CS 5135/6035 Learning Probabilistic Models

Exercise Questions for Lecture 10: Latent Variables, Mixture Models, Expectation Maximization

Gowtham Atluri

2/18/2020

Questions

1. **Scenario:** PetalLength of 150 flowers that belong to one of the three different species (setosa, versicolor, and virginica) are available in the iris dataset (as part of the Julia's RDatasets package).

   Assumptions: Assume PetalLength (denoted as x) is only variable available for this data. Treat Species (denoted as z) as a latent variable. Assume that this data is a mixture of three Gaussian distributions. Assume that probabilities of selecting individual species components is available (0.34, 0.33, 0.33). Assume that the variance σ^2 of all the components is the same and is 0.54.

   Goal: Your goal is to estimate the means of the individual components. The following questions will guide you in writing the EM algorithm (with E and M Steps), implementing it and demonstrating that it has estimated the correct parameters. **[10*2.5 = 25 points]**

   a. Write the expression for the marginal probability $p(z_i)$.
   b. Write the expression for the conditional probability $p(x_i|z_i)$.
   c. Write the expression for the probability density for one data point $p(x_i)$.
   d. Write the expression for the likelihood.
   e. Write the expression for the log-likelihood.
   f. Write the expressions for posterior probabilities $p(z_i = setosa|x_i)$, $p(z_i = versicolor|x_i)$, and $p(z_i = virginica|x_i)$.
   g. Derive the update equations for the three component means.
   h. Write the EM algorithm.
   i. Implement it in Julia.
   j. Plot the final estimated components on top of the histogram and comment on the goodness of the fit based on visual inspection.

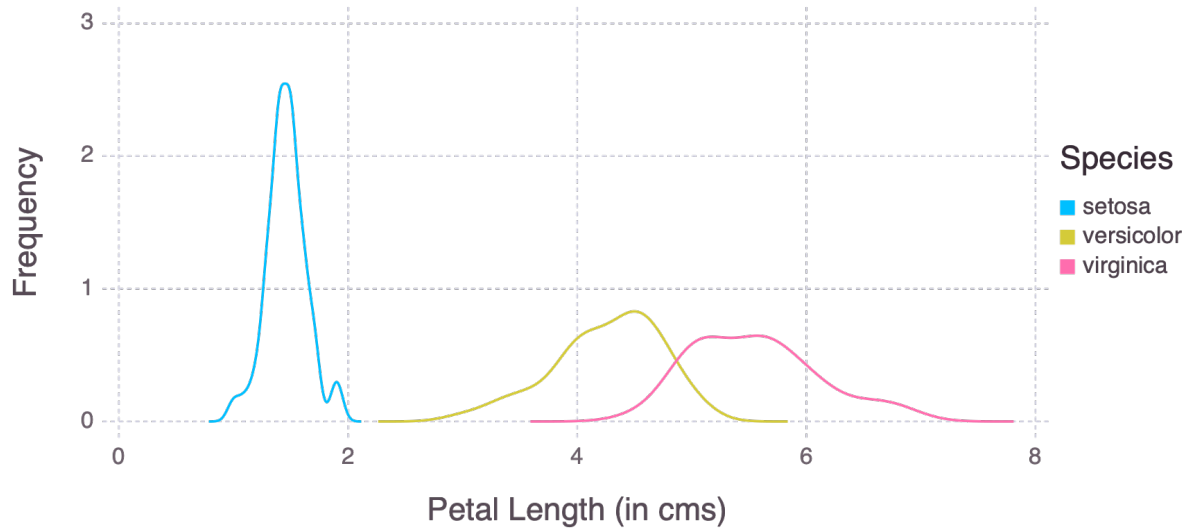
Bonus questions

1. Answer the above question by not assuming that the variance shared by the components is known. That is, the variance σ^2 needs to be estimated.
2. Answer the above question by not assuming that the variance is shared by the components. That is, each component will have its own variance. So, σ^2_{setosa}, $\sigma^2_{versicolor}$, and $\sigma^2_{virginica}$ needs to be estimated.
3. Answer the above question by considering PetalLength and PetalWidth as observed variables. That is, $\boldsymbol{x}$ needs to be modeled as a mixture of bivariate Gaussian distributions, unlike the above cases where it is modeled as a mixture of univariate Gaussians.

