Chapter12 Semi-Supervised Learning

1. Introduction
   a) Transductive Learning: Unlabeled data is the testing data
   b) Inductive Learning: Unlabeled data is not the testing data
   c) Collecting data is easy, but collecting 'labeled' data is expensive
   d) We do semi-supervised learning in our lives
   e) The distribution of the unlabeled data tells us something
2. Generative Model
   a) Supervised Generative model
      i) Prior Probability: $P(C_i)$ Class-Dependent Probability: $P(x|C_i)$
      ii) $P(x|C_i)$ is a Gaussian distribution parameterized by μ_i and Σ
      iii) $P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$
   a) Semi-Supervised Generative Model
      i) The unlabeled data x^u help re-estimate $P(C_1), P(C_2), \mu_1, \mu_2, \Sigma$
      ii) E-M Algorithm

         Step1: Compute the posterior probability of unlabeled data

         Step2: Update Model
3. Low-density Separation Assumption: Black or white
   a) Self-Training
      i) Train model f^* from labeled data set
      ii) Apply f^* to the unlabeled data, obtain $\{(x^u, y^u)\}_{u=R}^{R+U}$ (Pseudo Label)
      iii) Remove a set of data from unlabeled set, and add them into labeled set
      iv) Self-training uses hard label, and generative model uses soft label
   b) Entropy-based Regularization
      i) Entropy of y^u : evaluate how concentrate the distribution y^u is
      ii) $E(y^u) = -\sum_{m=1}^{5} y_m^u \ln(y_m^u)$ should be as small as possible
      iii) $L = \sum_{x^r} C(y^r, \hat{y}^r) + \lambda \sum_{x^u} E(y^u)$
   c) Semi-supervised SVM
      i) Enumerate all possible labels for the unlabeled data
      ii) Find a boundary that can provide the largest margin and least error
4. Smoothness Assumption: You are known by the company you keep
   a) Smoothness Assumption
      i) Assumption: "similar" x has the same $\hat{y}$
      ii) More precisely:

         x is not uniform

         If x^1 and x^2 are close in a high-density region, $\hat{y}^1$ and $\hat{y}^2$ are same
   b) Cluster and then label
   c) Graph-based Approach
      i) Represent the data points as a graph
      ii) Graph representation is nature sometimes

         Hyperlink of webpages

         Citation of papers
      iii) Sometimes you have to construct the graph yourself

iv) Define the similarity $s(x^i, x^j)$ between x^i and x^j
K Nearest Neighbor
e-Neighborhood

v) Labeled data influence their neighbors, propagate through the graph

vi) Define the smoothness of labels on the graph

$$S = \frac{1}{2}\sum_{i,j} w_{i,j}\left(y^i - y^j\right)^2 = y^T L y$$

Smaller means smoother

$$L = \sum_{x^r} C(y^r, \hat{y}^r) + \lambda S$$

5. Better Representation
   a) Find the latent factors behind the observation
   b) the latent factors are better representations

Chapter13 Deep Auto-Encoder

1. Auto-Encoder
   a) Compact representation of the input object
   b) Reconstruct the original object
2. PCA
   a) Bottleneck layer
   b) Decoding matrix is the transpose of the encoding matrix
3. Deep Auto-Encoder
   a) De-noising auto-encoder
   b) Text-Retrieval
      i) Vector Space Model
      ii) The documents talking about the same thing will have close code
   c) Similar Image Search
4. Auto-Encoder for CNN
   a) Unpooling
      i) Alternative: simply repeat the values
   b) Deconvolution
      i) Actually, deconvolution is convolution
5. Pre-train DNN