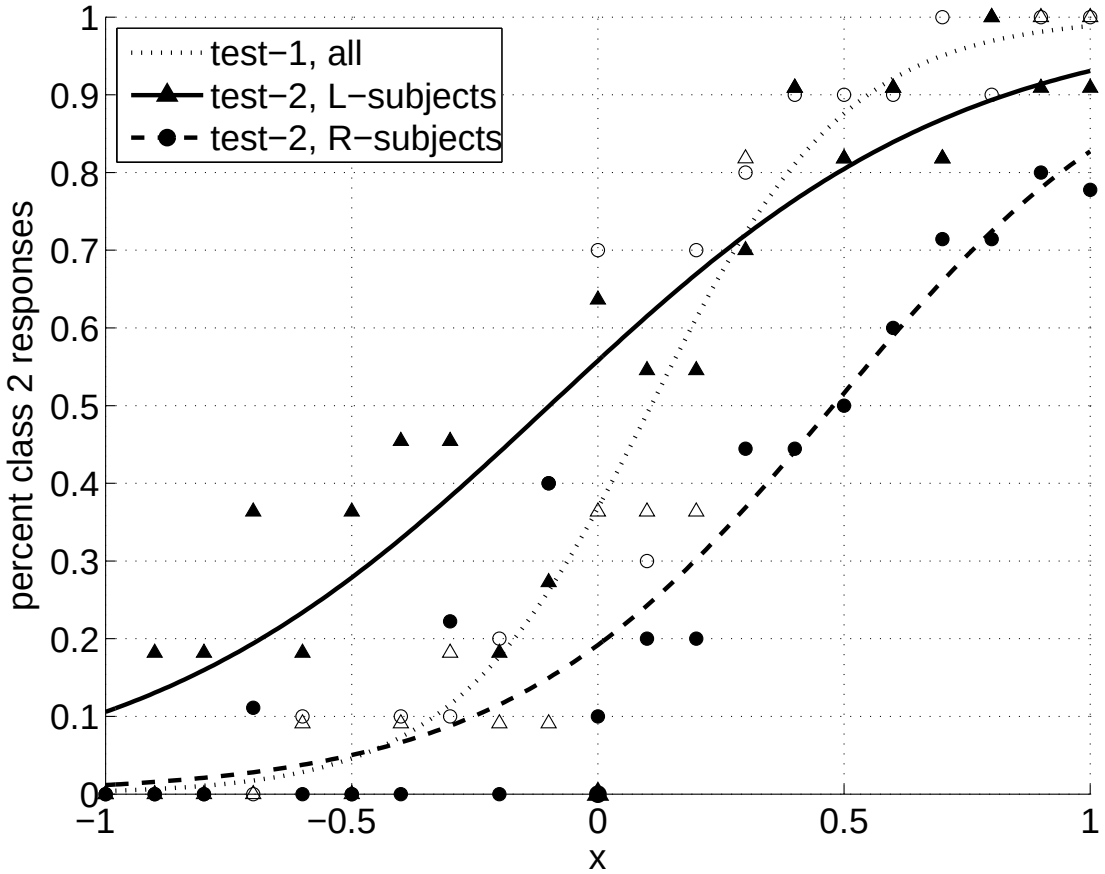
Observation 1: Unlabeled data affects decision boundary



Decision boundary:

- after labeled data (test-1): $x = 0.11$

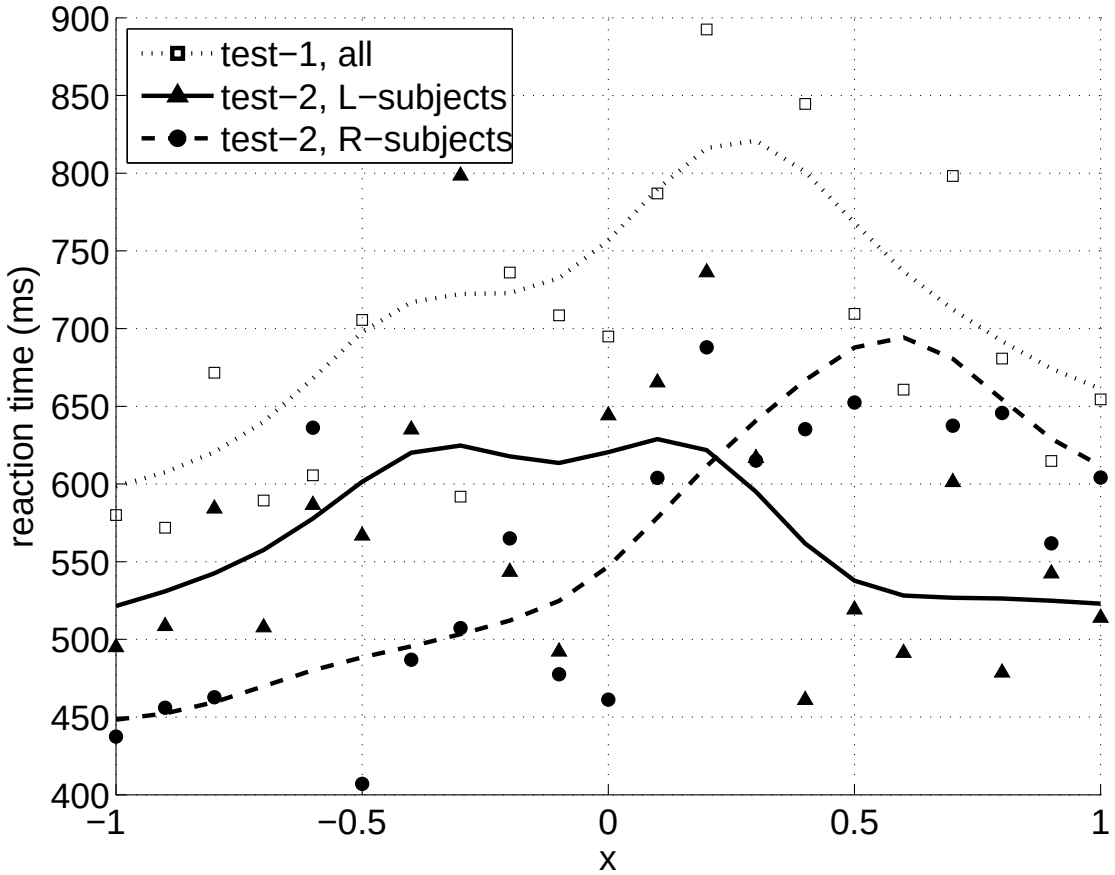
Observation 1: Unlabeled data affects decision boundary



