Lecture 02: Semi-Supervised Learning (Part II)

Readings:

- Survey Paper (optional)

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

Semi-Supervised Clustering (vs. Semi-Supervised Classification)

supervised learning (classification, regression) $\{(x_{1:n}, y_{1:n})\}$

$\updownarrow$

semi-supervised classification/regression $\{(x_{1:l}, y_{1:l}), x_{l+1:n}, x_{test}\}$

transductive 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) $\{x_{1:n}\}$

We will mainly discuss semi-supervised classification.

To Read more about Semi-Supervised Clustering see this article!

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

SSL Assumptions

With Semi-supervised Learning we make "assumptions" about input data

– If the assumptions hold we gain; otherwise we lose!

– The more ambitious the assumptions the higher the reward, if correct!

SSL Assumptions (cont'd)

Example assumptions:

- Separation Assumption: Unlabeled data from different classes are separated with large margin (TSVM)
- Continuity Assumptions: Points that are close to each other are more likely to share a label (Graph-based approaches)
- Cluster Assumption: The data tend to form discrete clusters, and points in the same cluster are more likely to share a label (Graph-based approaches, Cluster-and-Label approach)
- Consistency Assumption: Various data features consistently label the data.

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

Semi-Supervised Clustering (vs. Semi-Supervised Classification)

- Pseudo-labeling refers to approaches used to label the data when the existing labeled data is too small to sufficiently train a model
- SSL is a pseudo-labeling approach!

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

Self-Training

Assumption

One's own high confidence predictions are correct.

Self-training algorithm:

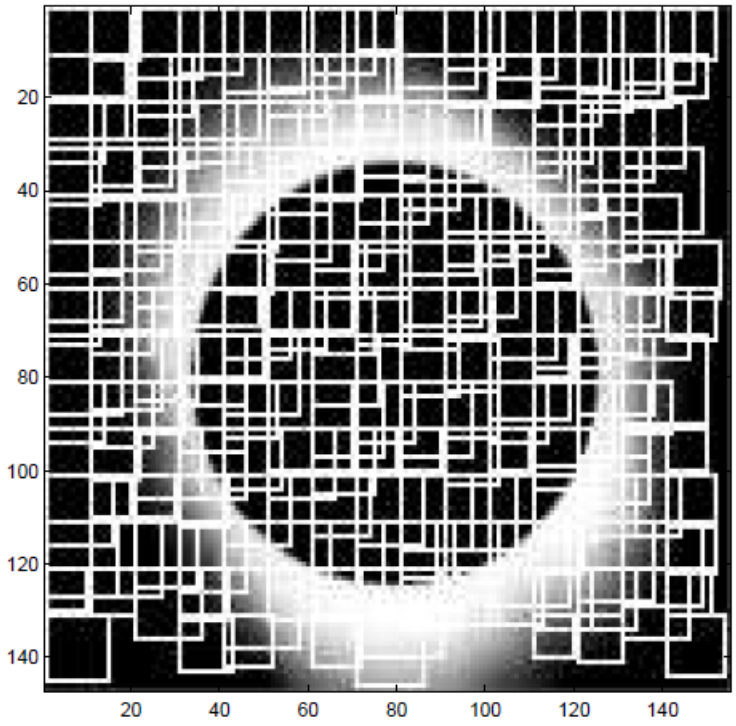
1. Train f from (X_l, Y_l)
2. Predict on $x \in X_u$
3. Add $(x, f(x))$ to labeled data
4. Repeat

Self-Training Variants

- Add a few most confident $(x, f(x))$ to labeled data
- Add all $(x, f(x))$ to labeled data
- Add all $(x, f(x))$ to labeled data, weigh each by confidence

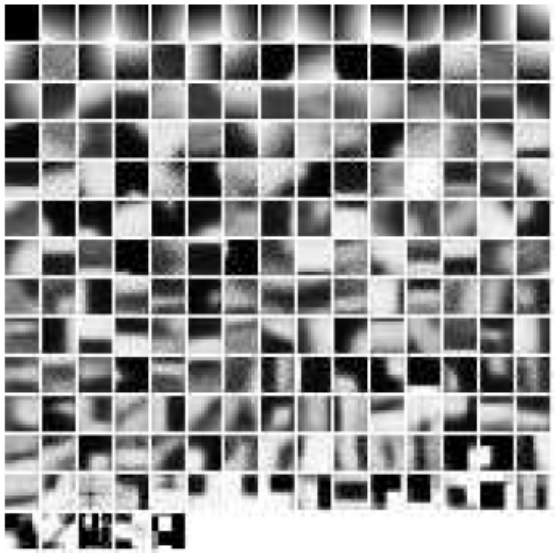
Self-Training Example: Image Categorization

- Each image is divided into small patches
- 10×10 grid, random size in $10 \sim 20$



Self-Training Example: Image Categorization
(cont'd)

- All patches are normalized.
- Define a dictionary of 200 'visual words' (cluster centroids) with 200-means clustering on all patches.
- Represent a patch by the index of its closest visual word.



The bag-of-word Representation of Images

→ 1:0 2:1 3:2 4:2 5:0 6:0 7:0 8:3 9:0 10:3 11:31 12:0 13:0 14:0 15:0 16:9 17:1 18:0 19:0 20:1 21:0 22:0 23:0 24:0 25:6

26:0 27:6 28:0 29:0 30:0 31:1 32:0 33:0 34:0 35:0 36:0 37:0 38:0 39:0 40:0 41:0 42:1 43:0 44:2 45:0 46:0 47:0 48:0 49:3 50:0

51:3 52:0 53:0 54:0 55:1 56:1 57:1 58:1 59:0 60:3 61:1 62:0 63:3 64:0 65:0 66:0 67:0 68:0 69:0 70:0 71:1 72:0 73:2 74:0 75:0

76:0 77:0 78:0 79:0 80:0 81:0 82:0 83:0 84:3 85:1 86:1 87:1 88:2 89:0 90:0 91:0 92:0 93:2 94:0 95:1 96:0 97:1 98:0 99:0 100:0

