EM Algorithm and Semi-Supervised Learning

WS 2015/2016

In this tutorial, you will experiment with the EM algorithm and get familiar with semi-supervised learning.

You are a teacher at a grammar school. You have just explained the notion of "arithmetic mean" and gave your pupils a homework to measure the average height of boys and girls in the school. They carried out the experiment, recorded everyone's height, but they were a bit sloppy and unfortunately forgot to note down the person's gender. They asked you for help.

Unless you wish to visit all lectures at our university with a measuring tape (feel free to do so!), create artificial data. Generate 200 samples from mixture of two normal distributions using the Matlab function `randn` and parameters $\mu_1 = 152\,\text{cm}$, $\sigma_1 = 9\,\text{cm}$, $\mu_2 = 178\,\text{cm}$, $\sigma_2 = 10\,\text{cm}$.[1] Assume there are as many boys as girls.

1 EM algorithm

1. First, we will simplify the task and estimate the height of *all* people in the school.

   Display empirical probability density (using `bar(centers, bins)`).[2]

   Function `hist` returns histogram with counts instead of relative frequencies. After normalisation you get empirical probability function. To get empirical density you should divide the empirical probability function by the width of the histogram's bins. Why?

   Display theoretical probability density in the same figure using matlab function `normpdf`.

   Estimate parameters of the normal distribution from the generated data and display corresponding probability density (using function `normpdf`) in the same figure with empirical and theoretical probability densities. Expected result is shown in figure 1.

2. In case of two different normal distributions (Gaussian Mixture Model) it is not possible to calculate mean and standard deviation for each component of the mixture as the empirical mean and empirical standard deviation from the whole dataset. You need to split the observed data into the particular normal distributions, each with own mean and standard deviation. One option how to get these parameters for particular normal deviations is to use EM algorithm.

   Find means and standard deviations for each component of the mixture using function "EM.m". Display empirical, theoretical and estimated probability density for

[1] Source: http://www.usablestats.com/lessons/normal

[2] Be careful about Matlab syntax: `[bins centers] = hist(x)`

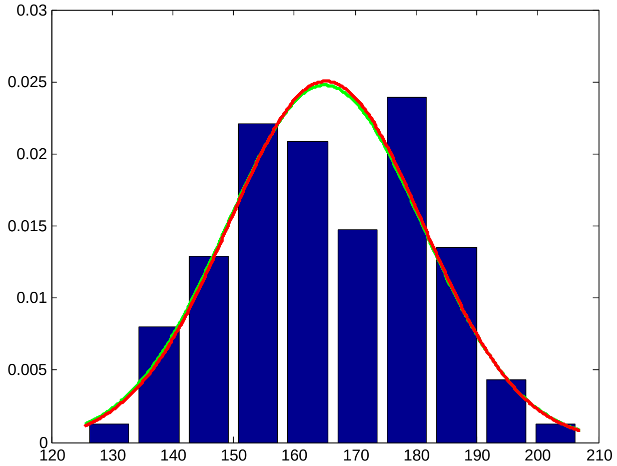

Figure 1: Expected result from task 1. Theoretical probability density is shown in red colour, estimated probability density is shown in green colour.

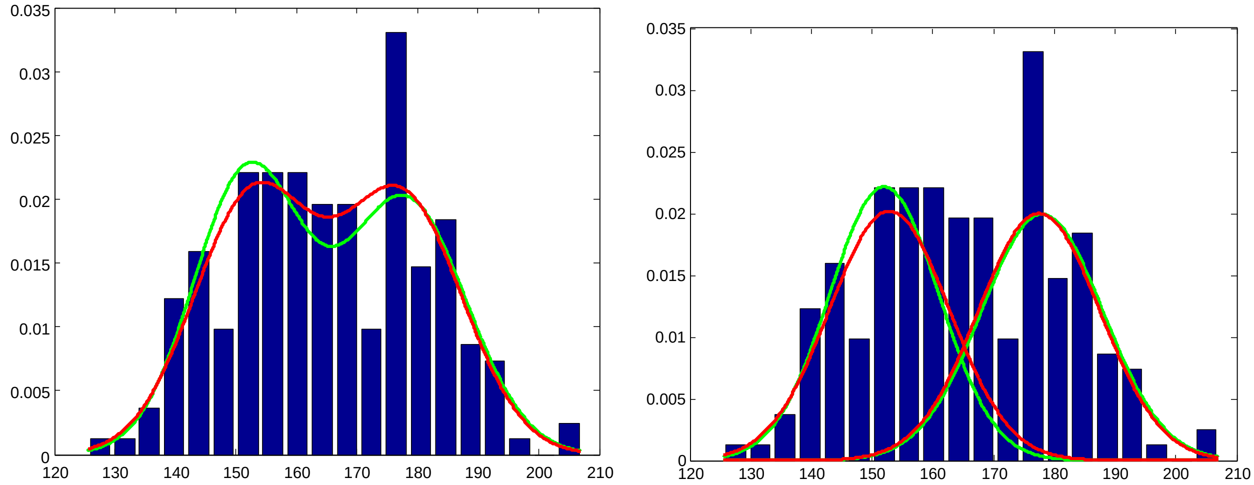

Figure 2: Expected results from task 2. Theoretical probability density os shown in red colour, estimated probability density is shown in green colour.

the mixture of normal distributions in one figure. (Expected result is shown in the left panel of figure 2.)