Decision boundary:

- after labeled data (test-1): $x = 0.11$
- after labeled and unlabeled data (test-2):
  L-subjects $x = -0.10$, R-subjects $x = 0.48$

Observation 2: Reaction time reflects boundary shift



- Longer reaction time $\longrightarrow$ closer to decision boundary
- Test-2 overall faster: familiarity with experiment
- L-, R-reaction time further support decision boundary shift

Outline

1 A Human Learning Experiment

2 A Machine Learning Explanation

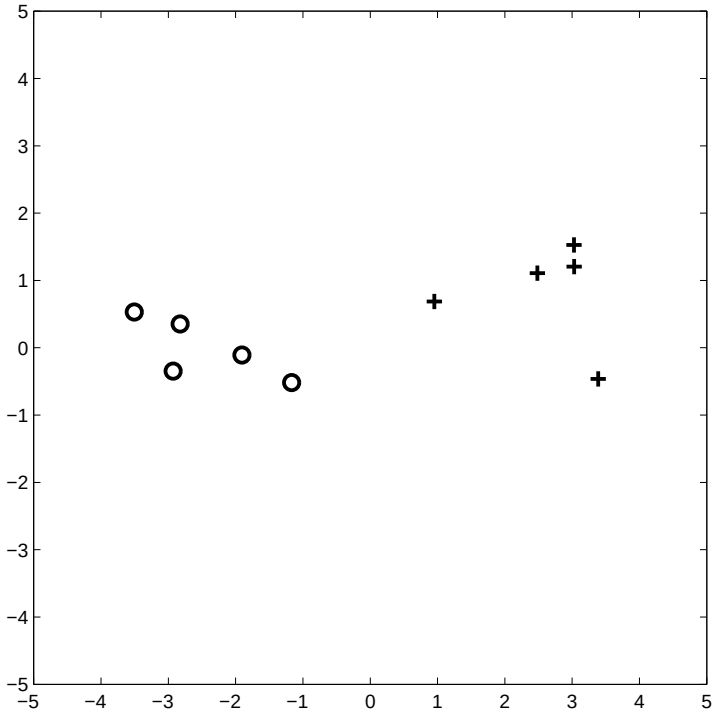
3 Opportunities for Further Research

A Machine Learning Explanation

A machine learning model (Gaussian mixture model) partially explains the human behavior.

A (slightly more interesting) example in 2D

Labeled data (X_l, Y_l):



Assuming each class has a 2D Gaussian distribution, what is the decision boundary?