101:1 102:0 103:0 104:0 105:1 106:0 107:0 108:0 109:0 110:3 111:1 112:0 113:3 114:0 115:0 116:0 117:0 118:3 119:0 120:0

121:1 122:0 123:0 124:0 125:0 126:0 127:3 128:3 129:3 130:4 131:4 132:0 133:0 134:2 135:0 136:0 137:0 138:0 139:0 140:0

141:1 142:0 143:6 144:0 145:2 146:0 147:3 148:0 149:0 150:0 151:0 152:0 153:0 154:1 155:0 156:0 157:3 158:12 159:4 160:0

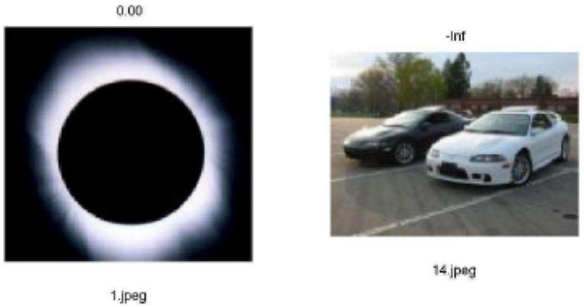
161:1 162:7 163:0 164:3 165:0 166:0 167:0 168:0 169:1 170:3 171:2 172:0 173:1 174:0 175:0 176:2 177:0 178:0 179:1 180:0

181:1 182:2 183:0 184:0 185:2 186:0 187:0 188:0 189:0 190:0 191:0 192:0 193:1 194:2 195:4 196:0 197:0 198:0 199:0 200:0

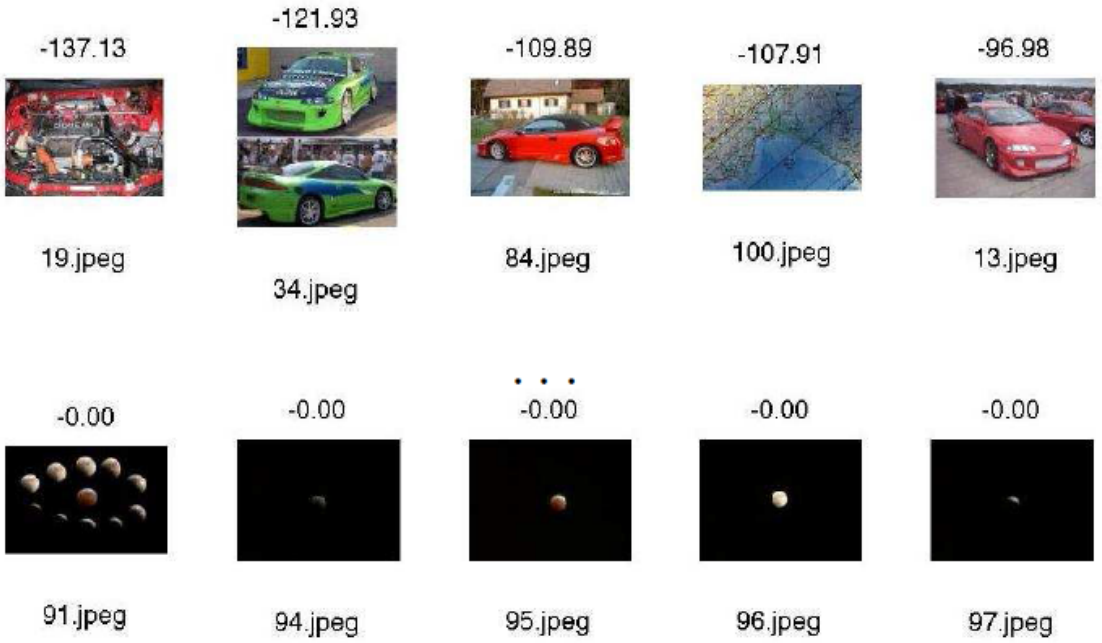
Self-Training Example: Image Categorization

(cont'd)

1. Train a naïve Bayes classifier on the two initial labeled images



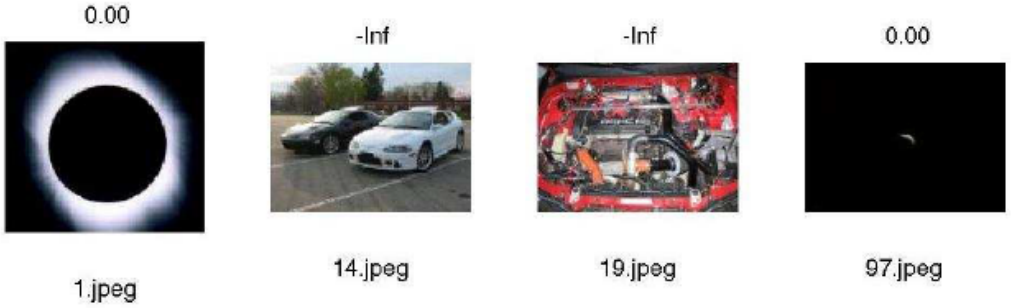
2. Classify unlabeled data, sort by confidence $\log p(y = \text{astronomy} | x)$



