Semi-Supervised Learning

an overview

Xiaojin "Jerry" Zhu

jerryzhu@cs.wisc.edu

Computer Science Department
University of Wisconsin, Madison

Outline

- background
- four methods
- open questions

A computer scientist's view ...

Why semi-supervised learning?

Because people want better performance for free.

Why semi-supervised learning?

Because people want better performance for free.

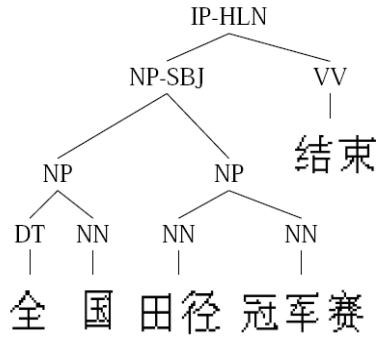
Exactly how hard can it be to obtain labeled data?

- Switchboard speech transcription, 400 hours for each hour of speech

  film ⇒ `f ih_n uh_gl_n m`

  be all ⇒ `bcl b iy iy_tr ao_tr ao l_dl`

- Penn Chinese Treebank, 2 years for 4000 sentences



'The National Track and Field Championship has finished.'

The landscape

supervised learning (classification, regression)

$\{(x_{1:n}, y_{1:n})\}$

$\updownarrow$

semi-supervised classification/regression

$\{(x_{1:l}, y_{1:l}), x_{l+1:n}\}$

$\updownarrow$

semi-supervised clustering $\{x_{1:n}$, must-, cannot-links$\}$

$\updownarrow$

unsupervised learning (clustering) $\{x_{1:n}\}$

transduction (limited to $x_{1:n}$) $\leftrightarrow$ **induction** (unseen data)

The problem

Goal:

Using both labeled and unlabeled data to build better learners (than using labeled data alone).

Notation:

- input features x, label y
- learner $f : \mathcal{X} \mapsto \mathcal{Y}$
- labeled data $(X_l, Y_l) = \{(x_{1:l}, y_{1:l})\}$
- unlabeled data $X_u = \{x_{l+1:n}\}$
- usually $l \ll n$

How can X_u help?

Method 1: generative models

Self-training:

1. Train f from (X_l, Y_l)
2. Predict on $x \in X_u$
3. Add $(x, f(x))$ to labeled data
4. Repeat

Naïve? error self-enforcing?

But if you set things just right, this is in fact the EM algorithm on mixture models ...

Method 1: generative models

Example: EM for Gaussian mixture models

$\theta = \{p(c), \mu, \Sigma\}_{1:C}$

Start from MLE θ on (X_l, Y_l), repeat:

1. E-step: compute the expected labels $p(y|x, \theta)$ for all $x \in X_u$
   - assign class 1 to $p(y = 1|x, \theta)$ fraction of x
   - assign class 2 to $p(y = 2|x, \theta)$ fraction of x
   - ...
2. M-step: update MLE θ with the original labeled and (now labeled) unlabeled data

Method 1: generative models

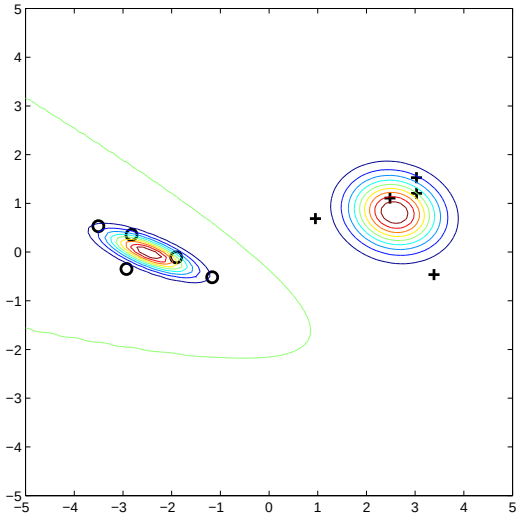
The MLE of θ without and with X_u is different.