A Gaussian mixture model with two components

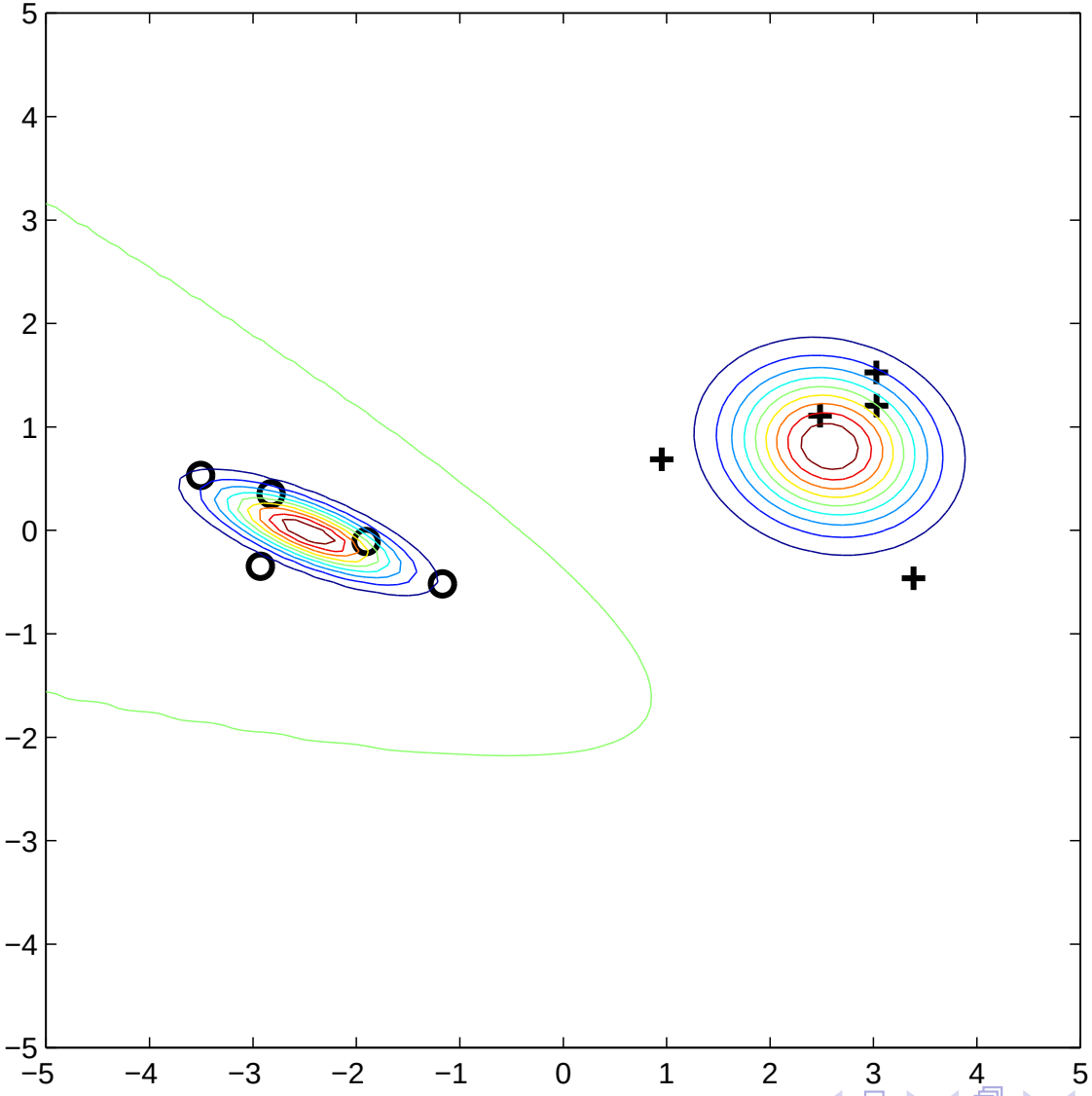
Parameters: $\theta = \{w_1, w_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$

$$\begin{aligned} p(x, y|\theta) &= p(y|\theta)p(x|y, \theta) \\ &= w_y \mathcal{N}(x; \mu_y, \Sigma_y) \end{aligned}$$

Classification: $p(y|x, \theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)}$

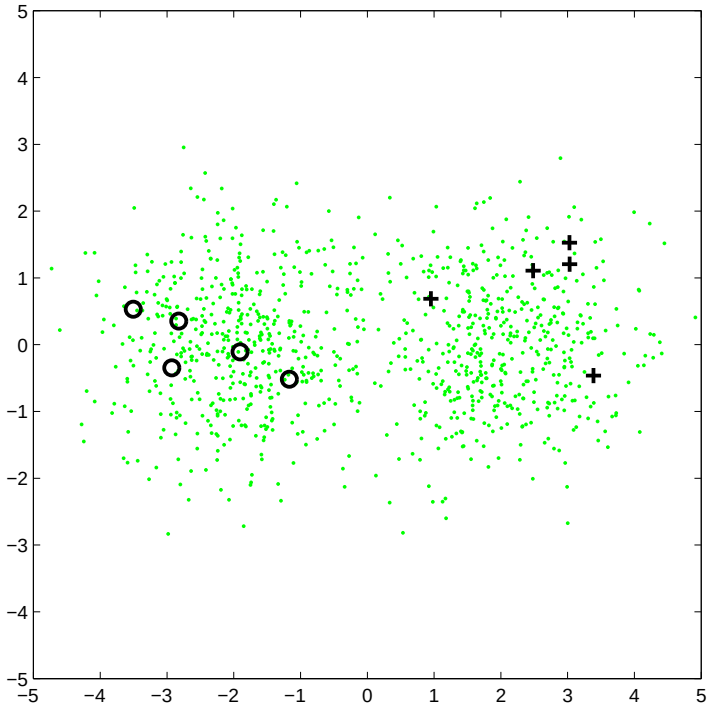
$\hat{\theta}^{MLE}$ on labeled data

Supervised decision boundary:



