Humans Perform Semi-Supervised Classification Too[*]

Xiaojin Zhu†, Timothy Rogers‡, Ruichen Qian† and Chuck Kalish‡
Department of Computer Sciences†
Department of Psychology‡
University of Wisconsin, Madison, WI 53706, USA
jerryzhu@cs.wisc.edu, ttrogers@wisc.edu, qian2@wisc.edu, cwkalish@wisc.edu

Abstract

We explore the connections between machine learning and human learning in one form of semi-supervised classification. 22 human subjects completed a novel 2-class categorization task in which they were first taught to categorize a single labeled example from each category, and subsequently were asked to categorize, without feedback, a large set of additional items. Stimuli were visually complex and unrecognizable shapes. The unlabeled examples were sampled from a bimodal distribution with modes appearing either to the left (left-shift condition) or right (right-shift condition) of the two labeled examples. Results showed that, although initial decision boundaries were near the middle of the two labeled examples, after exposure to the unlabeled examples, they shifted in different directions in the two groups. In this respect, the human behavior conformed well to the predictions of a Gaussian mixture model for semi-supervised learning. The human behavior differed from model predictions in other interesting respects, suggesting some fruitful avenues for future inquiry.

Introduction

Semi-supervised learning–the effort to develop classifiers that can capitalize on both labeled and unlabeled training data–has attracted considerable interest in the machine learning community. New semi-supervised methods have significantly improved machine learning in various applications, including text categorization, computer vision, and bioinformatics, see (Chapelle, Zien, & Schölkopf 2006; Zhu 2005) for recent reviews. Given these successes, we ask the fundamental question: *Do humans perform semi-supervised classification?* That is, do humans use "unlabeled data" in addition to "labeled data" to learn categories? If so, can we explain such behavior with mathematical models developed for semi-supervised machine learning? Answers to these questions may shed light on the cognitive process behind human learning, which may in turn lead to novel machine learning approaches (Mitchell 2006; Langley 2006).

Many people would agree that the first answer seems to be "yes". After all, a child learns with supervision from parents and teachers, as well as without supervision by silently observing the world around her. Despite a significant amount of research in psychology on supervised and unsupervised learning (e.g., (Love 2002) and the references therein), semi-supervised learning is not well studied. Although some prior research indirectly supports the above intuitions, e.g., (Graf Estes *et al.* 2006; Tenenbaum & Xu 2000), we are aware of just one previous study that directly investigates semi-supervised learning in humans. Specifically, Stromsten (2002, Chapter 3) used drawings of artificial fish to show that human categorization behavior can be influenced by the presence of unlabeled examples. Though certainly suggestive, this experiment had two limitations. First, Stromsten used a single positive labeled example and no negative labeled examples, making it a one-class setting similar to novelty detection or quantile estimation. Recent semi-supervised machine learning research has, in contrast, focused primarily on two-class classification with positive and negative examples. Second, since Stromsten used stimuli[1] that correspond to a familiar real-world concept (i.e. fish), it is difficult to know whether his results reflect prior knowledge about the category, or new learning obtained over the course of the experiment.

The current work describes a new study that clearly demonstrates one form of semi-supervised classification in humans. In a two-class learning paradigm, we show that the learned decision boundary is determined by both labeled and unlabeled data. In our experiment, participants view a series of visually complex shapes, and must guess to which of 2 categories each stimulus belongs. "Labeled" examples consist of trials for which the participant gets accurate feedback, and "unlabeled" examples consist of trials without feedback. Given the same labeled data but different unlabeled data, people form different decision boundaries. To account for this behavior, we propose that semi-supervised category learning in humans can be described with a generative mixture model, a traditional machine learning method (Nigam *et al.* 2000). Our paper thus takes the first steps toward designing and interpreting human learning experiments based on semi-supervised machine learning models.

[*] We thank Tom Mitchell, Joshua Tenenbaum and Sean Stromsten for helpful discussions. Research supported in part by Wisconsin Alumni Research Foundation.

[1] We use *stimulus* and *example* interchangeably.