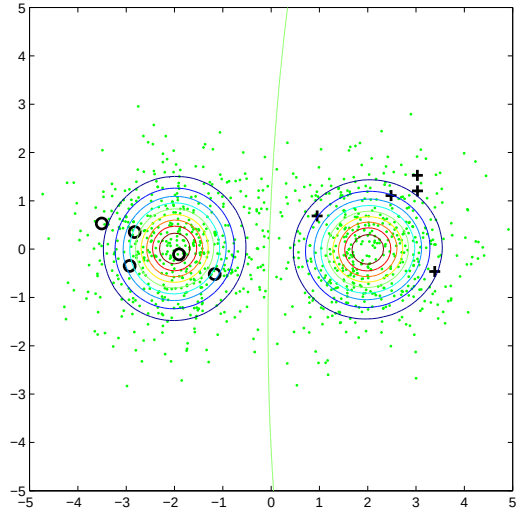
labeled data only

$$\log p(X_l, Y_l|\theta) = \sum_{i=1}^{l} \log p(y_i|\theta)p(x_i|y_i,\theta)$$

labeled and unlabeled

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

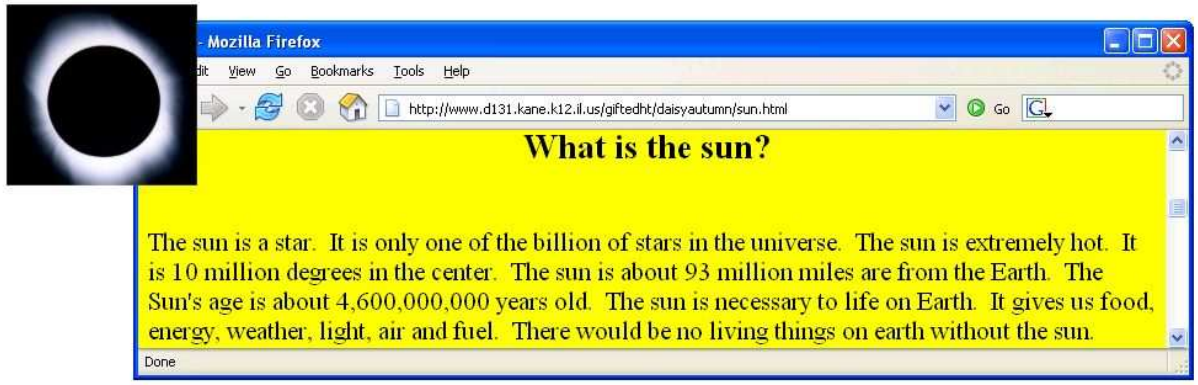




In principle X_u is useful for other generative models too.

Method 2: multi-view

Two views of an item: image and HTML text



Feature split $x = [x^{(1)}; x^{(2)}]$

Method 2: multi-view

Co-training:

1. Train $f^{(1)}$ 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 most-confident $(x, f^{(1)}(x))$ to $f^{(2)}$'s labeled data.
4. Add most-confident $(x, f^{(2)}(x))$ to $f^{(1)}$'s labeled data.
5. Repeat.

Encourages agreement between two classifiers.