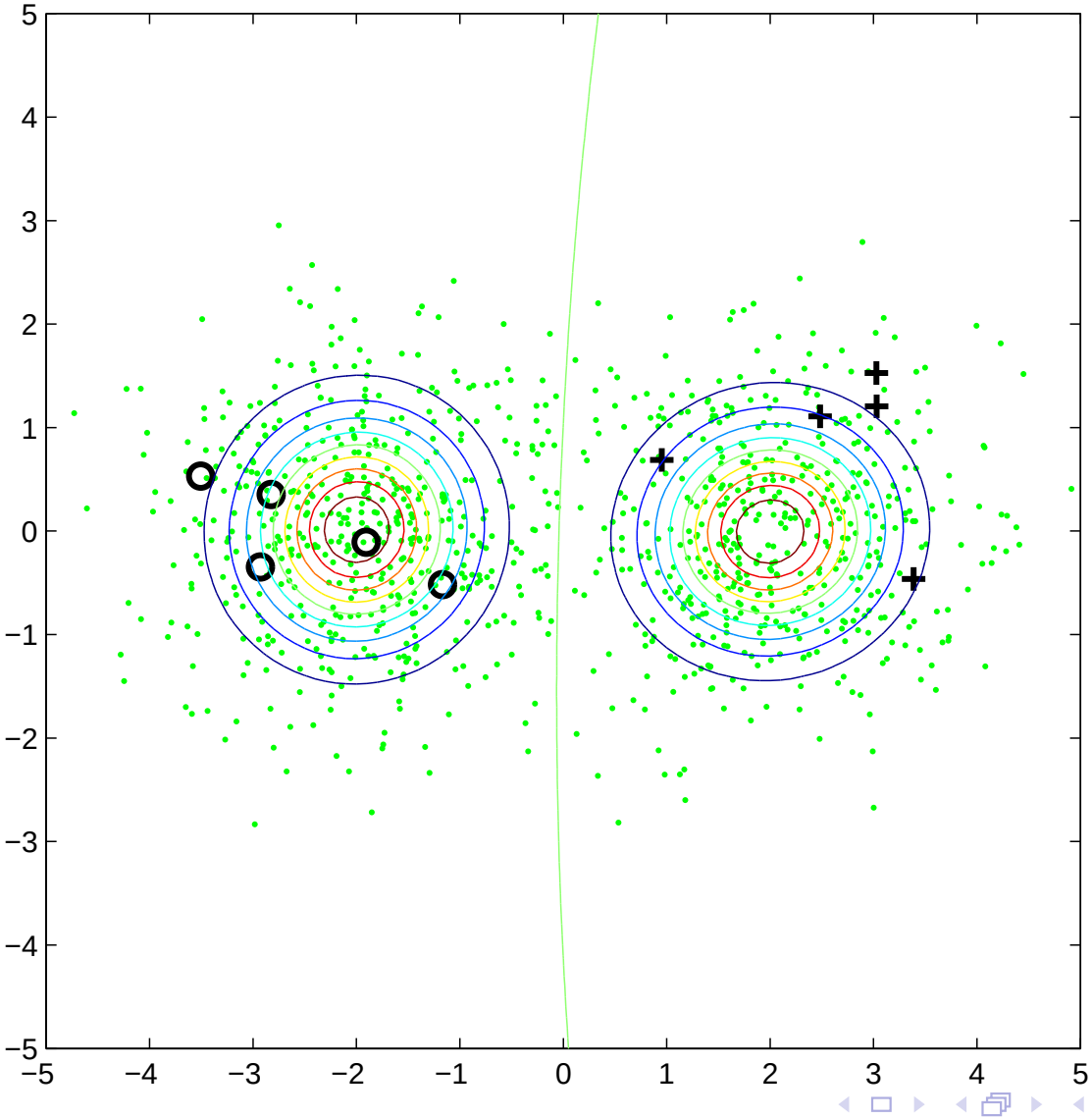
Semi-supervised learning

Add unlabeled data (green dots):



$\hat{\theta}^{MLE}$ on labeled and unlabeled data

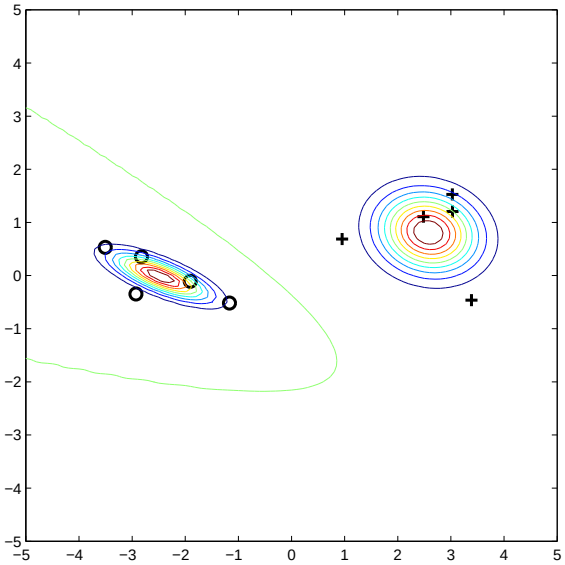
Semi-supervised decision boundary:



