Lecture #20: Unsupervised Learning

Instructor: Aditya Bhaskara *Scribe: Maryam Barouti*

CS 5966/6966: Theory of Machine Learning

March 29th, 2017

Abstract

1 Introduction

In last session, we talked about regularization which is a way to avoid overfitting. Explicit techniques which can be used for regularization are dropout and l_1/l_2 regularization. We should be careful about network structure when applying these techniques.

2 Definition of Unsupervised Learning and Examples

2.1 definition. In contrast with *supervised learning*, in unsupervised learning there is *NO* labeling for the data points.

Here is examples of unsupervised learning:

- Example 1: (Matrix completion or movie recommendation problem)
  A set of users has initially rated some subset of movies (on the scale 1 to 5). The goal is to recommend movies to a user that he/she might be interested.

$$\begin{array}{c} \\ M_1 \\ M_2 \\ \vdots \\ M_k \end{array} \begin{array}{c} \begin{array}{cccc} u_1 & u_2 & \dots & u_n \end{array} \\ \begin{bmatrix} & & 5* & \\ 2* & & & \\ & & 4* & \\ 3* & & & \end{bmatrix} \end{array}$$

  In this example, there is no labeling for the movies. There is some information about the preferences of users.

- Example 2: (Cocktail Party Problem)
  Suppose you are at a cocktail party and the microphone is the superposition of a whole bunch of signals. The goal is to break up the microphone sound to its components.
  In this problem, we do not know the labeling for each person's sound.

- Example 3: (Sparse coding)
  This problem answers this question in neuroscience: How humans visual system can interpret images? We know every image is a sparse combination of a few patterns. In this problem, there's a bunch of signals (all

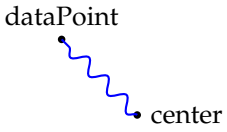
the images) and the goal is to find a set of patterns which satisfies these signals.

Now we come up with a question: How we think about unsupervised learning? The answer would be *Generative Model*.

3 Generative Model

3.1 definition. In generative model, we *assume* that data is generated from a random process with few parameters.

In clustering problems, we assume the data are chosen from a guassian distribution. In figure below, the point "dataPoint" is chosen by a guassian distribution with mean "center" .



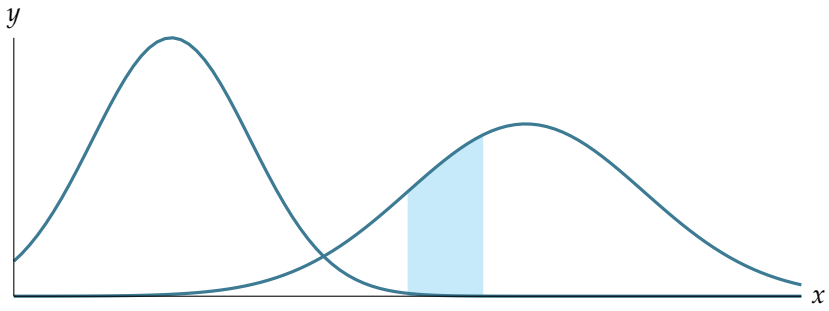
In movie recommendation systems example, suppose each movie is a sparse combination of generes, for example movie M_i is $0.4 \times action + 0.6 \times drama$ and every user j has preferences l_j. We can "pretend" that the user preferences are generated from linear combination of generes and l_j's.
Now we can ask: Is there any right model? The answer is NO. And aslo we should consider, all the issues such as overfitting and trade off between complexity and overfitting will happen in unsupervised learning.

4 Unsupervised Learning: Clustering

In this lecture, we will mostly talk about the data clustering which is unsupervised learnig. Suppose there is some points on a line, what is a simple way to model this?

One way to think about this data points is generative model. Now we should find what is the reasonable model and what are the parameters?
In this problem, polynomial models are a bad choice, since they blow up very easily. A good model for clustering problem is to think of the distribution of points as a superposition of gaussian distribution, but how many different gaussian distrubution?
Suppose the actual data distribution is like this $0.5g_1 + 0.3g_2 + 0.2g_3$ which is known as *Mixture of Guassian*. We know every guassian distribution has *two* parameters: mean (μ) and variance (σ^2).
In the mixture of guassian model, we need more parameters: the number of guassian distributions (k), the mean and variance for each guassian distribution (μ_i and σ_i^2) and also the weight for each guassian distribution (ω_i). Once

we find the parameters, we will have a way to generate (potentially infinity) data points.
What does data distribution mean in gaussian mixture model?