Method 2: multi-view

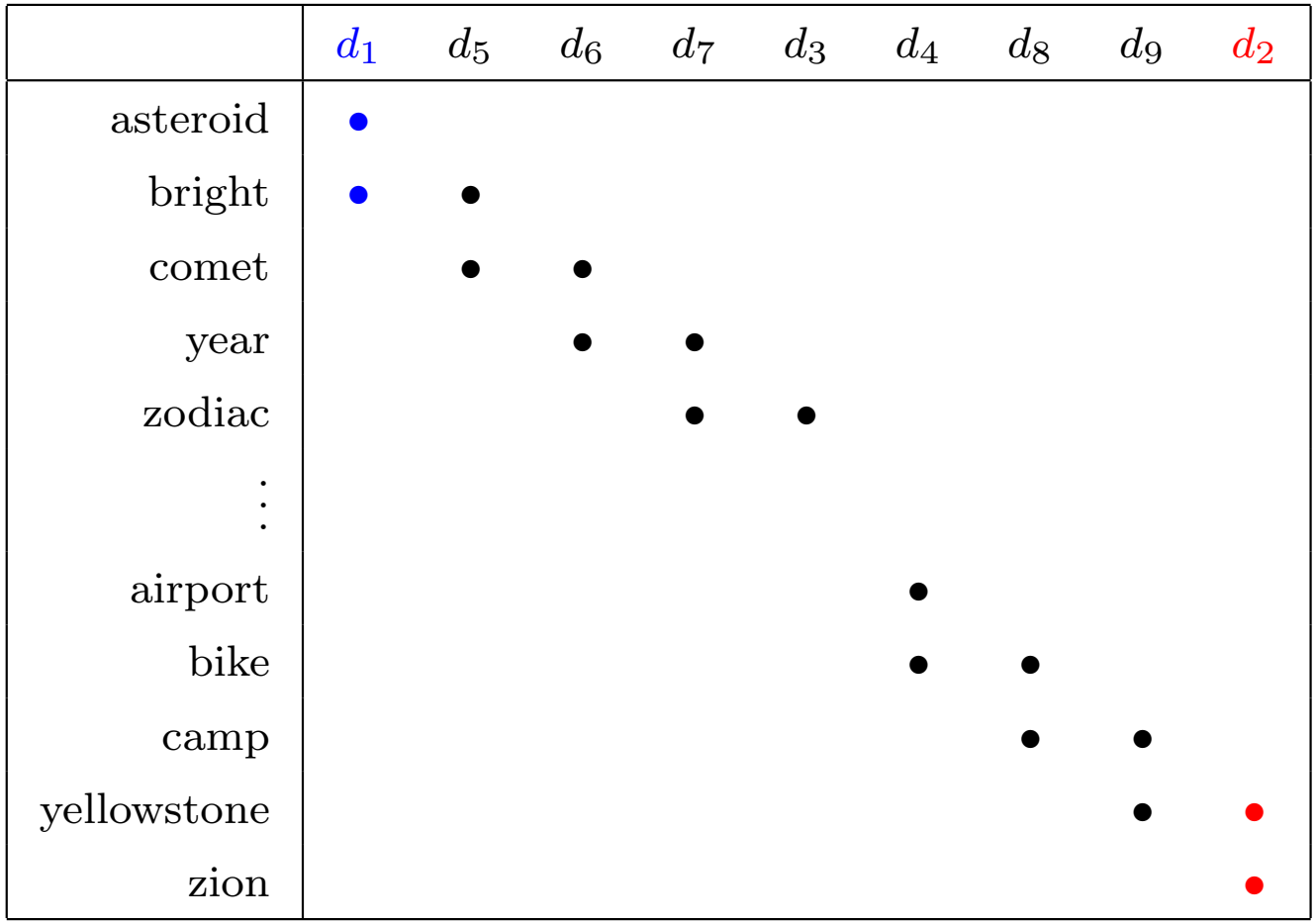
A regularized risk minimization framework to encourage multi-learner agreement:

$$\min_f \sum_{v=1}^{M} \left(\sum_{i=1}^{l} c(y_i, f_v(x_i)) + \lambda_1 \|f\|_K^2 \right) + \lambda_2 \sum_{u,v=1}^{M} \sum_{i=l+1}^{n} \left(f_u(x_i) - f_v(x_i)\right)^2$$

M learners, c loss function (e.g., hinge)

Method 3: graph-based

Example: Classify astronomy vs. travel articles.