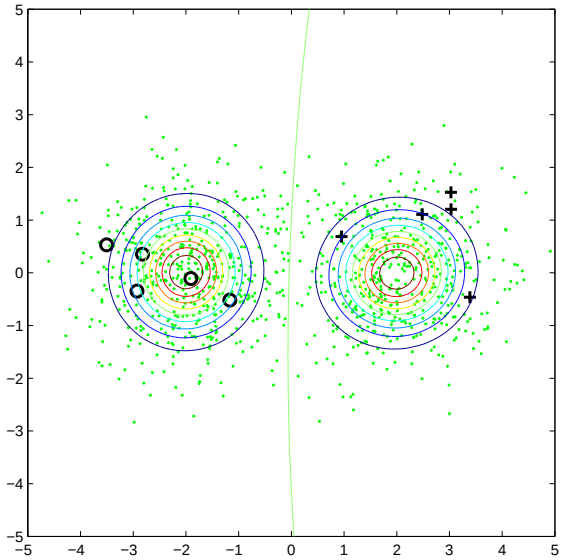
Supervised vs. semi-supervised

They are different because they maximize different objectives.

$p(X_l, Y_l|\theta)$



$p(X_l, Y_l, X_u|\theta)$



The objectives

- labeled data only
  - $\log p(X_l, Y_l|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i, \theta)$
  - MLE for θ trivial (sample mean, sample covariance)
- labeled and unlabeled data

$$\log p(X_l, Y_l, X_u|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i, \theta) + \sum_{i=l+1}^{l+u} \log \left(\sum_{y=1}^{2} p(y|\theta)p(x_i|y, \theta) \right)$$

  - MLE harder (hidden variables)
  - The Expectation-Maximization (EM) algorithm is one method to find a local optimum.

The EM algorithm for Gaussian mixture models

1. Start from MLE $\theta = \{w, \mu, \Sigma\}_{1:2}$ on (X_l, Y_l)
2. The E-step: compute the expected label $p(y|x,\theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)}$ for all $x \in X_u$
   - label $p(y=1|x,\theta)$-fraction of x with class 1
   - label $p(y=2|x,\theta)$-fraction of x with class 2
3. The M-step: update MLE θ with (now labeled) X_u
   - w_c=proportion of class c
   - μ_c=sample mean of class c
   - Σ_c=sample cov of class c

The EM algorithm for Gaussian mixture models

1. Start from MLE $\theta = \{w, \mu, \Sigma\}_{1:2}$ on (X_l, Y_l)
2. The E-step: compute the expected label $p(y|x,\theta) = \frac{p(x,y|\theta)}{\sum_{y'} p(x,y'|\theta)}$ for all $x \in X_u$
   - label $p(y=1|x,\theta)$-fraction of x with class 1
   - label $p(y=2|x,\theta)$-fraction of x with class 2
3. The M-step: update MLE θ with (now labeled) X_u
   - w_c=proportion of class c
   - μ_c=sample mean of class c
   - Σ_c=sample cov of class c

Repeat E-step and M-step until convergence.

GMM for the human experiment

Prior: $w_c \sim \mathrm{U}[0,1], \mu_c \sim \mathrm{N}(0,\infty), \sigma_c^2 \sim \mathrm{Inv}{-}\chi^2(\nu, s^2), c = 1, 2$

Objective ($\lambda \leq 1$ weight on unlabeled example):

$$\log p(\theta) + \sum_{i=1}^{l} \log p(x_i, y_i|\theta) + \lambda \sum_{i=l+1}^{n} \log p(x_i|\theta)$$

E-step

$$q_i(c) \propto w_c \mathrm{N}(x_i; \mu_c, \sigma_c^2), \quad i = l+1, \ldots, n; c = 1, 2$$

M-step

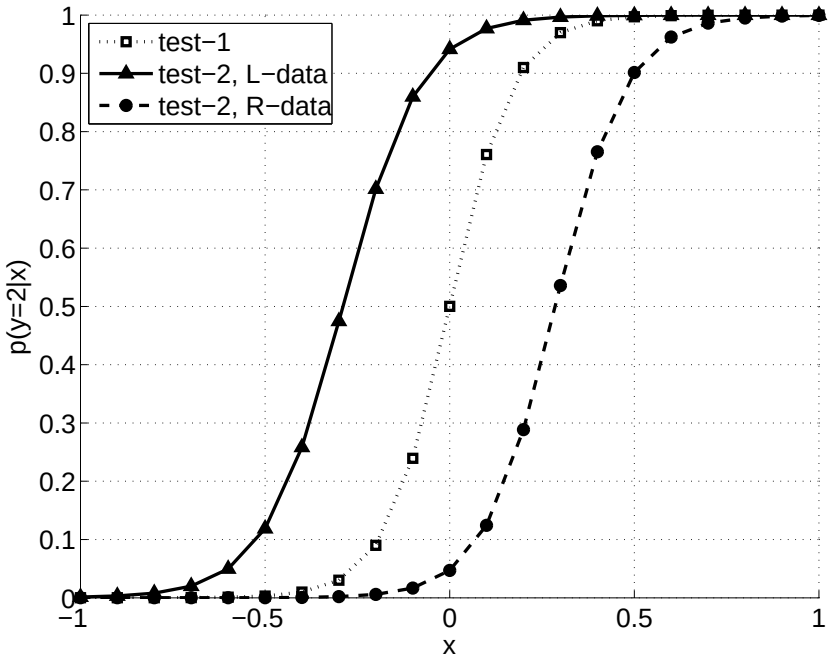
$$\mu_c = \frac{\sum_{i=1}^{l} \delta(y_i, c) x_i + \lambda \sum_{i=l+1}^{n} q_i(c) x_i}{\sum_{i=1}^{l} \delta(y_i, c) + \lambda \sum_{i=l+1}^{n} q_i(c)}$$