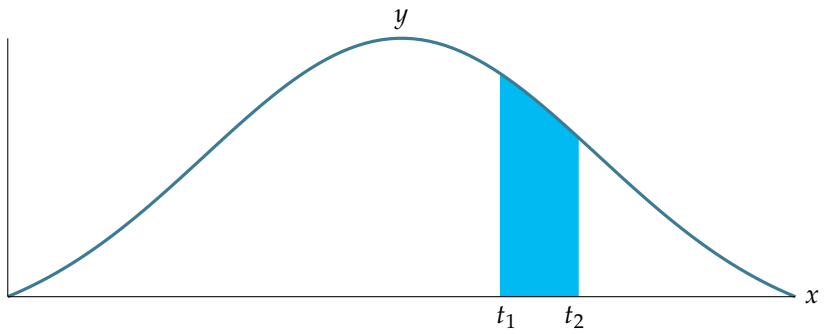
If we assume the probability mass is 1, the data distrubtion means the fraction of data points that belong to a specific interval (For example, blue area in figure above) in the gaussion mixture plot.

5 Data Distribution in Mixture of Gaussian Model: Likelihood

A formal way to measure how good a model fits data points is *likelihood*. Likelihood is defined as:

$$Pr[\text{generating data}|\mu_i, \sigma_i^2, \omega_i, k]$$

Suppose the parameters $\mu_i, \sigma_i^2, \omega_i, k$ is associated with a guassian mixture model, this measure calculates the likelihood of generating given data points using this model.
For one guassian distribution which mean and variance are 0 and 1.



The probability of generating points t_1 and t_2 is equal to $\frac{1}{\sqrt{2\pi\sigma^2}}.e^{-\frac{(t_1^2)}{2\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}}.e^{-\frac{(t_2^2)}{2\sigma^2}}$. In general, if f is the probability density function of the distribution, the probability of generating points $\{x_1, x_2, \ldots, x_n\}$ would be:
$f(x_1).f(x_2)\ldots.f(x_n)$.
A small probability is not good, since we want to find parameters that maximize this probability. *Maximum Likelihood Estimator* which is defind as:

$$\underset{\theta}{\text{argmax}}\, Pr(\text{generating data}|\theta)$$

will help us find a model which is most likely to generate the data points. Now we want to find to use MLE to find the best singular gaussian to generate data points $\{x_1, x_2, \ldots, x_n\}$:

$$Pr(x_1, x_2, \ldots, x_n | \mu, \sigma^2) = (\frac{1}{\pi\sigma})^{\frac{n}{2}} \times e^{-(\sum_i (\frac{x_i - \mu}{\sigma^2}))}$$

Which is equivalent to minimizing the $\sum_i (\frac{x_i - \mu}{\sigma^2})$. Therefore:

$$\underset{\mu,\sigma}{\operatorname{argmax}}\, Pr(X|\mu, \sigma) = \underset{\mu,\sigma}{\operatorname{argmin}} \sum_i (\frac{x_i - \mu}{\sigma^2})$$

For a model which is a mixture of two gaussian distribution, if f is the probability density function, the probability of generating data points $\{x_1, x_2, \ldots, x_n\}$ would be:

$$Pr(x_1, x_2, \ldots, x_n | \mu_i, \sigma_i^2) = \Pi_i (f(x_i))$$

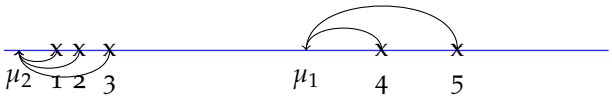
Considering both ω_1 and ω_2 equal to $\frac{1}{2}$:

$$f(x_i) = \frac{1}{2} . \frac{1}{\sqrt{2\pi\sigma_1^2}} . e^{-} \frac{(x_i - \mu_1)^2}{2\sigma_1^2} + \frac{1}{2} . \frac{1}{\sqrt{2\pi\sigma_2^2}} . e^{-} \frac{(x_i - \mu_2)^2}{2\sigma_2^2}$$

Which the distance to the closest mean matters.

6 Kmeans Problem

6.1 definition. you have a bunch of points $x_1, x_2, \ldots, x_n$; the goal of *kmeans problem* is to find centers $\mu_1, \mu_2, \ldots, \mu_k$ subject to $\sum_{i=1}^{n} (x_i - \mu_{c(i)})^2$, which $c_{(i)}$ is the closest center to x_i, is minimized.



Points $1, 2, 3$ have assigned to center μ_2 and points $4, 5$ have assigned to μ_1. By changing μ_i's, the clustering would change.