3. EM Algorithm

Extension Note: Homework 4 due date has been extended by 1 day to **August 17 23:59UTC** .

Consider the following mixture of two Gaussians:

$$p(x;\theta) = \pi_1 \mathcal{N}(x;\mu_1,\sigma_1^2) + \pi_2 \mathcal{N}(x;\mu_2,\sigma_2^2)$$

This mixture has parameters $\theta = \{\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2\}$. They correspond to the mixing proportions, means, and variances of each Gaussian. We initialize θ as $\theta_0 = \{0.5, 0.5, 6, 7, 1, 4\}$.

We have a dataset $\mathcal{D}$ with the following samples of x: $x^{(0)} = -1$, $x^{(1)} = 0$, $x^{(2)} = 4$, $x^{(3)} = 5$, $x^{(4)} = 6$.

We want to set our parameters θ such that the data log-likelihood $l(\mathcal{D};\theta)$ is maximized:

$$\underset{\theta}{\operatorname{argmax}} \sum_{i=0}^{4} \log p(x^{(i)};\theta).$$

Recall that we can do this with the EM algorithm. The algorithm optimizes a lower bound on the log-likelihood, thus iteratively pushing the data likelihood upwards. The iterative algorithm is specified by two steps applied successively:

1. E-step: infer component assignments from current $\theta_0 = \theta$ (complete the data)

$$p(y=k \mid x^{(i)}) := p(y=k \mid x^{(i)};\theta_0), \text{ for } k=1,2, \text{ and } i=0,\dots,4.$$

2. M-step: maximize the expected log-likelihood

$$\tilde{l}(D;\theta) := \sum_i \sum_k p(y=k \mid x^{(i)}) \log \frac{p(x^{(i)}, y=k;\theta)}{p(y=k \mid x^{(i)})}$$

with respect to θ while keeping $p(y=k \mid x^{(i)})$ fixed.

To see why this optimizes a lower bound, consider the following inequality:

$$\begin{aligned}
\log p(x;\theta) &= \log \sum_y p(x,y;\theta) \\
&= \log \sum_y q(y|x) \frac{p(x,y;\theta)}{q(y|x)} \\
&= \log \mathbb{E}_{y\sim q(y|x)} \left[\frac{p(x,y;\theta)}{q(y|x)}\right] \\
&\geq \mathbb{E}_{y\sim q(y|x)} \left[\log \frac{p(x,y;\theta)}{q(y|x)}\right] \\
&= \sum_y q(y|x) \log \frac{p(x,y;\theta)}{q(y|x)}
\end{aligned}$$

where the inequality comes from **Jensen's inequality** . EM makes this bound tight for the current setting of θ by setting $q(y|x)$ to be $p(y \mid x; \theta_0)$.

Note: If you have taken *6.431x Probability–The Science of Uncertainty*, you could review the video in *Unit 8: Limit Theorems and Classical Statistics, Additional Theoretical Material, 2. Jensen's Inequality*.

Likelihood Function

1/1 point (graded)
What is the log-likelihood of the data $l(\mathcal{D}; \theta)$ given the initial setting of θ? Please round to the nearest tenth.

Note: You will want to write a script to calculate this, using the natural log (np.log) and np.float64 data types.

-24.512532330086678 ✔ **Answer:** -24.5

Solution:

The likelihood can be written as:

$$\begin{aligned} P(\mathcal{D}; \theta) &= \prod_{i=0}^{4} p(x; \theta) \\ &= \prod_{i=0}^{4} \pi_1 \mathcal{N}(x^{(i)}; \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x^{(i)}; \mu_2, \sigma_2^2) \end{aligned}$$

Taking the log gives:

$$l(\mathcal{D}; \theta) = \sum_{i=0}^{4} \log(\pi_1 \mathcal{N}(x^{(i)}; \mu_1, \sigma_1^2) + \pi_2 \mathcal{N}(x^{(i)}; \mu_2, \sigma_2^2))$$

We then evaluate each Gaussian using the standard formulation:

$$\mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Submit You have used 1 of 3 attempts

Answers are displayed within the problem

E-Step

1/1 point (graded)
What is the formula for $p(y = k \mid x, \theta)$? Write in terms of π_k, π_1, π_2, N_k, N_1, and N_2 (where $N_k = \mathcal{N}(x \mid \mu_k, \sigma_k^2)$).

pi_k*N_k/(pi_1*N_1+ pi_: ✔ **Answer:** (pi_k * N_k) / (pi_1 * N_1 + pi_2 * N_2)