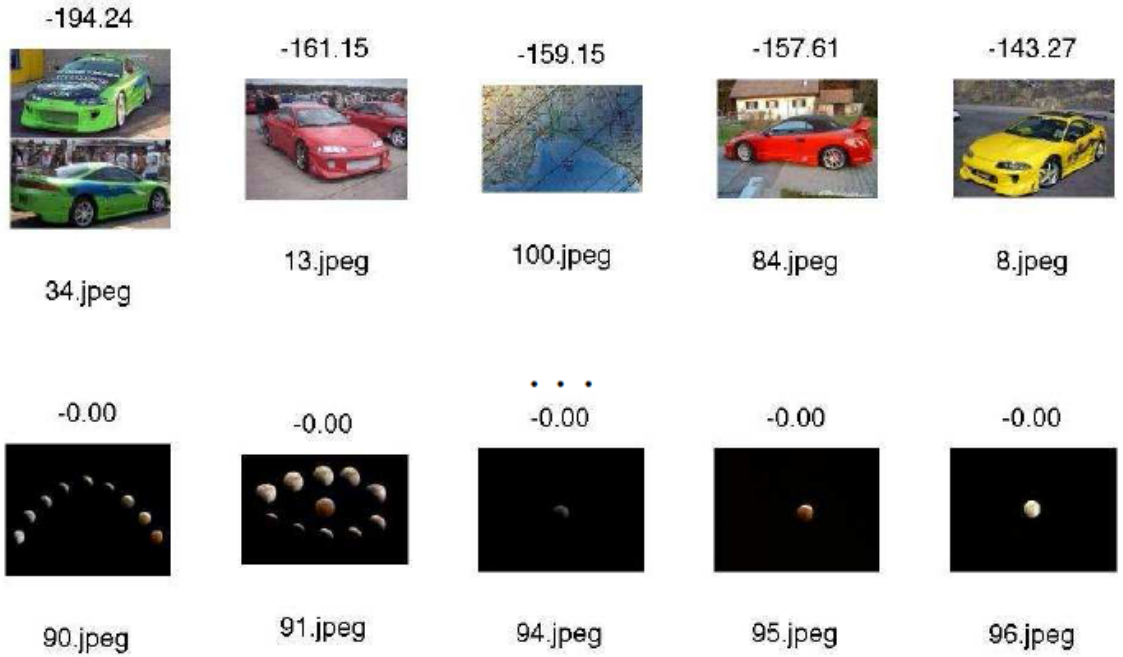
Self-Training Example: Image Categorization

(cont'd)

3. Add the most confident images and predicted labels to labeled data



4. Re-train the classifier and repeat



Advantages of Self-Training

- The simplest semi-supervised learning method.
- A wrapper method, applies to existing (complex) classifiers.
- Often used in real tasks like natural language processing.

Main Disadvantage of Self-Training

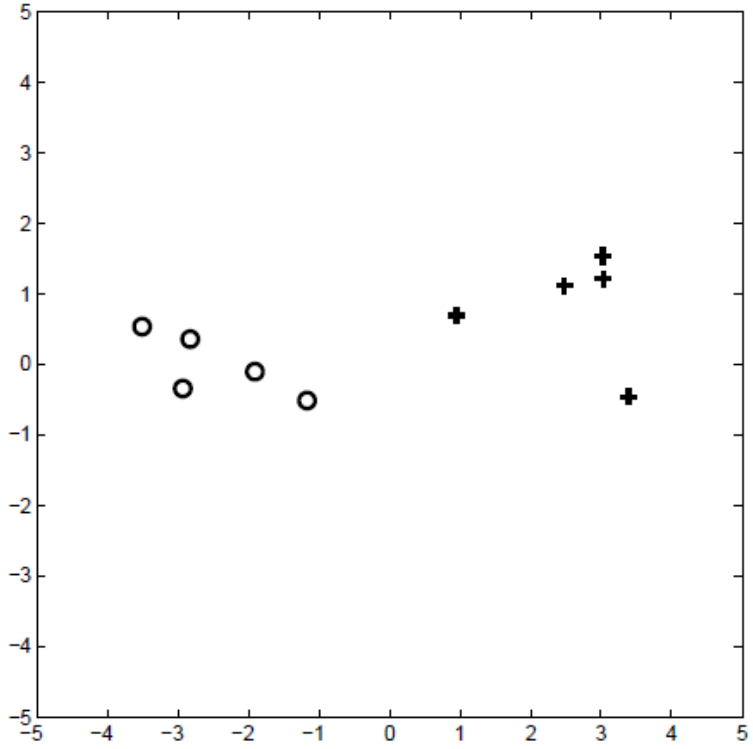
- Early mistakes could reinforce themselves.
  - Heuristic solutions, e.g. "un-label" an instance if its confidence falls below a threshold.

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

A simple example of generative models

Labeled data (X_l, Y_l):



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

A simple example of generative models
(cont'd)