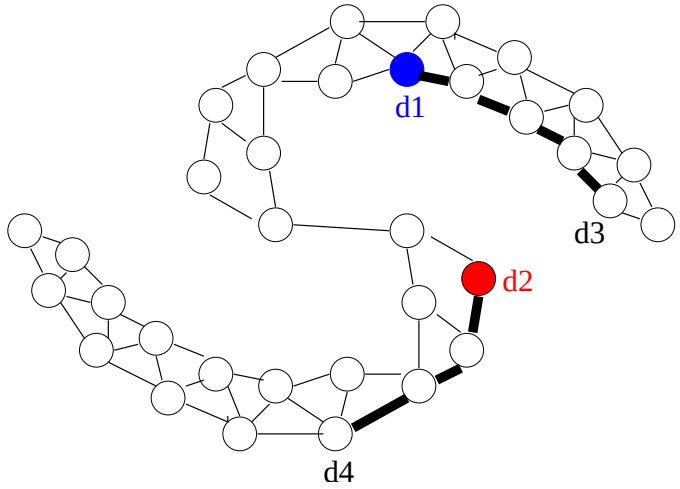
Unlabeled articles are stepping stones.

Method 3: graph-based

The graph has nodes $X_l \cup X_u$ and edges:

- k-nearest-neighbor unweighted graph
- fully connected graph, weights decay with distance $w = \exp\left(-\|x_i - x_j\|^2/\sigma^2\right)$
- other (expert knowledge)

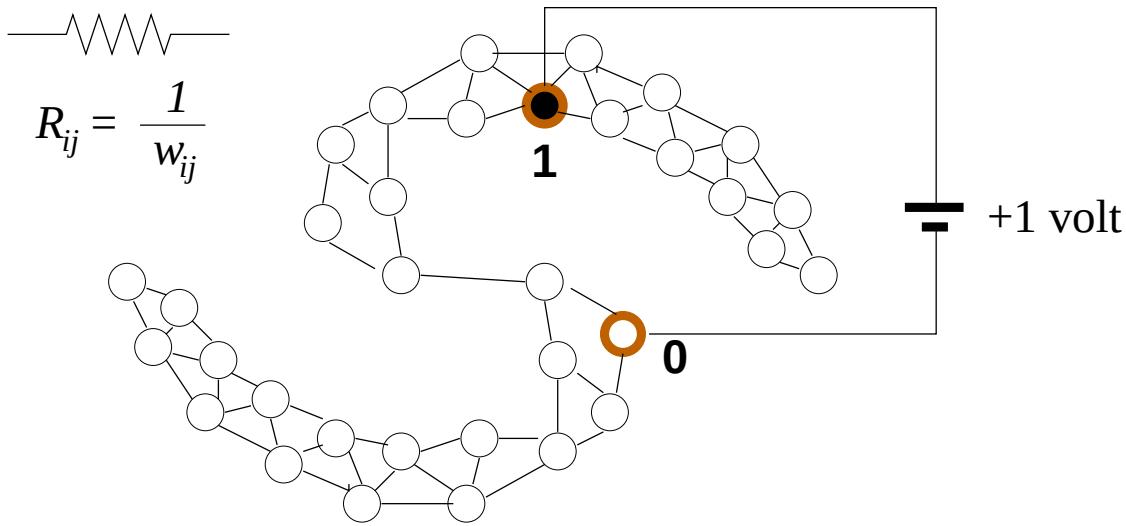
The graph is represented by an $n \times n$ weight matrix.



Method 3: graph-based

An electric network view:

- Edges are resistors with conductance w_{ij}
- 1-volt battery connects to labeled points $y = 0, 1$
- The voltage at the nodes is the harmonic function f



The harmonic function minimizes $\sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2$.