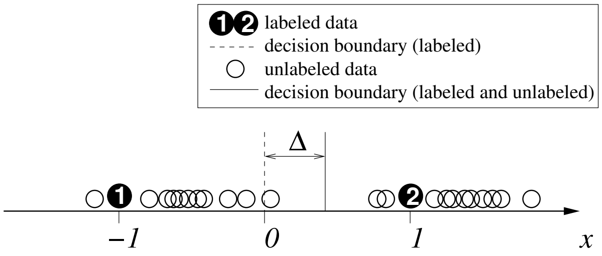

❶❷ labeled data
decision boundary (labeled)
○ unlabeled data
decision boundary (labeled and unlabeled)
Δ
❶
❷
−1
0
1
x

Figure 1: The additional knowledge of unlabeled data produces a better decision boundary.

The Semi-Supervised Learning Task

We start by introducing a classic classification task in semi-supervised machine learning. For simplicity, let us assume each example is represented by a one-dimensional feature $x \in \mathbb{R}$. There are two classes $y \in \{1, 2\}$. Consider the following two scenarios:

1. We are given only two labeled training examples $(x_1, y_1) = (-1, 1)$ and $(x_2, y_2) = (1, 2)$. The best estimate of the decision boundary is $x = 0$: everything to the left should be classified as $y = 1$, while others as $y = 2$.

2. In addition to the two labeled examples above, we are also given a large number of unlabeled examples $x_3, \ldots, x_n$. The correct class labels for these unlabeled examples are unknown. However we observe that they form two groups as in Figure 1. Under *the assumption* that examples in each class form a coherent group (e.g., follow a Gaussian distribution), our estimate of the decision boundary should be between the two groups instead (solid line in Figure 1).

Comparing the two scenarios in Figure 1, we expect to see a shift Δ in the decision boundaries. The amount of shift depends on the particular distributions of labeled and unlabeled data. Intuitively, the decision boundary estimated from only labeled data may be unreliable, since the number of labeled examples is small. One can show that if the "coherent group" assumption is correct, unlabeled data will lead to a better estimate of the decision boundary[2]. This is a well-studied semi-supervised machine learning method (Castelli & Cover 1996; Ratsaby & Venkatesh 1995), and has shown empirical successes (Nigam *et al.* 2000; Baluja 1998).

A Behavioral Experiment

Our study of human semi-supervised learning closely follows the above setting. We compare two scenarios: 1) the participant receives only labeled examples, which happen to be off the true class centers, versus 2) the participant also receives unlabeled examples sampled from the true class conditional feature distributions. Our goal is to determine whether the participant's category decision boundary shifts between the two scenarios. An appropriate shift would indicate that the participant's mental representations of the two categories take into account distributional information from the unlabeled data.

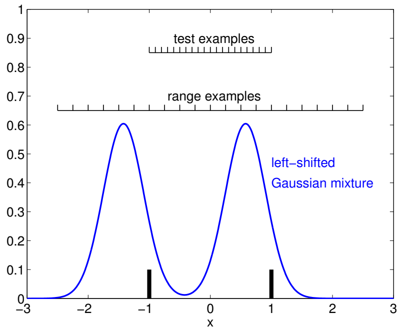


Figure 3: Data used in our behavioral experiment.

Participants and Materials

Participants were 22 students from University of Wisconsin, participating for partial course credit.

In our experiment, each example is a 3D shape displayed to the subject on a computer screen. To keep later analysis simple, the examples are parameterized by a single parameter x. A similar setting was described by Mozer, Jones, & Shettel (2006), who used circles of different sizes as the examples. Size, however, is a less than ideal parameter, first because people may bring relevant prior knowledge to the task (for instance, the knowledge that size varies continuously along an infinite range), and second because size is limited by what can be displayed on a computer screen. To avoid these difficulties, we generated novel, artificial 3D stimuli based on "supershapes" introduced in (Gielis 2003). The shapes change with x smoothly in several aspects simultaneously. Figure 2 shows a few shapes and their x values. In our experiment the examples are organized into six sequential blocks. We refer to Figure 3 in the following description.

Block 1 (labeled) consists of 2 labeled examples ($x = -1, y = 1$ and $x = 1, y = 2$) each appearing 10 times, with the total of 20 trials appearing in a different random order for each participant. The repetition of 2 items in this block ensures quick learning of the 2 distinct labeled examples.