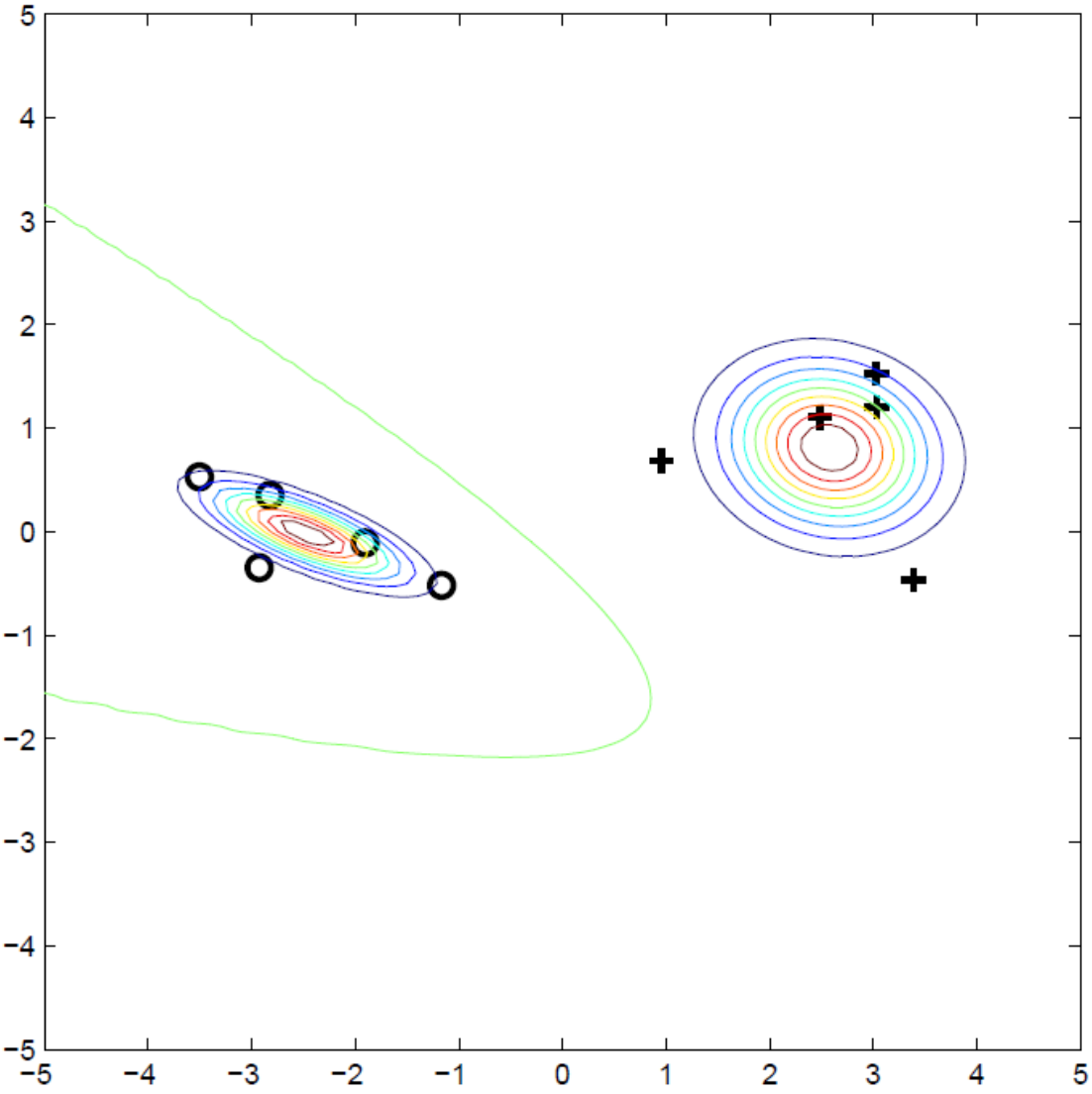
Model parameters: $\theta = \{w_1, w_2, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$
The GMM:

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

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

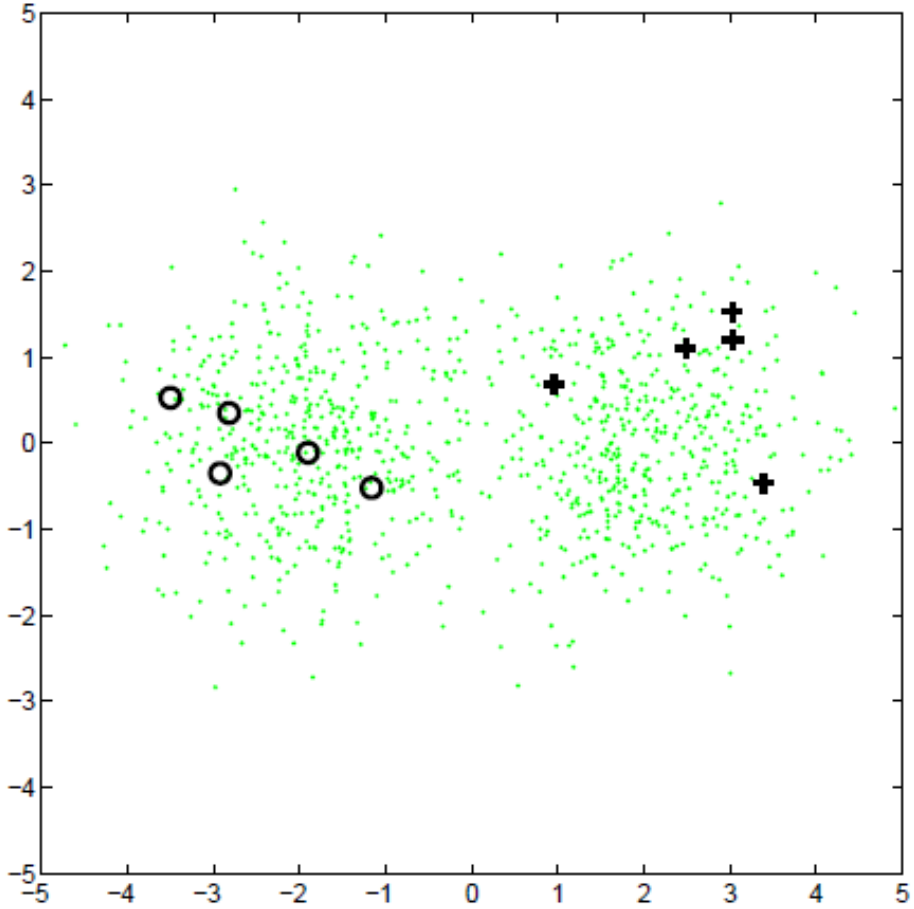
A simple example of generative models
(cont'd)

The most likely model, and its decision boundary:



A simple example of generative models
(cont'd)