Method 3: graph-based

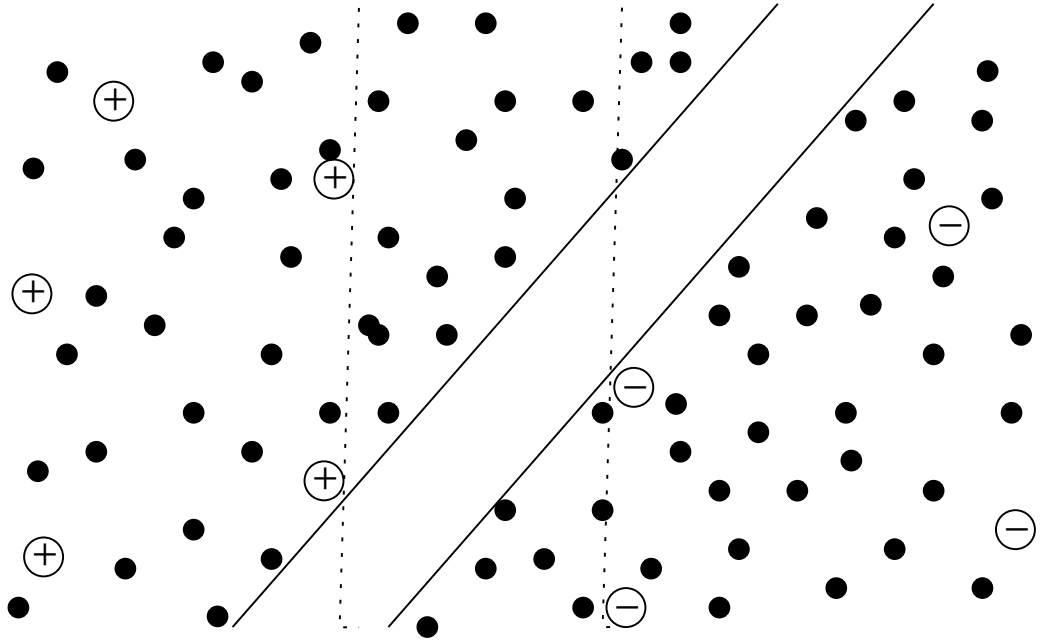
Manifold regularization extends the harmonic solution to handle

- noisy labels
- unseen examples

$$\min_f \sum_{i=1}^{l} c(y_i, f(x_i)) + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2$$

Method 4: S3VMs

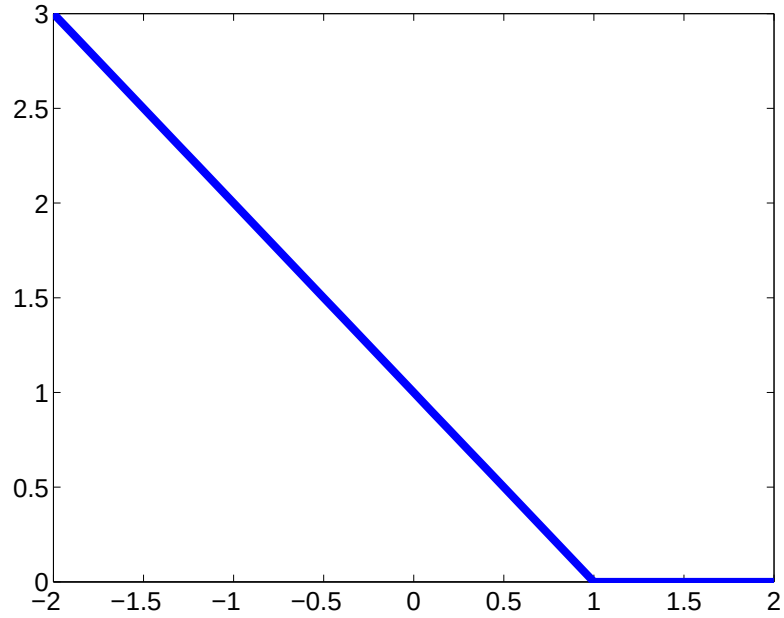
Semi-supervised SVMs (S3VMs, transductive SVMs): maximizing "unlabeled margin"



Method 4: S3VMs

standard SVM with hinge loss

$$\min_f \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda \|h\|_K^2$$



$y_i f(x_i)$

Prefers labeled points on the 'correct' side.

Method 4: S3VMs

How to incorporate unlabeled points?