Block 2 (test-1) consists of 21 evenly spaced unlabeled examples $x = -1, -0.9, -0.8, \ldots, 1$, appearing in a different random order for each participant. We use them to test the learned decision boundary after Block 1.

Block 3 (unlabeled-1) is one of the three unlabeled data blocks. We sample 230 unlabeled examples from an equal mixture of two Gaussian distributions, representing the "true" concepts to be learned. Importantly, the means are shifted away from the labeled examples at $x = -1$ and $x = 1$: for 12 participants the two Gaussian distributions are shifted to the left, and for the other 10 participants they are shifted to the right, so that in both groups the labeled examples from Block 1 are not prototypical examples of each class. For the left-shifted mixture, we use

$$x \sim \frac{1}{2}\mathrm{N}(-1 - 1.28\sigma, \sigma^2) + \frac{1}{2}\mathrm{N}(1 - 1.28\sigma, \sigma^2), \qquad (1)$$

where $\mathrm{N}(\mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. We set the standard deviation $\sigma = 1/3$. We

[2] Some cautionary notes are provided in (Cozman, Cohen, & Cirelo 2003) when the assumption is wrong.

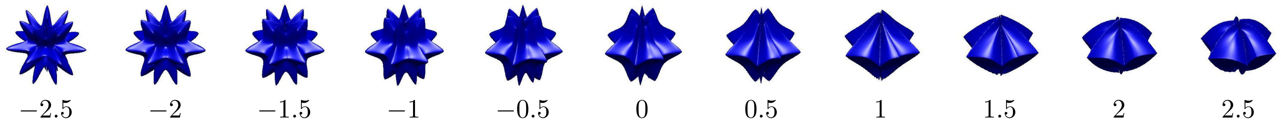


Figure 2: Our experiment uses a large number of 3D shape visual stimuli, parameterized by a continuous scalar x. A few examples are shown above with the corresponding x values.

choose the shift $\Delta = 1.28\sigma$ because 80% of the area under the Normal curve is within 1.28 standard deviation of the mean, which puts the labeled examples off the center, but not so extreme as to be outliers. Similarly, for the right-shifted mixture, we use $x \sim \frac{1}{2}\mathrm{N}(-1+1.28\sigma, \sigma^2) + \frac{1}{2}\mathrm{N}(1+1.28\sigma, \sigma^2)$. In addition, we add 21 "range examples" evenly spaced in the interval $x \in [-2.5, 2.5]$. The range examples ensure that the unlabeled examples for both groups span the same range, so that any measured shift in the decision boundary cannot be explained by differences in the range of examples viewed.

Block 4,5 (unlabeled-2,3) are identical to Block 3, each with the same 21 range examples, but with a different 230 random samples from the Gaussian mixture in Block 3. Blocks 3,4,5 are always all left-shifted or all right-shifted.

Block 6 (test-2) is identical to Block 2, consisting of 21 unlabeled examples evenly spaced in $[-1, 1]$. They are used to test whether the participant's decision boundary has changed after seeing the unlabeled blocks.

Procedure

Participants were told that they would see microscopic images of pollen particles from one of two fictitious flowers ("Belianthus" or "Nortulaca"), and were asked to classify each image by pressing the B or N key, respectively. They were instructed that they would receive audio feedback for the first 20 trials, after which they must make their best guess for a large set of items without any feedback. To ensure useful measurements of both speed and accuracy, participants were asked to respond as quickly as possible without making too many mistakes.

All participants saw the 815 stimuli in blocks 1 through 6 presented in order, but order within each block was randomized separately for each subject. In addition, 12 of the participants received Blocks 3,4,5 with left-shifted unlabeled stimuli (L-subjects), while the other 10 received right-shifted stimuli (R-subjects)[3].

Stimuli were displayed on a 15-inch CRT monitor in a darkened room at a normal viewing distance. The stimulus remained on-screen until a response was detected, after which the screen went blank for a duration of 1 second. Decisions and response times (time from onset of the stimulus to the detection of a key-press measured in milliseconds) were recorded for each trial. For each of the 20 stimuli in Block 1, the participants received an affirmative sound if they made the correct classification, or a warning sound if they were wrong. There was no audio feedback for the remaining stimuli.