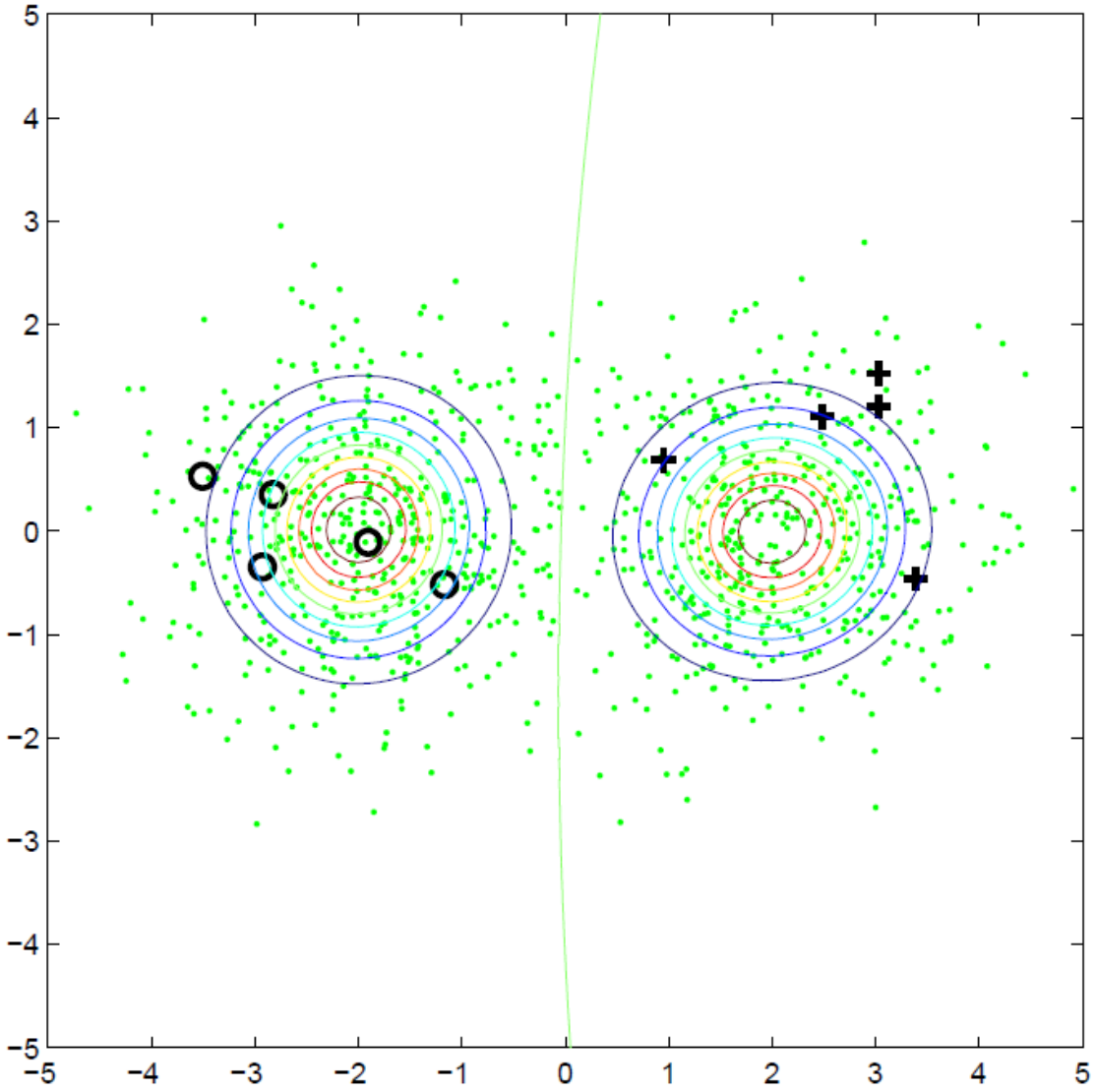
Adding unlabeled data:



A simple example of generative models

(cont'd)

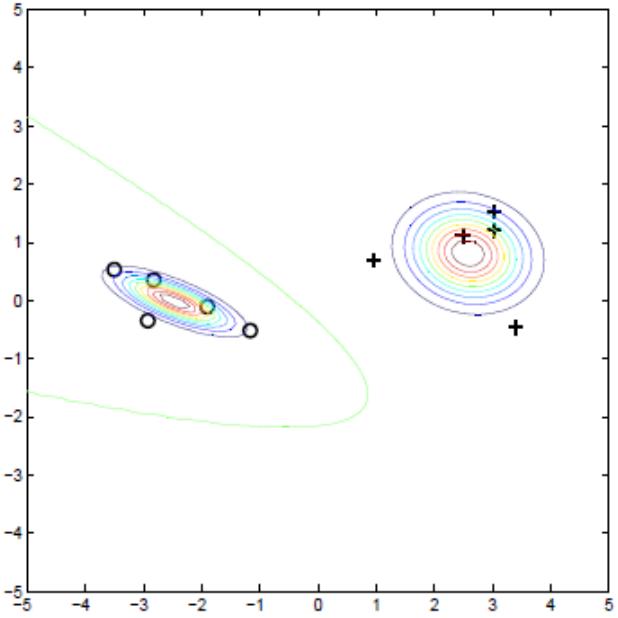
With unlabeled data, the most likely model and its decision boundary:



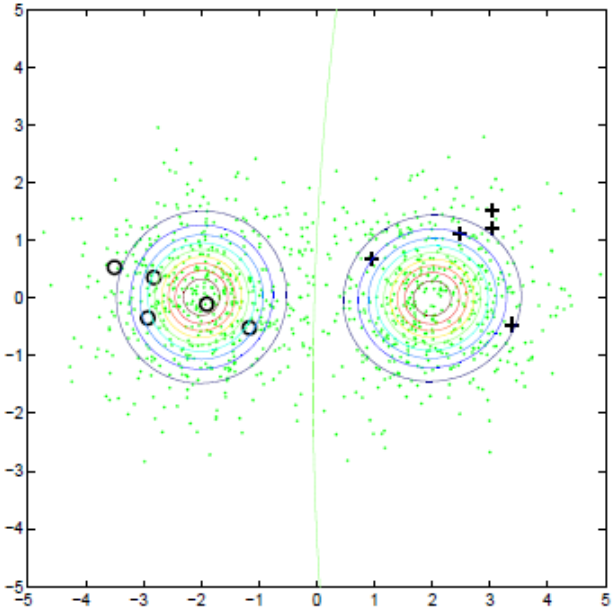
A simple example of generative models
(cont'd)

They are different because they maximize different quantities.

$p(X_l, Y_l|\theta)$



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



Generative Models for Semi-Supervised Learning

Assumption

The full generative model $p(X, Y|\theta)$.