$$\sigma_c^2 = \frac{\nu s^2 + \sum_{i=1}^{l} \delta(y_i, c) e_{ic} + \lambda \sum_{i=l+1}^{n} q_i(c) e_{ic}}{\nu + 2 + \sum_{i=1}^{l} \delta(y_i, c) + \lambda \sum_{i=l+1}^{n} q_i(c)}$$

$$w_c = \frac{\sum_{i=1}^{l} \delta(y_i, c) + \lambda \sum_{i=l+1}^{n} q_i(c)}{l + \lambda(n-l)}$$

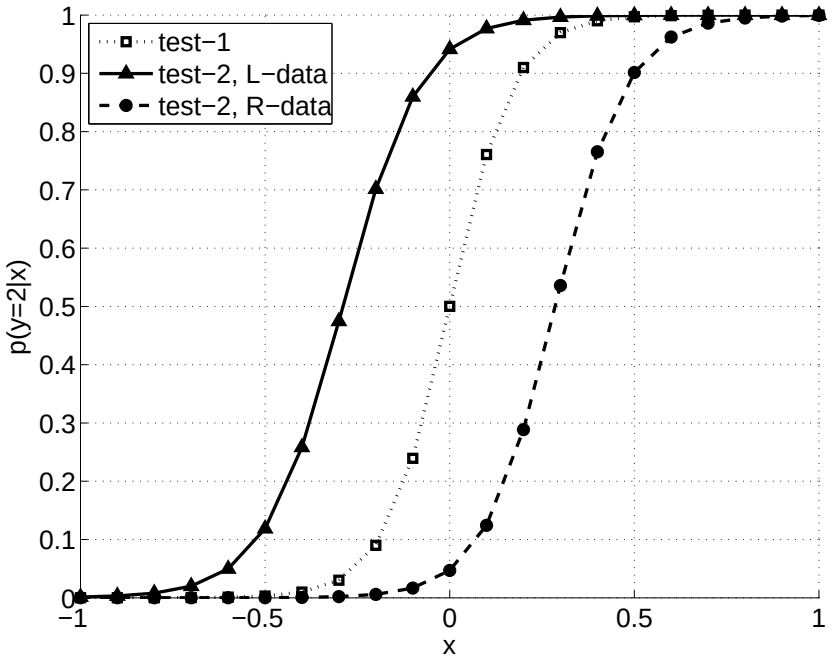
Model fitting result (1)

Comparing supervised vs. semi-supervised $\hat{\theta}^{MLE}$
GMM predicts decision boundary shift:



Model fitting result (1)

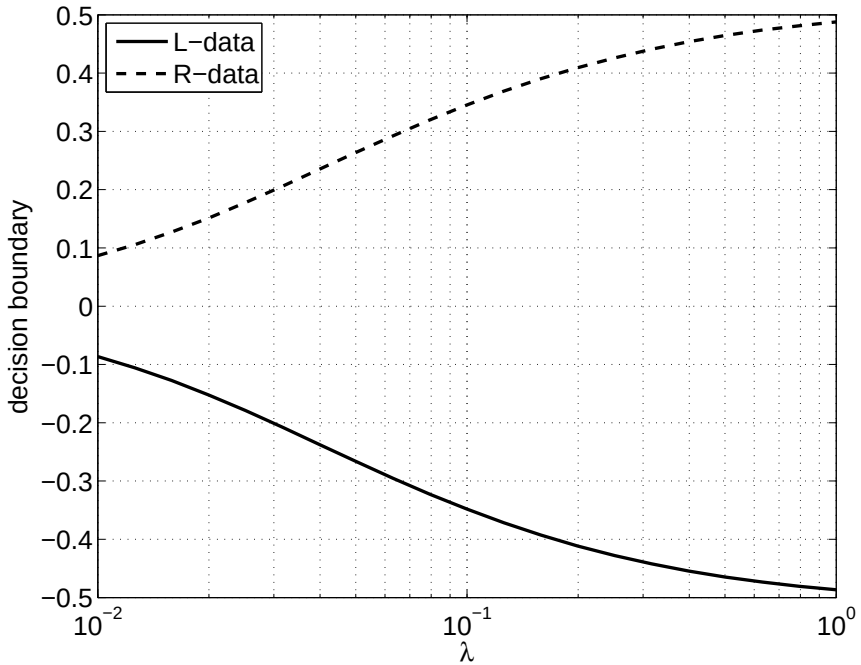
Comparing supervised vs. semi-supervised $\hat{\theta}^{MLE}$
GMM predicts decision boundary shift:



(But notice the slope is much steeper...)