The experiment manipulates one within-subjects factor (the category boundary is assessed either *before* or *after* exposure to the unlabeled data) and one between-subjects factor (unlabeled data distributions are shifted to the *left* or *right* of the labeled examples).

Results and discussion

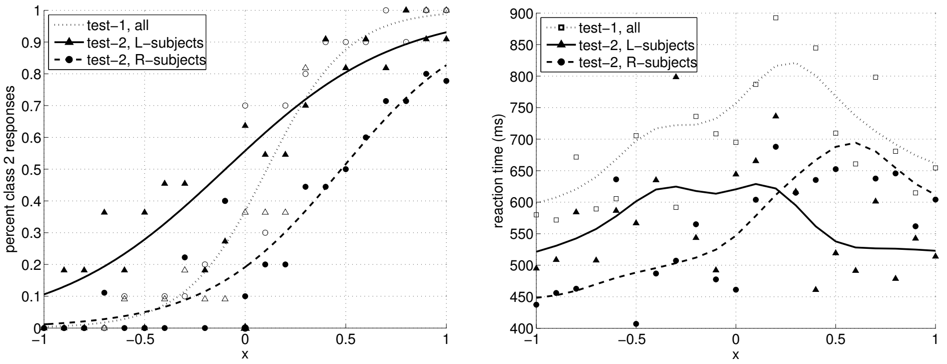
Data from 1 subject in the left-shift condition were discarded, as the participant appeared to "give up" halfway through the experiment, making the same "N" response for virtually every stimulus. From the remaining subjects (11 in left-shift condition and 10 in right-shift condition), we make the following observations:

Unlabeled data helps determine the decision boundary. We compared the participants' classification on blocks test-1 vs. test-2. In test-1, we expect the decision boundary to be around $x = 0$ for all participants, because they have just seen the same 20 labeled examples at $x = -1$ and $x = 1$ (Figure 3). If unlabeled data helps learning, the decision boundary in test-2 should shift towards left for L-subjects (Figure 3), or right for R-subjects. To quantify the decision boundary, we fit logistic regression functions $p(y = 2|x) = 1/(1 + \exp(-(\beta x + \beta_0)))$ to the data. For all participants on the test-1 block, the data consists of $(x, \hat{y})$ pairs, where $\hat{y} \in \{1, 2\}$ is each participant's classification on x within test-1. Figure 4(a) shows the best fit ($\beta = 4.99, \beta_0 = -0.54$, the dotted curve). The decision boundary is $x = -\beta_0/\beta = 0.11$ where $p(y|x) = 0.5$. This decision boundary is close to zero as expected[4]. The curve is also relatively steep, showing that the participants are highly consistent on their classifications.

The best fit for R-subjects after seeing the unlabeled data is shown by the dashed curve ($\beta = 3.00, \beta_0 = -1.44$). The decision boundary is at $x = 0.48$. This represents a shift to the right of $\Delta_R = 0.37$ on average, compared to the test-1 decision boundary. *This shift represents the effect of unlabeled data on the R-subjects, and fits the expectation of semi-supervised classification.* For L-subjects on test-2, the best fit is the solid curve ($\beta = 2.37$, $\beta_0 = 0.23$). The decision boundary is at $x = -0.10$, which represents a shift to the left by $\Delta_L = -0.21$, also consistent with semi-supervised learning. For visual inspection, we also show the empirical percentage of class 2 responses with different symbols in Figure 4(a).

[3] Data from 2 additional participants in the right-shift condition were lost when the computer crashed halfway through the experiment.

[4] It is not exactly zero because of small sample size, and potentially because the perceptual distance along the x-axis is not completely uniform. This, however, does not affect our conclusions.



(a) classification

(b) reaction time

Figure 4: (a) Unlabeled data helps classification. The curves are fitted logistic regression functions $p(y = 2|x)$. Clearly the test-2 decision boundaries shift toward the expected directions. Symbols represent empirical fraction of participants classifying a particular x as class 2, within: L-subjects in test-1 (△), R-subjects in test-1 (○), L-subjects in test-2 (▲), and R-subjects in test-2 (●). (b) Seeing unlabeled data shifts the perception of 'difficult stimuli', as revealed by reaction time. The difficult stimuli in each case correspond well with the decision boundaries.