- Assign putative labels $\text{sign}(f(x))$ to $x \in X_u$
- $\text{sign}(f(x))f(x) = |f(x)|$
- The hinge loss on unlabeled points

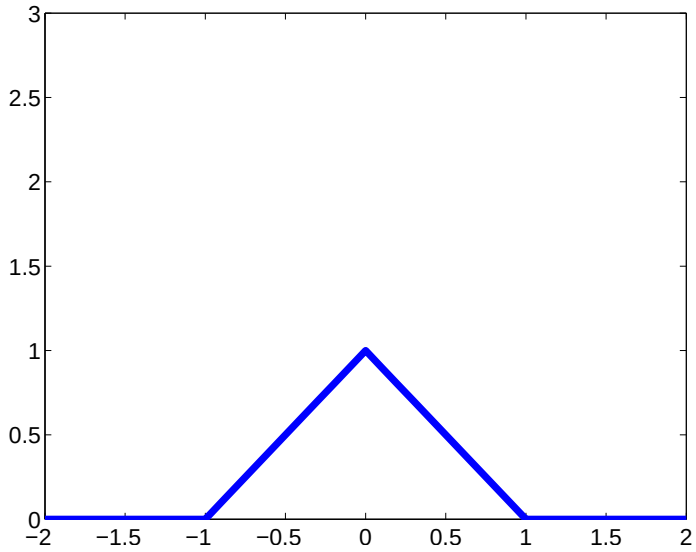
$$(1 - y_i f(x_i))_+ = (1 - |f(x_i)|)_+$$

Semi-supervised SVMs

$$\min_f \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

Method 4: S3VMs

The hat loss $(1 - |f(x_i)|)_+$ prefers $f(x) \geq 1$ or $f(x) \leq -1$.



$f(x_i)$

$$\min_f \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

The third term prefers unlabeled points outside the margin.

Question 1

What can we say about convergence, consistency, etc.?

- $n \to \infty$, l fixed
- $l \to 0$
- $l \to \infty$, $n \to \infty$, $l/n \to 0$

Question 2

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

$$\min_f \sum_{v=1}^{M} \left(\sum_{i=1}^{l} c(y_i, f_v(x_i)) + \lambda_1 \|f\|_K^2 \right) + \lambda_2 \sum_{u,v=1}^{M} \sum_{i=l+1}^{n} \left(f_u(x_i) - f_v(x_i)\right)^2$$

$$\min_f \sum_{i=1}^{l} c(y_i, f(x_i)) + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i,j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2$$

$$\min_f \sum_{i=1}^{l} (1 - y_i f(x_i))_+ + \lambda_1 \|f\|_K^2 + \lambda_2 \sum_{i=l+1}^{n} (1 - |f(x_i)|)_+$$

Question 2

- Why 4 methods?

Question 2

- Why 4 methods?
- Why not just 1 method?
  - What is the model that unifies semi-supervised learning?

Question 2

- Why 4 methods?
- Why not just 1 method?
  - What is the model that unifies semi-supervised learning?
- Why not 40 methods?
  - What are new semi-supervised learning approaches?

Question 3

no pain, no gain

Question 3

no model assumption, no gain

Question 3

no model assumption, no gain

wrong model assumption, no gain

Question 3

no model assumption, no gain

wrong model assumption, no gain, a lot of pain

Question 3

no model assumption, no gain

wrong model assumption, no gain, a lot of pain

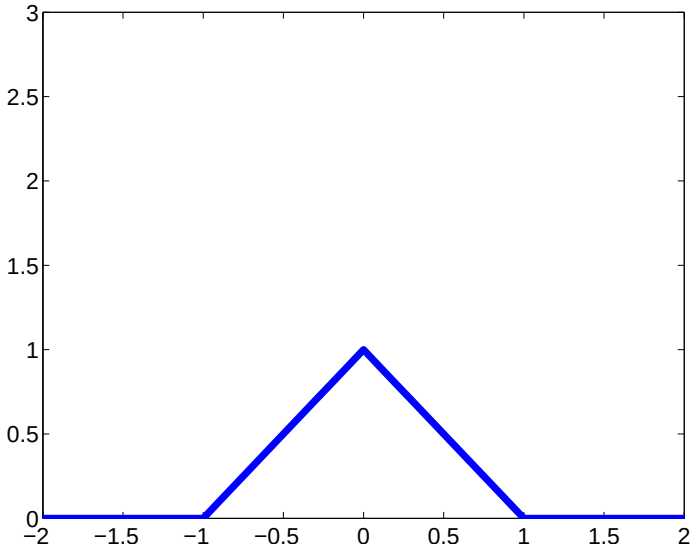
How do we know that we are making the right model assumptions?