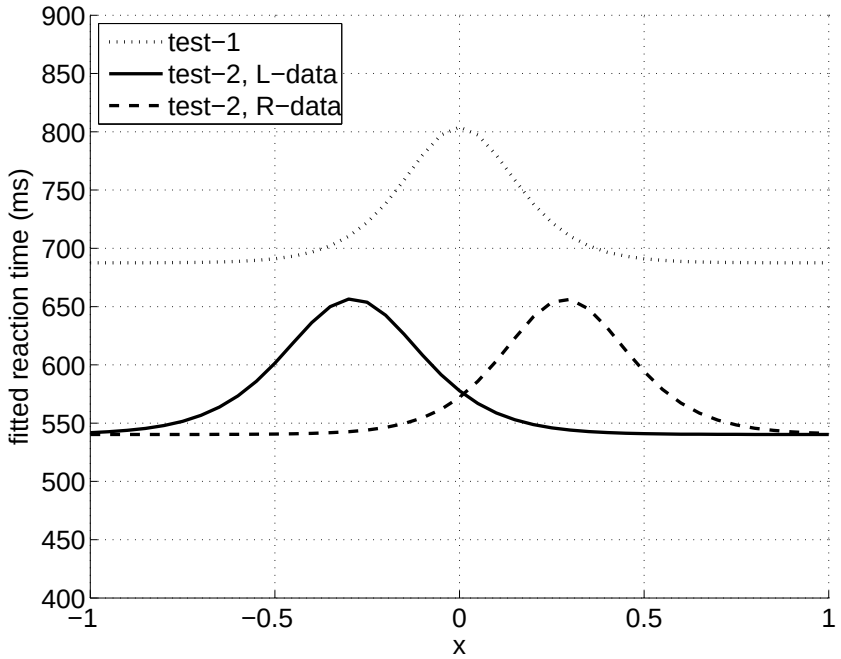
Model fitting result (2)

Unlabeled data seem to worth less than labeled data ($\lambda = 0.06$)



Model fitting result (3)

GMM explains reaction time: $t = aH(x) + b$



$H(x)$: entropy of $p(y|x, \theta)$

So far so good

GMM fits human behavior data.

So we calculate Gaussian mixture models in our heads?

So far so good

GMM fits human behavior data.

So we calculate Gaussian mixture models in our heads?

- Maybe so.
- There are rich alternatives in machine learning literature.
- They should be tested on humans.

Outline

1 A Human Learning Experiment

2 A Machine Learning Explanation

3 Opportunities for Further Research

Co-Training in humans?

Feature split: Each instance is represented by two subsets of features $x = [x^{(1)}; x^{(2)}]$

- $x^{(1)}$ = visual shape
- $x^{(2)}$ = audio sound

Co-training idea:

- Train a visual classifier and an audio classifier
- The two classifiers teach each other

Co-Training algorithm

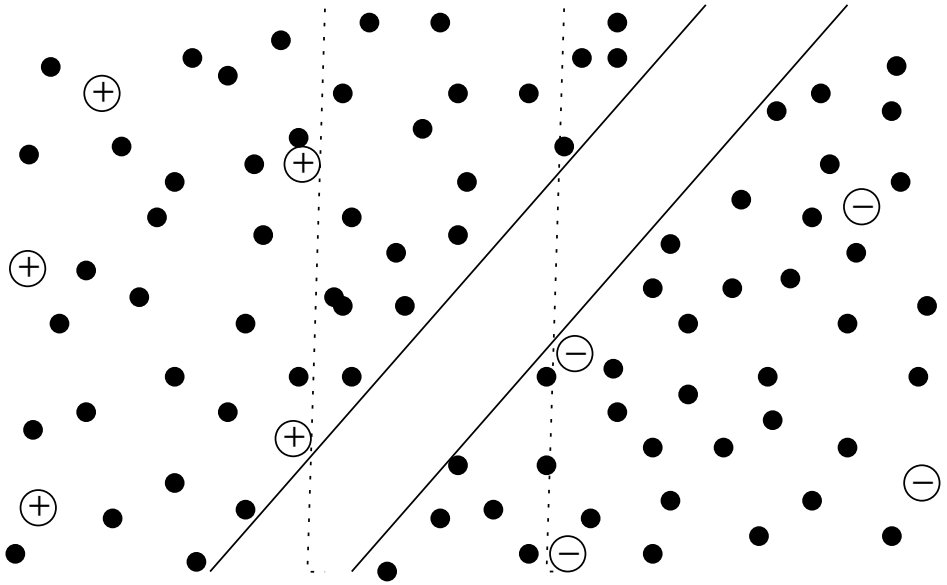
1. Train two classifiers: $f^{(1)}$ from $(X_l^{(1)}, Y_l)$, $f^{(2)}$ from $(X_l^{(2)}, Y_l)$.
2. Classify X_u with $f^{(1)}$ and $f^{(2)}$ separately.
3. Add $f^{(1)}$'s k-most-confident $(x, f^{(1)}(x))$ to $f^{(2)}$'s labeled data.
4. Add $f^{(2)}$'s k-most-confident $(x, f^{(2)}(x))$ to $f^{(1)}$'s labeled data.
5. Repeat.