To test the statistical reliability of these observations, we fit a separate logistic function to each individual participant's decision data for blocks test-1 and test-2, and from these curves computed the decision boundary for each subject on each test. Thus for each subject we obtained an estimate of the decision boundary *before* and *after* exposure to the unlabeled data. These data were subject to a repeated-measures analysis of variance assessing the influence of test block (1 versus 2, within-sj factor) and group (left-shift versus right-shift, between-sjs factor) on the location of the decision boundary. The results showed a significant interaction between the two factors (F(1,18) = 7.82, $p < 0.02$), indicating that after exposure to the unlabeled data, the decision boundary shifted in significantly different directions for the two groups.

Reaction time reflects decision boundary shift. Reaction time is the time elapsed between the appearance of the stimulus and the detection of a response–a long reaction time implies that the stimulus is relatively difficult to classify. It follows that stimuli near the decision boundary will be associated with longer reaction times. If the unlabeled data shift the participant's mental representation of the decision boundary, this shift should be reflected by a shift in the peak reaction time. Figure 4(b) verifies this hypothesis. The Figure shows mean reaction times (excluding outliers more than ± 3 beyond log reaction time) for each stimulus in the first test block, computed over all participants (squares in Figure 4(b)). The dotted curve shows the same data smoothed with a Gaussian kernel smoother. After seeing just the labeled examples at -1 and 1, people react quickly to examples near these points (rt $\approx$ 600ms), but are much slower ($\approx$800ms) for examples 'in the middle', that is, near the decision boundary. The peak is slightly to the right of the nominal decision boundary $x = 0$ for an unknown reason, which is consistent with Figure 4(a).

We then compute the average reaction time on block test-2 separately for L-subjects (black triangles and solid curve in the Figure) and R-subjects (black dots and dashed curve in the Figure). The overall reaction time on test-2 is faster than on test-1, reflecting the participants' greater familiarity with the experiment. More importantly, L-subjects have a reaction time plateau around $x = -0.1$, which is left-shifted compared to test-1, whereas R-subjects have a reaction time peak around $x = 0.6$, which is right-shifted. In line with the accuracy data, the reaction times suggest that exposure to the unlabeled data has shifted the decision boundary in different directions in the two groups.

A Semi-Supervised Model

In this section we consider whether the boundary-shifts reflected in the behavioral data are consistent with those predicted by a model developed for semi-supervised machine learning. We assume humans represent each category with a central prototype and a spread around the prototype. This allows us to model each category using a Gaussian distribution, whose mean and variance characterizes the prototype and the spread, respectively. Therefore our binary classification experiment can be modeled with a Gaussian mixture model (GMM) with two components[5], parameterized by $\theta = \{w_1, \mu_1, \sigma_1^2, w_2, \mu_2, \sigma_2^2\}$. The w's are non-negative component weights that sum to 1.

Before seeing any examples, we assume a prior distribution over θ. Learning involves updating the GMM parameters to best explain the observed labeled and unlabeled examples. One approach is to perform Bayesian analysis and compute the posterior distribution of θ (Tenenbaum 1999). However, to keep the model comparable with the existing semi-supervised learning literature (Nigam *et al.* 2000), we instead compute the *maximum a posteriori* (MAP) point estimate of θ. We assume exchangeability and let $D = \{(x_1, y_1), \ldots, (x_l, y_l), x_{l+1}, \ldots, x_n\}$ be the set of l

[5] This GMM is a cognitive model, not to be confused with the unlabeled-data-generating GMM in Blocks 3,4,5.