$\frac{\pi_k \cdot N_k}{\pi_1 \cdot N_1 + \pi_2 \cdot N_2}$

STANDARD NOTATION

Solution:

Following Bayes Rule we have:

$$p(y \mid x) = \frac{p(y)\, p(x \mid y)}{\sum_{y'} p(y')\, p(x|y')}$$

For this problem, this equates to:

$$p(y = k \mid x; \theta) = \frac{\pi_k \mathcal{N}(x; \mu_y, \sigma_y^2)}{\sum_{i=1}^{2} \pi_i \mathcal{N}(x; \mu_i, \sigma_i^2)}$$

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

E-Step Weights

5/5 points (graded)

For each of the given data points say which Gaussian (1 or 2) they are given more weight towards in the first E-step using the given setting of θ_0. This is, answer 2 if $p(y = 2 \mid x, \theta_0) > p(y = 1 \mid x, \theta_0)$and 1 otherwise.

$x^{(0)}$: 2 Answer: 2

$x^{(1)}$: 2 Answer: 2

$x^{(2)}$: 2 Answer: 2

$x^{(3)}$: 1 Answer: 1

$x^{(4)}$: 1 Answer: 1

Solution:

Note that x will more likely be assigned to Gaussian 2 ($y = 2$) instead of Gaussian 1 ($y = 1$) when the following is true:

$$\begin{aligned}
\frac{P(y=2|x^{(i)}, \theta_0)}{P(y=1|x^{(i)}, \theta_0)} &> 1 \\
\Leftrightarrow \frac{P(x^{(i)}|y=2)\, P(y=2)}{P(x^{(i)}|y=1)\, P(y=1)} &> 1 \\
\Leftrightarrow \frac{\frac{1}{\sqrt{(2\pi\sigma_2^2)}} exp\{-\frac{1}{2}(x-\mu_2)^2/\sigma_2^2\}}{\frac{1}{\sqrt{(2\pi\sigma_1^2)}} exp\{-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2\}} &> 1 \\
\Leftrightarrow \frac{\frac{1}{\sqrt{(2\pi \times 4)}} exp\{-\frac{1}{2}(x-7)^2/4\}}{\frac{1}{\sqrt{(2\pi \times 1)}} exp\{-\frac{1}{2}(x-6)^2\}} &> 1 \\
\Leftrightarrow \frac{1}{2} exp\{-\frac{1}{2}((x-7)^2/4 - (x-6)^2)\} &> 1 \\
\Leftrightarrow \frac{1}{2} exp\{\frac{1}{8}(x-5)(3x-19)\} &> 1 \\
\Leftrightarrow \log(\frac{1}{2}) + \frac{1}{8}(x-5)(3x-19) &> 0
\end{aligned}$$

The x-intercepts of this parabola are $x_1 \approx 4.1525, x_2 \approx 7.1809$. Thus, we can see that all points $x \in [4.15, 7.18]$ have higher probability under class $y = 1$, and all other points have higher probability under $y = 2$. Thus, $x^{(0)}$, $x^{(1)}$, and $x^{(2)}$ are more likely (but not entirely) assigned to Gaussian 2, and the rest of the points ($x^{(3)}$, $x^{(4)}$) are more likely (but not entirely) assigned to Gaussian 1.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

M-Step

3/3 points (graded)

Fixing $p(y=k \mid x, \theta_0)$, we want to update θ such that our lower bound is maximized.

What is the optimal $\hat{\mu}_k$? Answer in terms of $x^{(1)}$, $x^{(2)}$, and γ_{k1}, γ_{k2}, which are defined to be $\gamma_{ki} = p(y=k \mid x^{(i)}; \theta_0)$

(For ease of input, use subscripts instead superscripts, i.e. type `x_i` for $x^{(i)}$. Type `gamma_ki` for γ_{ki}.)

(gamma_k1*x_1 + gamma_k2*x_2)/(gamma_k1+ gamma_k2) ✔

Answer: (gamma_k1 * x_1 + gamma_k2 * x_2) / (gamma_k1 + gamma_k2)

$$\frac{\gamma_{k1} \cdot x_1 + \gamma_{k2} \cdot x_2}{\gamma_{k1} + \gamma_{k2}}$$

What is the optimal $\hat{\sigma}_k^2$? Answer in terms of $x^{(1)}$, $x^{(2)}$, γ_{k1} and γ_{k2}, which are defined as above to be $\gamma_{ki} = p(y=k \mid x^{(i)}; \theta_0)$, and $\hat{\mu}_k$.