Generative model for semi-supervised learning:

- quantity of interest: $p(X_l, Y_l, X_u|\theta) = \sum_{Y_u} p(X_l, Y_l, X_u, Y_u|\theta)$
- find the maximum likelihood estimate (MLE) of θ, the maximum a posteriori (MAP) estimate, or be Bayesian

Example of Generative Models

Often used in semi-supervised learning:

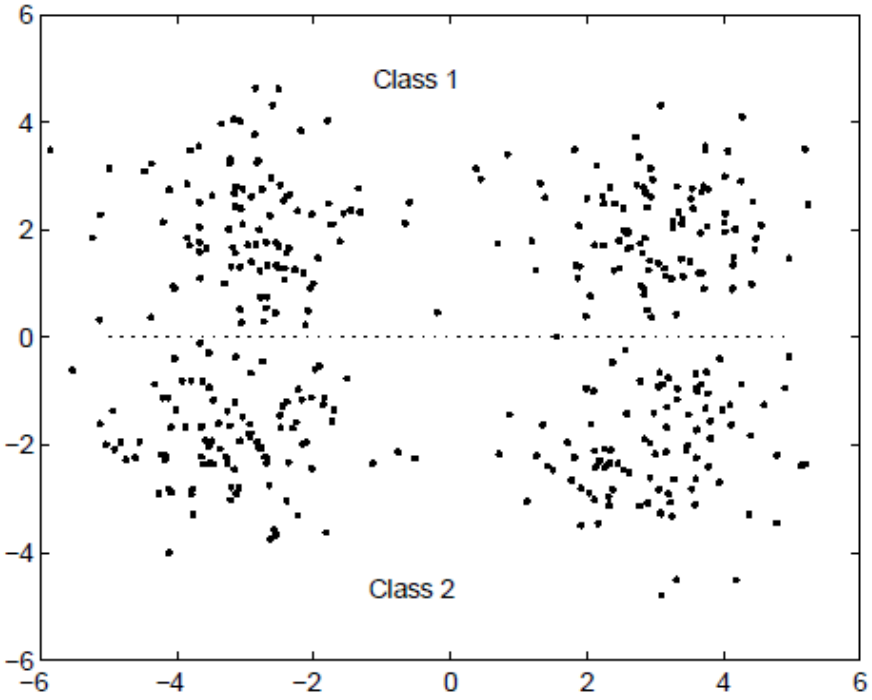
- Mixture of Gaussian distributions (GMM)
  - image classification
  - the EM algorithm
- Mixture of multinomial distributions (Naïve Bayes)
  - text categorization
  - the EM algorithm
- Hidden Markov Models (HMM)
  - speech recognition
  - Baum-Welch algorithm

Advantages of Generative Models

- Clear, well-studied probabilistic framework
- Can be extremely effective, if the model is close to correct

Disadvantages of Generative Models

- Often difficult to verify the correctness of the model
- Unlabeled data may hurt if generative model is wrong

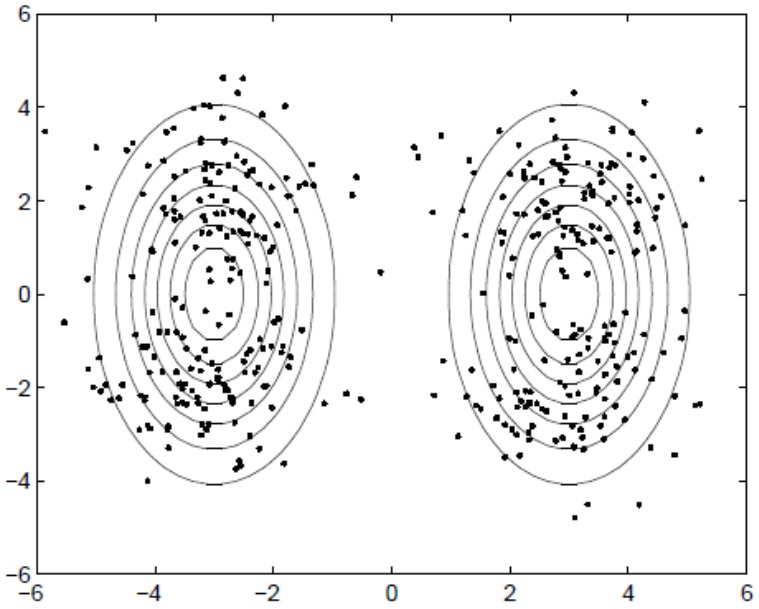


For example, classifying text by topic vs. by genre.

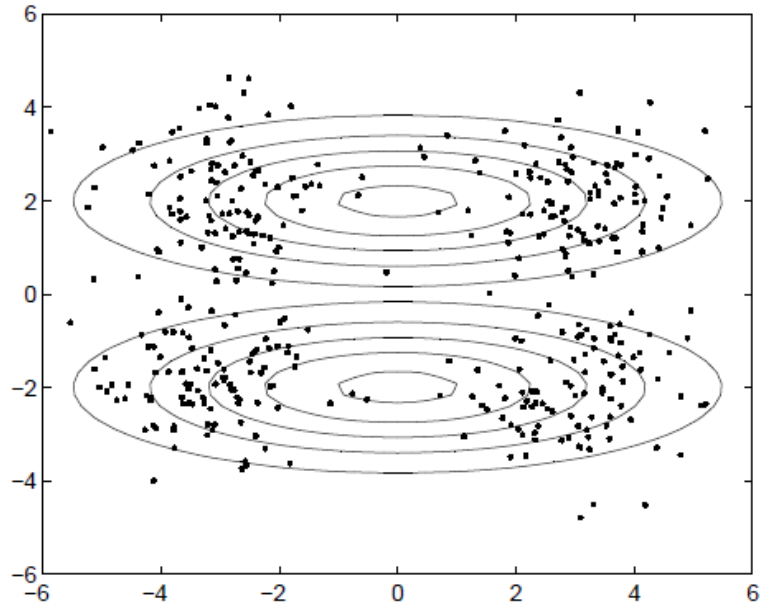
Disadvantages of Generative Models (cont'd)