labeled and $n-l$ unlabeled examples seen so far. The MAP estimate $\mathrm{argmax}_\theta p(\theta|D)$, which represents the updated internal model, can be found (up to local maxima) with the standard EM algorithm (Dempster, Laird, & Rubin 1977). We use a factored, semi-conjugate prior distribution (Gelman *et al.* 2004) on θ: $p(\theta) = \prod_{k=1}^{2} p(w_k)p(\mu_k)p(\sigma_k^2)$, with $w_k \sim \mathrm{Uniform}[0,1]$, $\mu_k \sim \mathrm{N}(0,\infty)$, and $\sigma_k^2 \sim \mathrm{Inv}-\chi^2(\nu, s^2), k = 1,2$. Our priors are fairly benign: w_k is uniform over its range, and μ_k has a non-informative prior, making no assumption on what it might be. σ_k^2 has a scaled inverse-χ^2 distribution with scale s^2 and ν degrees of freedom, which is equivalent to ν pseudo observations with average squared deviation s^2. This prevents degeneracy in Gaussian variances. In our experiment we set $\nu = 1$, and $s^2 = \frac{25}{12}$ which is the variance of the uniform distribution over the range $[-2.5, 2.5]$.

We want to find θ that maximizes the posterior, which is equivalent to maximizing $\log p(\theta)p(D|\theta)$. The latter equals

$$\log p(\theta) + \sum_{i=1}^{l} \log p(x_i, y_i|\theta) + \lambda \sum_{i=l+1}^{n} \log p(x_i|\theta). \quad (2)$$

This objective is 'semi-supervised' because unlabeled data helps learning through the last term. To account for the possibility that an unlabeled example is perceptually 'worth less' than a labeled example, we introduced a weight λ above that can down-scale the contribution of unlabeled data. Such weight is common in prior work (Corduneanu & Jaakkola 2001; Nigam *et al.* 2000).

The objective (2) is difficult to optimize directly because the parameters are coupled in the $\log p(x_i|\theta)$ term. It is, however, not hard to derive the EM updates for our specific model. The derivation is standard and omitted for space considerations. We introduce hidden label distributions for each unlabeled example: $q_i(k) = p(y_i = k|x_i, \theta)$ for $i = l+1, \ldots, n$ and $k = 1, 2$. EM consists of iterating between the E-step and the M-step until convergence, which is guaranteed since our prior is log-concave. The **E-step** finds the expected distribution on hidden labels, given current model parameters θ:

$$q_i(k) \propto w_k \mathrm{N}(x_i; \mu_k, \sigma_k^2), \;\; i = l+1, \ldots, n; k = 1, 2. \quad (3)$$

The **M-step** updates the model parameters, given $q_i(k)$ above:

$$\begin{aligned}
\mu_k &= \frac{\sum_{i=1}^{l} \delta(y_i, k)x_i + \lambda \sum_{i=l+1}^{n} q_i(k)x_i}{\sum_{i=1}^{l} \delta(y_i, k) + \lambda \sum_{i=l+1}^{n} q_i(k)} \\
\sigma_k^2 &= \frac{\nu s^2 + \sum_{i=1}^{l} \delta(y_i, k)e_{ik} + \lambda \sum_{i=l+1}^{n} q_i(k)e_{ik}}{\nu + 2 + \sum_{i=1}^{l} \delta(y_i, k) + \lambda \sum_{i=l+1}^{n} q_i(k)} \\
w_k &= \frac{\sum_{i=1}^{l} \delta(y_i, k) + \lambda \sum_{i=l+1}^{n} q_i(k)}{l + \lambda(n-l)}, \quad (4)
\end{aligned}$$

where $\delta(y_i, k) = 1$ if $y_i = k$, and 0 otherwise; $e_{ik} = (x_i - \mu_k)^2$. Once the MAP θ is found through EM, prediction can be made with the Bayes rule, $p(y|x) = \frac{w_y \mathrm{N}(x;\mu_y,\sigma_y^2)}{\sum_{k=1,2} w_k \mathrm{N}(x;\mu_k,\sigma_k^2)}$. The corresponding decision boundary x can be found through the equation $w_1\mathrm{N}(x;\mu_1,\sigma_1^2) - w_2\mathrm{N}(x;\mu_2,\sigma_2^2) = 0$.