Sample code

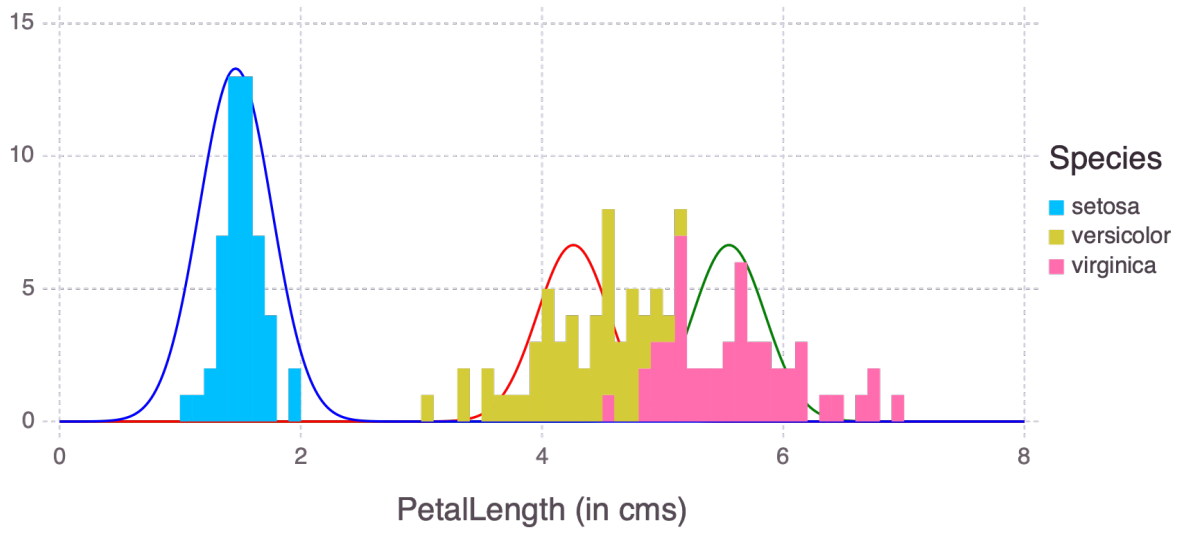
1. To plot the densities of the data

```
    using RDatasets, Gadfly, Distributions;
    data = dataset("datasets","iris");
    myplot = plot(data,x=:PetalLength,color=:Species,Geom.density,
                  Guide.xlabel("Petal Length (in cms)"),
                  Guide.ylabel("Frequency"), major_label_font_size=18pt,
                  minor_label_font_size=14pt,
                  key_title_font_size = 18pt,
                  key_label_font_size = 14pt,
                  major_label_color=colorant"black",
                  minor_label_color=colorant"black",
                  Coord.Cartesian(xmin=0, xmax=8));
    draw(PNG("./figs/iris_densities.png", 6inch, 3inch,dpi=300), myplot);
```



2. Sample code for plotting histogram and EM results

```
myplot = plot(layer(data,x=:PetalLength,color=:Species,Geom.histogram,
                    Theme(default_color=colorant"purple")),
        layer(x=0:0.02:8,y=pdf.(Normal(1.46,0.3),0:0.02:8)*10,Geom.line,
                    Theme(default_color=colorant"blue")),
        layer(x=0:0.02:8,y=pdf.(Normal(4.26,0.3),0:0.02:8)*5,Geom.line,
                    Theme(default_color=colorant"red")),
        layer(x=0:0.02:8,y=pdf.(Normal(5.55,0.3),0:0.02:8)*5,Geom.line,
                    Theme(default_color=colorant"green")),
        Guide.xlabel("PetalLength (in cms)"),Guide.ylabel(""), major_label_font_size=18pt,
        minor_label_font_size=14pt,
        key_title_font_size = 18pt,
        key_label_font_size = 14pt,
        major_label_color=colorant"black",
        minor_label_color=colorant"black",Coord.Cartesian(xmin=0, xmax=8));
draw(PNG("./figs/iris_em.png", 6inch, 3inch,dpi=300), myplot);
```



3. EM code

```
function E_step(x,mu_M,mu_F,sigma,p)
    numerator = p*pdf.(Normal(mu_M,sigma),x)
    denom = numerator .+ (1-p).* pdf.(Normal(mu_F,sigma),x);
    post_x = numerator ./denom;
    return post_x;
end
```

```
## E_step (generic function with 1 method)
```

```
function M_step(x,post_x)
    mu_M = (post_x'*x)./sum(post_x);
    mu_F = ((1.-post_x)'*x)./sum((1.-post_x));
    return mu_M, mu_F;
end
```

```
## M_step (generic function with 1 method)
```

```
function EM(x,mu_M,mu_F,p,sigma)
    maxIter = 1000;
    for i=1:maxIter
        print(i,"\n");
        post_x = E_step(x,mu_M,mu_F,sigma,p); #print(post_x,"\n");
        mu_M_new, mu_F_new = M_step(x,post_x); print(mu_M_new," ",mu_F_new,"\n");
        if(abs(mu_M-mu_M_new)<0.001 && abs(mu_F-mu_F_new)<0.001)
            break;
        end;
        mu_M = mu_M_new;
        mu_F = mu_F_new;
    end
    return mu_M, mu_F;
end
```

```
## EM (generic function with 1 method)
```

```
data = dataset("car","Davis");
x = data[:Height];
mu_M=190;
mu_F=150;
p = 0.5;
sigma=7;
EM(x,mu_M,mu_F,p,sigma)
```

```
## (176.38366334360896, 163.46901164793636)
```