If the generative model is wrong:

high likelihood
wrong



low likelihood
correct

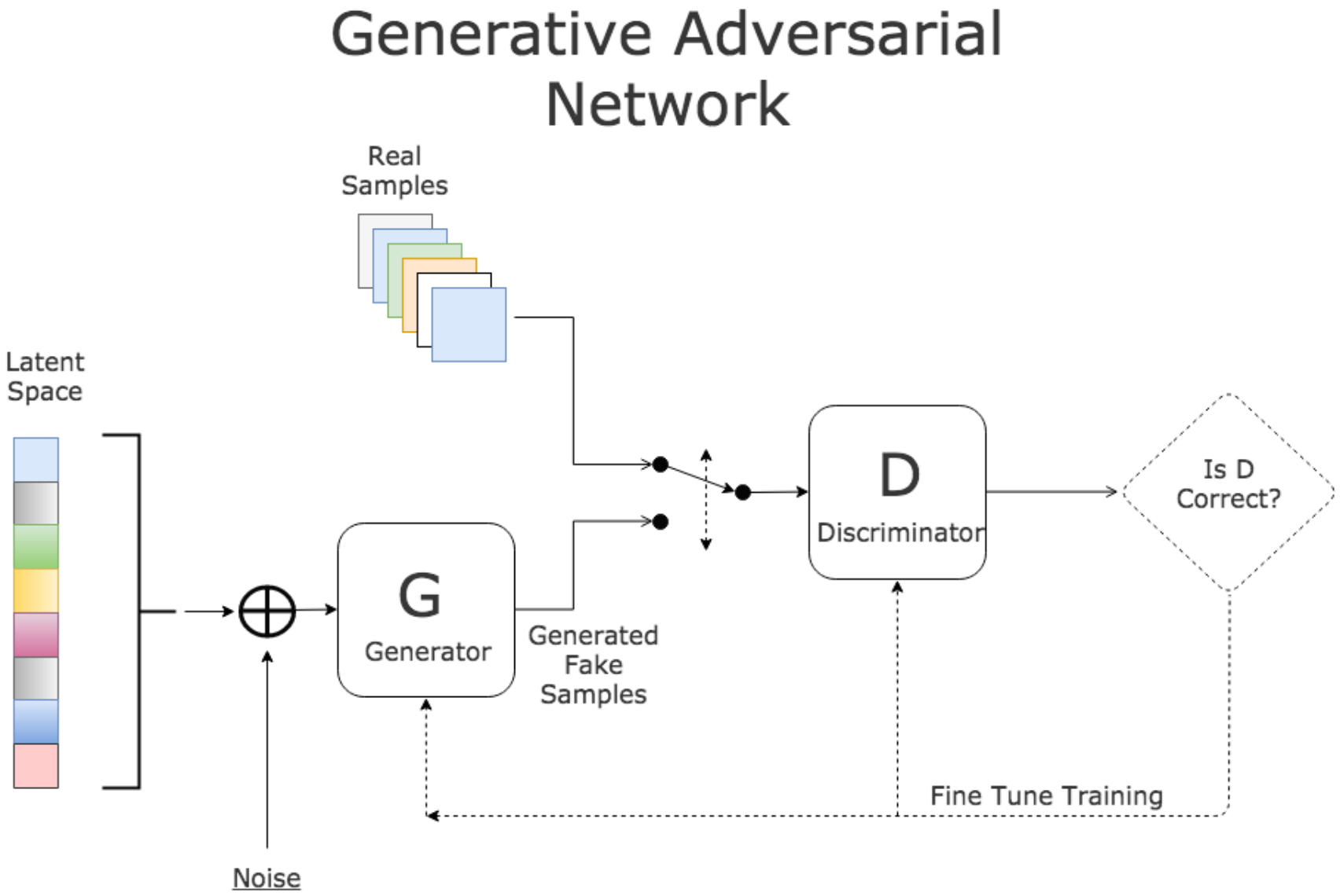


A Related Method: Cluster-and-Label

Instead of probabilistic generative models, any clustering algorithm can be used for semi-supervised classification too:

- Run your favorite clustering algorithm on X_l, X_u.
- Label all points within a cluster by the majority of labeled points in that cluster.
- Pro: Yet another simple method using existing algorithms.
- Con: Can be difficult to analyze.

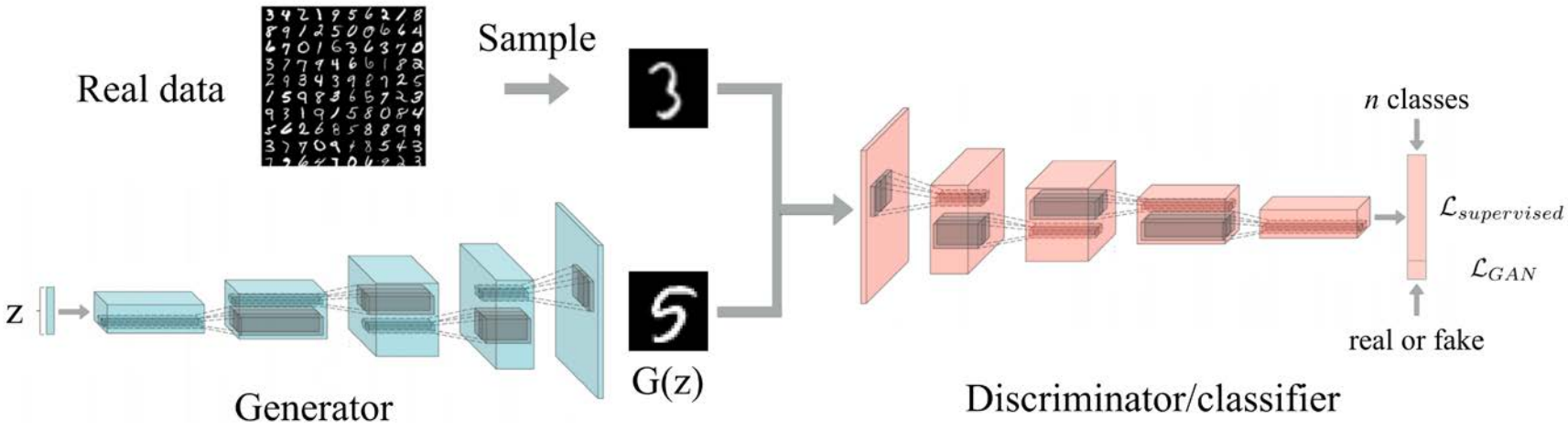
A Modern Generative Model Example:



GAN Logic

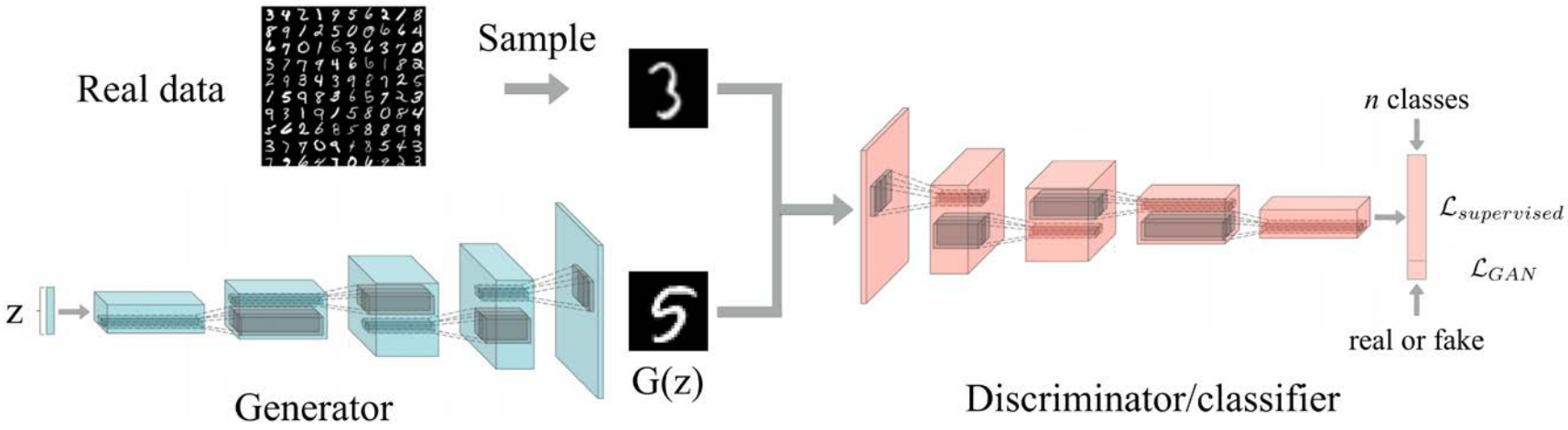
- The basic idea of GAN is to set up a **game between two players**:
  - A ***generator G***: Takes random noise ***z*** as input and outputs an image ***x***. Its parameters are tuned to get a high score from the discriminator on fake images that it generates.
  - A ***discriminator D***: Takes an image ***x*** as input and outputs a score which reflects its confidence that it is a real image. Its parameters are tuned to have a high score when it is fed by a real image, and a low score when a fake image is fed from the generator.

Using GAN for SSL



The vanilla architecture of discriminator has only one output neuron for classifying the R/F probabilities. We train both the networks simultaneously and discard the discriminator after the training as it was used only for improving the generator.

Using GAN for SSL (cont'd)



For the semi-supervised task, in addition to R/F neuron, the discriminator will now have 10 more neurons for classification of MNIST digits. Also, this time their roles change and we can discard the generator after training, whose only objective was to generate unlabeled data to improve the discriminator's performance.

Using GAN for SSL (cont'd)

Now the discriminator is turned into an 11-class classifier with 1 neuron (R/F neuron) representing the fake data output and the other 10 representing real data with classes. The following are true:

- The R/F neuron will output label = 0, when real unsupervised data from dataset is fed
- The R/F neuron will output label= 1, when fake unsupervised data from generator is fed
- The R/F neuron will output label = 0 and corresponding label output = 1, when real supervised data is fed

This combination of different sources of data will help the discriminator classify more accurately than, if it had been only provided with a portion of labeled data.

See this for implementation details for MNIST dataset.

Outline

- Some Generalizations
  - Semi-supervised Clustering
  - SSL assumptions
  - Another view: pseudo-labeling to learn from small labeled datasets
- Other SSL Approaches
  - Self-training
  - Generative Models
- SSL vs. Human Learning

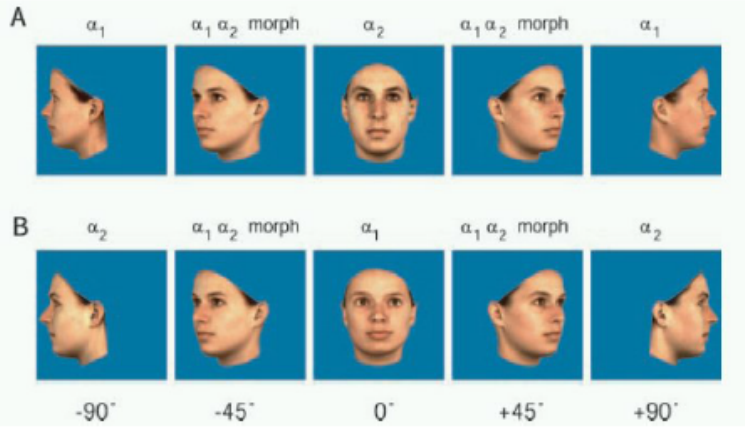
Does Human Learn from both Labeled and Unlabeled Data?

Learning exists long before machine learning.

- Do humans perform semi-supervised learning?
- Yes, it seems. We discuss two human experiments:
  1. visual recognition with temporal association
  2. infant word-object mapping

Visual Recognition with Temporal Association

- A face from two angles are very different, but we can easily associate it.
- The image sequence (unlabeled data) 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.



Infant Word-Object Mapping

- 17-month infants listen to a word, see an object
- Measure their ability to associate the word and object
  - If the word heard many times before (without seeing the object; unlabeled data), association is stronger.
  - If the word not heard before, association is weaker.

Similar to cluster-then-label.