Model Fitting Results

The model predicts decision boundary shift. To model the participants' behavior on block test-1, we fit a GMM with EM on labeled and unlabeled data in blocks 1,2 (with initial parameters $w_k = 0.5$, $\mu_k = 0$, $\sigma_k^2 = 1$, and unlabeled data weight $\lambda = 0.06$, see below). This corresponds to a hypothetical subject who just saw blocks 1,2[6]. The GMM is $0.5\mathrm{N}(-0.97, 0.17) + 0.5\mathrm{N}(0.97, 0.17)$, whose classification is shown as the dotted curve in Figure 5(a), which corresponds to the empirical data (also dotted curve) in Figure 4(a). Then for the behavior on block test-2, we fit two GMMs on blocks 1–6 (results on blocks 1–5 are similar): For L-subjects who saw left-shifted unlabeled data in blocks 3,4,5, the fitted GMM is $0.49\mathrm{N}(-1.26, 0.20) + 0.51\mathrm{N}(0.71, 0.21)$. For R-subjects the GMM is $0.51\mathrm{N}(-0.74, 0.20) + 0.49\mathrm{N}(1.26, 0.18)$. *These two GMMs thus predict shifts of the decision boundary after seeing unlabeled data.* We show their classification curves (solid and dashed) in Figure 5(a), which qualitatively explains the empirical behavior in Figure 4(a).

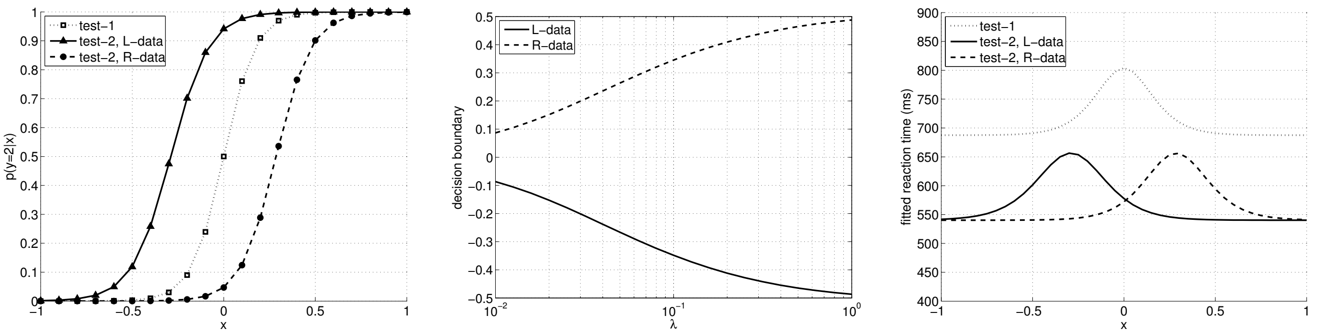
Unlabeled example weight λ controls the amount of decision boundary shift. The predicted amount of decision boundary shift is controlled by λ, the unlabeled example weight. By assigning an unlabeled example a small weight ($\lambda < 1$ in (2)), the shift is reduced as in Figure 5(b). This makes intuitive sense: As $\lambda \to 0$, the effect of unlabeled blocks diminishes, and both GMMs converge to the GMM trained on block 1 only. To account for the observed distance of 0.58 between L, R decision boundaries in Figure 4(a), $\lambda \approx 0.06$. This seems to indicate that people treat unlabeled examples less importantly than labeled examples.

The model explains reaction time. We model the reaction time with a sum of two parts: The first part is a base reaction time which decreases with experience. Let it be b_1 at block test-1, and a smaller b_2 later at test-2. The second part is proportional to the difficulty of each particular example. We assume if $p(y|x)$ is close to 0 or 1, the example x is easy because the classification is clear; x is difficult if $p(y|x)$ is close to 0.5. A natural measure of difficulty is the entropy of the prediction $h(x) = -\sum_{k=1}^{2} p(y = k|x) \log p(y = k|x)$, which is zero for $p(y|x) = 0$ or 1, and one for $p(y|x) = 0.5$. Our reaction time model is thus $ah(x) + b_i$ for block test-i. We find the parameters $a = 168$, $b_1 = 688$, $b_2 = 540$ with least squares from the empirical data in Figure 4(b). Our reaction time model is plotted in Figure 5(c), *which explains the empirical peaks before and after seeing unlabeled data in Figure 4(b).*

Conclusions and Discussion

We have designed and conducted a behavioral experiment that clearly demonstrates a form of semi-supervised learning in humans. Participants quickly learned from labeled data and set a stable category boundary midway between

[6] Alternatively, we can fit a GMM on block 1 only, the result is similar and not reported here. We can also fit a sequence of GMMs per subject one example at a time, following the exact randomized data stream in the blocks. Such detailed modeling gives very similar results here and below, and is not reported.