Co-training assumptions:

- $x^{(1)}$ or $x^{(2)}$ alone is sufficient to train a good classifier
- $x^{(1)}$ and $x^{(2)}$ are conditionally independent given the class

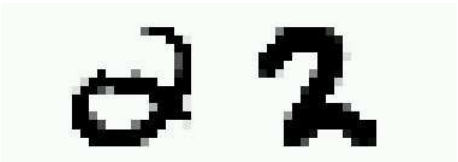
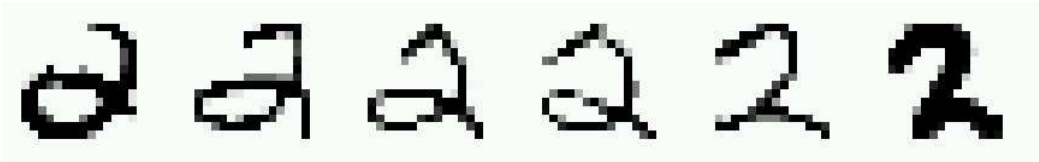
Large margin separation in humans?

Semi-supervised SVMs (S3VMs) maximize the "unlabeled data margin":



Graph-based learning in humans?

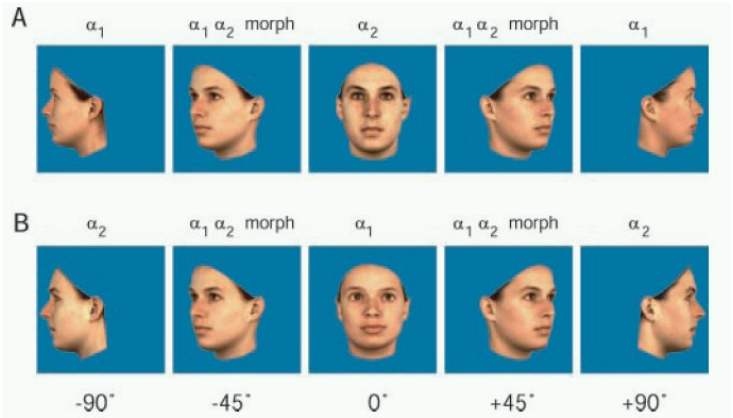
Handwritten digits recognition with pixel-wise Euclidean distance

	
not similar	'indirectly' similar with stepping stones

Graph-based learning in humans?

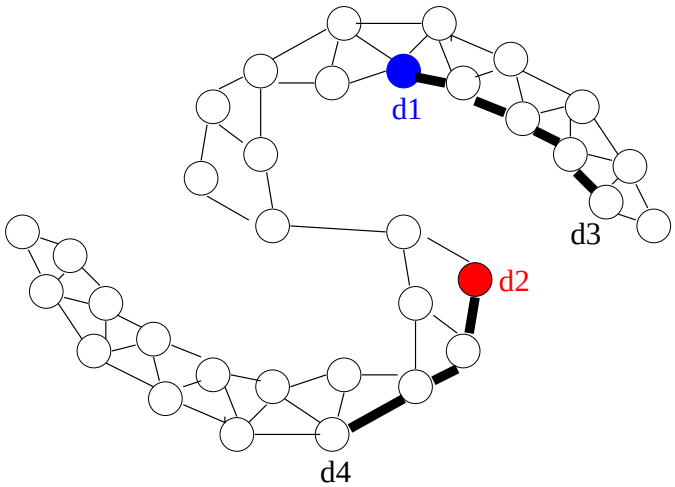
- A face from two angles are very different
- But we can easily associate them
- Unlabeled temporal image sequence might be the "glue"
- Artificial wrong sequences (person A's profile morphs to B's frontal) damage people's ability to match test profile and frontal images.

  [Wallis and Bülthoff 01]



The graph

- Nodes: $X_l \cup X_u$
- Edges: direct similarity between two instances (e.g., kNN graph)
- Want: implied similarity via all paths



Active learning in humans?

- What if one can actively query the label of any instance?
- Can they beat passive learning?
- Can they beat active machine learning algorithms?

Conclusions

- Humans and machines both perform semi-supervised learning.

Conclusions

- Humans and machines both perform semi-supervised learning.
- Other models (Co-training, graph-based, S3VMs, active learning, etc.) in humans should be explored.

Conclusions

- Humans and machines both perform semi-supervised learning.
- Other models (Co-training, graph-based, S3VMs, active learning, etc.) in humans should be explored.
- Machine learning + cognitive psychology = new discoveries.

Conclusions

- Humans and machines both perform semi-supervised learning.
- Other models (Co-training, graph-based, S3VMs, active learning, etc.) in humans should be explored.
- Machine learning + cognitive psychology = new discoveries.

Acknowledgments:
Timothy Rogers, Ruichen Qian, Chuck Kalish, Rob Nowak