Display empirical, theoretical and estimated probability density for each component of the mixture in another figure. (Expected result is shown in the right panel of figure 2.)

2 Semi-Supervised Learning

3. Modify EM algorithm for semi-supervised clustering. Add a new input parameter *classification* to the function EM, which will contain assignment of some samples to clusters (1,2,...). Use 0 for samples for which you do not have assignment. In EM algorithm, fix assignment to clusters for those samples, for which the assignment is known from the input.

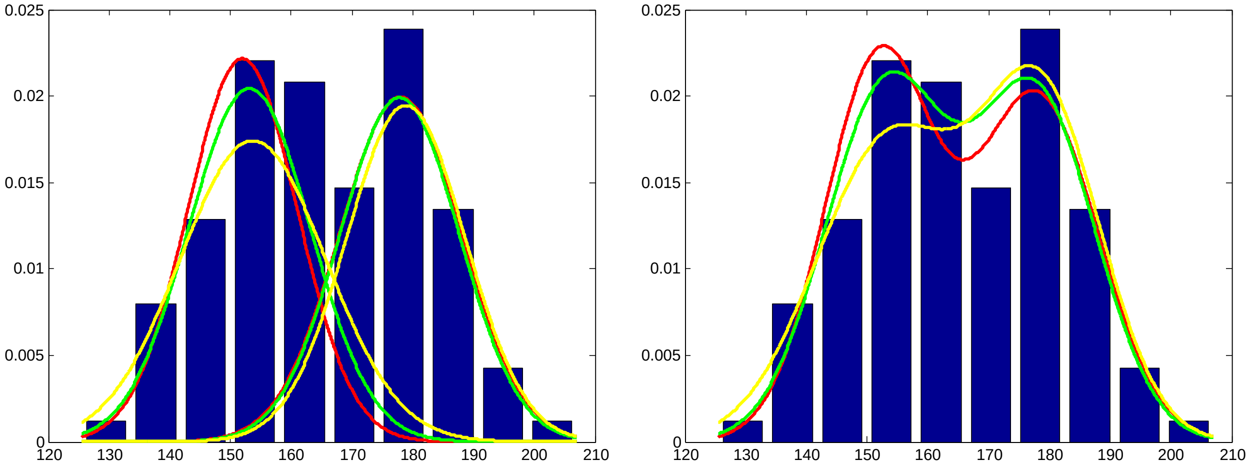

Figure 3: Expected result from task 3. Theoretical probability density is shown in red colour, probability density estimated using EM algorithm is shown in green colour, probability density estimated only from annotated samples is shown in yellow colour.

Perform the following experiment. Modify generation of data - add classification, i.e. information about membership of samples to clusters. Divide your data into two subsets - training subset and testing subset. Estimate parameters of the mixture in two ways.

(a) Estimate parameters only from training samples, for which you know classification (here, you do not use EM algorithm) - use the information about classification class for 10%, 20% and 50% of samples.

(b) Estimate parameters from all training samples using modified EM algorithm.

Compare accuracy of sample assignment to clusters using these two methods.[3]

Expected result is shown in figure 3.

3 Theory questions

1. The provided EM algorithm uses the Gaussian distribution as the likelihood. What is the name of the prior distribution?

2. In Elementary Statistics[4], the statistical properties of men's and women's height is as follows: $\mu_1 = 162\,\text{cm}$, $\sigma_1 = 7\,\text{cm}$, $\mu_2 = 176\,\text{cm}$, $\sigma_2 = 7\,\text{cm}$.[5] Try estimating the average height from these parameters.

3. The EM algorithm with Gaussian probabilities finds a local optimum of the posterior probability. Does a local optimum always exists? In other words, does the algorithm always converge?

[3] Calculate accuracy as the fraction of correctly assigned samples for which you did not know the assignment at the beginning.

[4] *Elementary Statistics: Picturing the World* by Ron Larson and Elizabeth Farber; Prentice Hall 2005; ISBN: 0-13-148316-1; http://esminfo.prenhall.com/samplechps/larson/

[5] The data about the height of 20- to 29- year-old people originate from the American *National Center for Health Statistics*.