(a) predicted decision boundary shift

(b) λ determines shift amount

(c) reaction time fit

Figure 5: Our semi-supervised Gaussian mixture models (GMMs) explain experimental data.

the labeled items. After exposure to a set of unlabeled examples, however, the category boundaries shifted to reflect the distributions from which the unlabeled examples were drawn. The boundary-shifts were reflected both in the categorization decisions and in the mean reaction times.

We have also suggested that the boundary-shifts are well accounted for by a Gaussian mixture model of semi-supervised learning that has been successfully applied in machine learning. The GMM suggests that mental representations of categories consist of both the central tendency and the spread, and that these parameters are estimated from both labeled and unlabeled data.

One aspect of the behavioral data is not well explained by the GMM: decision curves after exposure to the unlabeled data were noticeably flatter than predicted. This apparent flattening is not an artifact of averaging across subjects–the slope of the logistic function estimated separately for each subject was significantly steeper before exposure to the unlabeled data than afterward (F(1,18) = 5.3, $p < 0.04$). In fact this flattening effect would be expected if participants systematically over-estimated the variance associated with each category. This interesting discrepancy in model and human behavior may therefore indicate important differences in human and machine memory. For instance, the current model retains a faithful representation of all past examples, and uses this perfect record to generate optimal estimates of the corresponding distributions. In human memory, traces of individual examples may degrade with time or may be subject to interference, so that decisions in the moment strongly weight more recent experiences. Future work will investigate these possibilities.

References

Baluja, S. 1998. Probabilistic modeling for face orientation discrimination: Learning from labeled and unlabeled data. *NIPS*.

Castelli, V., and Cover, T. 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Trans. Information Theory* 42(6):2101–2117.

Chapelle, O.; Zien, A.; and Schölkopf, B., eds. 2006. *Semi-supervised learning*. MIT Press.

Corduneanu, A., and Jaakkola, T. 2001. Stable mixing of complete and incomplete information. Technical Report AIM-2001-030, MIT.

Cozman, F.; Cohen, I.; and Cirelo, M. 2003. Semi-supervised learning of mixture models. In *ICML-03*.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*.

Gelman, A.; Carlin, J. B.; Stern, H. S.; and Rubin, D. B. 2004. *Bayesian data analysis*. Chapman & Hall/CRC, second edition.

Gielis, J. 2003. A generic geometric transformation that unifies a wide range of natural and abstract shapes. *American Journal of Botany* 90(3):333–338.

Graf Estes, K.; Evans, J. L.; Alibali, M. W.; and Saffran, J. R. 2006. Can infants map meaning to newly segmented words? Statistical segmentation and word learning. *Psychological Science*.

Langley, P. 2006. Intelligent behavior in humans and machines. Technical report, Stanford University.

Love, B. 2002. Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review* 9(4):829–835.

Mitchell, T. 2006. The discipline of machine learning. Technical Report CMU-ML-06-108, Carnegie Mellon University.

Mozer, M.; Jones, M.; and Shettel, M. 2006. Context effects in category learning: An investigation of four probabilistic models. In *NIPS*.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning* 39(2/3):103–134.

Ratsaby, J., and Venkatesh, S. 1995. Learning from a mixture of labeled and unlabeled examples with parametric side information. *COLT*.

Stromsten, S. B. 2002. *Classification learning from both classified and unclassified examples*. Ph.D. Dissertation, Stanford.

Tenenbaum, J. B., and Xu, F. 2000. Word learning as Bayesian inference. In *Proc. Cognitive Science Society*.

Tenenbaum, J. B. 1999. *A Bayesian framework for concept learning*. Ph.D. Dissertation, MIT.

Zhu, X. 2005. Semi-supervised learning literature survey. Technical Report 1530, Univ. Wisconsin-Madison.