CMPE 320/Spring 08/Project 3: Mixing Densities

Samir Chettri

April 21, 2008

Abstract

We have discussed many types of densities in class - Normal, Uniform, Cauchy, Chi–Square etc. In each of these cases, we have an individual density. In this project we will study what happens when we mix densities. In particular we will study the combination of Normal densities. A key problem in density estimation from a set of data is obtaining the mean and standard deviations and the proportions if we are given the number of Normals in the mixture/data.

Students will first understand and obtain derivations followed by actual mixture density calculations. They will hand in their program that does the calculations. The maximum grade a student can get on this project is one hundred points.

Keywords: Normal density, mixtures of densities, probability and statistics, MATLAB.

Contents

1 Background on mixtures

Consider $\mathbf{X}$ and $\mathbf{Y}$, both independent normal RV's with means $\mu_{\mathbf{X}} = 50, \mu_{\mathbf{Y}} = 150$ and standard deviations $\sigma_{\mathbf{X}} = \sigma_{\mathbf{Y}} = 10$. The term "mixture distribution" is used in probability and statistics in the following ways:

1. Consider the RV $\mathbf{A} = 0.5(\mathbf{X} + \mathbf{Y})$. The meaning implied here is that a realization of $\mathbf{X}$ and an independent realization of $\mathbf{Y}$ are produced and half of each value is added to produce $\mathbf{A}$. In this sense, mixture means that $\mathbf{A}$ is a 50/50 combination of $\mathbf{X}$ and $\mathbf{Y}$. In the general case this is written as (p refers to the proportion)
$$\mathbf{A} = p\mathbf{X} + (1-p)\mathbf{Y} \tag{1}$$
2. Now consider an RV $\mathbf{B}$ that is stochastically chosen from $\mathbf{X}$ with probability p and and from $\mathbf{Y}$ with $1-p$. Therefore to produce a single RV $\mathbf{B}$, we first "flip a biased coin," that has a probability p of heads showing up and $1-p$ of tails. (For the case discussed in the previous item, $p = 0.5$.) If we get a heads, we generate a normal RV from $\mathbf{X}$ otherwise (we got tails) and so we generate a normal RV from $\mathbf{Y}$. The manner in which this is represented is
$$f_{\mathbf{B}}(b) = pf_{\mathbf{X}}(x) + (1-p)f_{\mathbf{Y}}(y). \tag{2}$$
For the specific case under discussion $f_{\mathbf{X}}$ and $f_{\mathbf{Y}}$ would be normal distributions.

1. Using the specific values of means and standard deviations given, sketch the densities $\mathbf{A}$ and $\mathbf{B}$. Is $\mathbf{A}$ normally distributed? Is $\mathbf{B}$? Indicate means and variances for both.

2. Using general (i.e., non–specific) values for $\mu_{\mathbf{X}}, \mu_{\mathbf{Y}}$ and $\sigma_{\mathbf{X}}, \sigma_{\mathbf{Y}}$ obtain the means and variances of $\mathbf{A}$ and $\mathbf{B}$ respectively.

2 Separating stochastically mixed distributions

Consider the stochastically generated mixture distribution, i.e.,

$$f_{\mathbf{B}}(b) = p f_{\mathbf{X}}(x) + (1-p) f_{\mathbf{Y}}(y). \tag{3}$$

We are going to generalize this to multi–variate normal random variables and more than two component densities.

Start with a density that is a linear combination of component densities $f(\mathbf{x}|j)$. From now on, I am going to drop the subscript $\mathbf{X}$ with the intention of typing less. This mixture is

$$f(\mathbf{x}) = \sum_{j=1}^{M} f(\mathbf{x}|j) P(j). \tag{4}$$

Since this is a mixture, $0 \leq P(j) \leq 1$ and

$$\sum_{j=1}^{M} P(j) = 1. \tag{5}$$

Using Bayes' Theorem (consider $P(j)$ to be the prior) we can write

$$P(j|\mathbf{x}) = \frac{f(\mathbf{x}|j)P(j)}{f(\mathbf{x})}, \tag{6}$$

with $f(\mathbf{x})$ from (4). Since $P(j|\mathbf{x})$ represents a probability,

$$\sum_{j=1}^{M} P(j|\mathbf{x}) = 1. \tag{7}$$

To simplify our problem, we are going to assume that

$$f(\mathbf{x}|j) = \frac{1}{(2\pi\sigma_j^2)^{d/2}} \exp\left\{ -\frac{(\mathbf{x}-\mu_j)^T(\mathbf{x}-\mu_j)}{2\sigma_j^2} \right\}, \tag{8}$$

that is, we have a multivariate (d–dimensional) gaussian with a covariance matrix that has entries σ_j^2 along the main diagonal and 0 everywhere else. Note that μ_j is a vector mean for the j^{th} normal.

An algorithm to obtain the parameters of the mixture density (4) is

$$\mu_j^{\text{new}} = \frac{\sum_{n=1}^{N} P^{\text{old}}(j|\mathbf{x}^n)\mathbf{x}^n}{P^{\text{old}}(j|\mathbf{x}^n)} \tag{9}$$

and

$$(\sigma_j^{\text{new}})^2 = \frac{1}{d} \frac{\sum_{n=1}^{N} P^{\text{old}}(j|\mathbf{x}^n)\mathbf{s}^n}{\sum_{n=1}^{N} P^{\text{old}}(j|\mathbf{x}^n)} \tag{10}$$

with $\mathbf{s}^n = (\mathbf{x}^n - \mu_j^{\text{new}})^T(\mathbf{x}^n - \mu_j^{\text{new}})$, and

$$P(j)^{\text{new}} = \frac{1}{N} \sum_{n=1}^{N} P^{\text{old}}(j|\mathbf{x}^n) \tag{11}$$

Notice where the "new" and "old" superscripts appear. Also note, when we say $\mathbf{x}^n$, we are not referring to the n^{th} power of $\mathbf{x}$, rather we are referring to the n^{th} input data vector $\mathbf{x}$.

Equations (4),(6),(8) and (9),(10),(11) form the backbone of an iterative algorithm to obtain the unknowns in a mixture distribution.

The student should encode this algorithm in MATLAB.

3 What to do?

This project has two parts. The first part deals with performing the relatively simple derivations needed to understand the basic theory while the latter one deals with programming in MATLAB.

Derivations and Theory - 30%.

1. Answer the questions asked in the section entitled **Background on mixtures**.
2. **Due date**: **28** April 2008 (no later than 11:59 pm). Your submissions should be neatly written (or typed). Make all assumptions clear to the reader and show all steps in derivations.

MATLAB - 70%

1. Write a MATLAB program called `EMD` that obtains the parameters of the mixture distribution, i.e., mean vectors and standard deviations and mixture fractions. You can create test data on your own. I will also generate test data for you so your program should be able to read the data I provide. The data format will be `Number of dimensions`

   `Number of densities in the mixture`

   `data vector 1`

   `data vector 2`

   `etc.`
2. **Due date**: 12 May 2008 by email only.

Remember, points will be taken off for not following the format described above, regardless of whether you get the right results or not. Late submissions will not be accepted and will be given a grade of zero points.