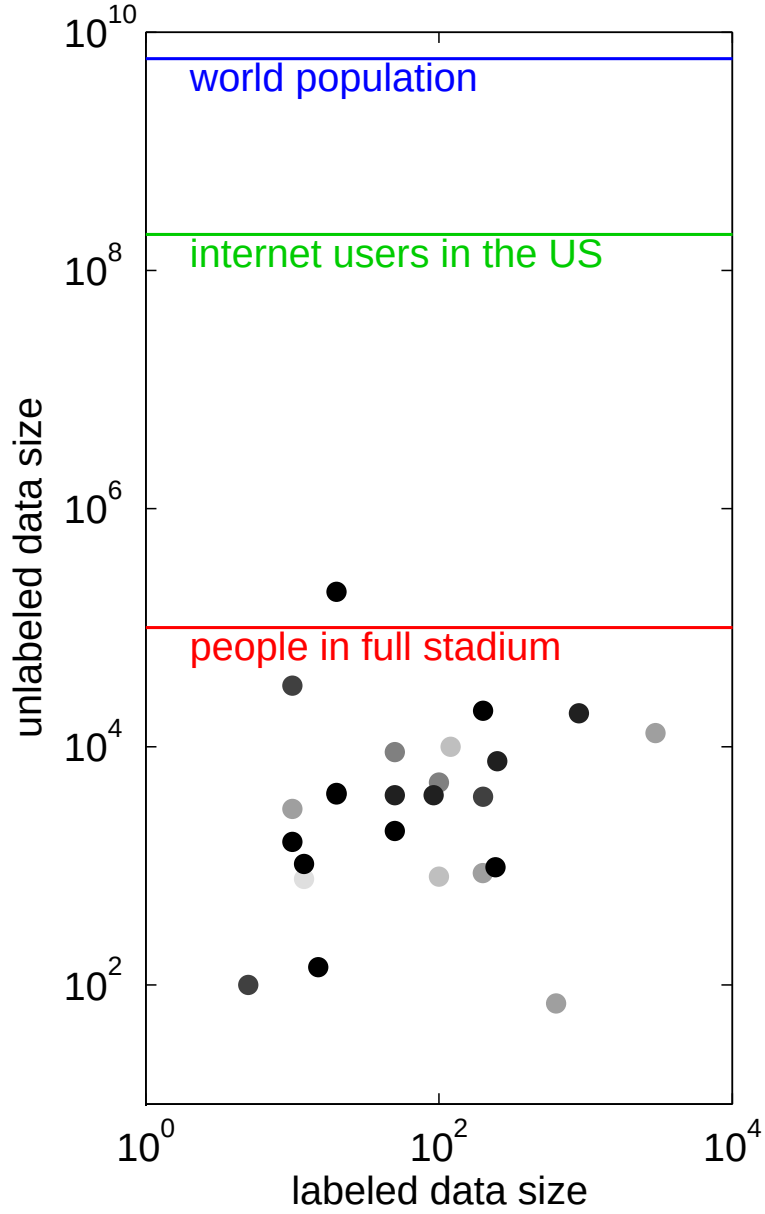
= Which semi-supervised learning method should I use?

Question 4

Are the methods practical?

e.g., S3VM is not convex





Question 5

Do humans do semi-supervised learning?

- 17-month-old infants word learning
- heard word before $\Rightarrow$ easier to associate the word with a visual object

References

google *semi-supervised learning survey*

thank you