(Type `hatmu_k` for $\hat{\mu}_k$. As above, for ease of input, use subscripts instead superscripts, i.e. type `x_i` for $x^{(i)}$. Type `gamma_ki` for γ_{ki}.)

(gamma_k1*(x_1 - hatmu_k)^2 + gamma_k2*(x_2-hatmu_k)^2)/(gamma_k1+ ga ✔

Answer: (gamma_k1 * (x_1 - hatmu_k)^2 + gamma_k2 * (x_2 - hatmu_k)^2) / (gamma_k1 + gamma_k2)

$$\frac{\gamma_{k1} \cdot (x_1 - hatmu_k)^2 + \gamma_{k2} \cdot (x_2 - hatmu_k)^2}{\gamma_{k1} + \gamma_{k2}}$$

What is the optimal $\hat{\pi}_k$? Answer in terms of γ_{k1} and γ_{k2}, which are defined as above to be $\gamma_{ki} = p(y=k \mid x^{(i)}; \theta_0)$.

(As above, type `gamma_ki` for γ_{ki}.)

Note: that you must account for the constraint that $\pi_1 + \pi_2 = 1$ where $\pi_1, \pi_2 \geq 0$.

Note: If you know that some aspect of your formula equals an exact constant, simplify and use this number, i.e. $\gamma_{11} + \gamma_{21} = 1$.

(gamma_k1+ gamma_k2)/2 ✔ **Answer:** (gamma_k1 + gamma_k2) / 2

$$\frac{\gamma_{k1} + \gamma_{k2}}{2}$$

STANDARD NOTATION

Solution:

The function we are optimizing is now:

$$\sum_i \sum_k \gamma_{ki} \log\left(\pi_k \mathcal{N}(x^{(i)}; \mu_k, \sigma_k^2)\right)$$

Taking $\frac{\partial}{\partial \mu_k}$ and setting to 0 gives:

$$\begin{aligned}
\frac{\partial}{\partial \mu_k} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k \mathcal{N}(x^{(i)}; \mu_k, \sigma_k^2)\right) &= \sum_i \gamma_{ki} \frac{\partial}{\partial \mu_k} \log\left(\pi_k \mathcal{N}(x^{(i)}; \mu_k, \sigma_k^2)\right) \\
&= \sum_i \gamma_{ki} \frac{\partial}{\partial \mu_k} \left(\log\left(\frac{1}{\sqrt{2\pi\sigma_k^2}}\right) - \frac{(x^{(i)} - \mu_k)^2}{2\sigma_k^2}\right) \\
&= \sum_i \gamma_{ki} \frac{x^{(i)} - \mu_k}{\sigma_k^2} = 0
\end{aligned}$$

Separating out μ_k gives:

$$\mu_k = \frac{\sum_i \gamma_{ki} x^{(i)}}{\sum_i \gamma_{ki}}$$

We can interpret this as a weighted average of the data points, normalized by the "total mass" assigned to Gaussian k. The weight is the probability that point $x^{(i)}$ "belongs" to Gaussian k.

Solving for σ_k^2 is similar:

$$\begin{aligned}
\frac{\partial}{\partial \sigma_k^2} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k \mathcal{N}\left(x^{(i)}; \mu_k, \sigma_k^2\right)\right) &= \sum_i \gamma_{ki} \frac{\partial}{\partial \sigma_k^2} \log\left(\pi_k \mathcal{N}\left(x^{(i)}; \mu_k, \sigma_k^2\right)\right) \\
&= \sum_i \gamma_{ki} \frac{\partial}{\partial \sigma_k^2} (\log(\frac{1}{\sqrt{2\pi\sigma_k^2}}) - \frac{\left(x^{(i)} - \mu_k\right)^2}{2\sigma_k^2}) \\
&= \sum_i \gamma_{ki} \left(-\frac{1}{2\sigma_k^2} + \frac{\left(x^{(i)} - \mu_k\right)^2}{2\sigma_k^4}\right) = 0
\end{aligned}$$

Separating out σ_k^2 gives:

$$\sigma_k^2 = \frac{\sum_i \gamma_{ki} \left(x^{(i)} - \mu_k\right)^2}{\sum_i \gamma_{ki}}$$

Finally we solve for π_k while including a lagrange multiplier for the constraint that $\sum_k \pi_k = 1$.

$$\begin{aligned}
\frac{\partial}{\partial \pi_k} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k \mathcal{N}\left(x^{(i)}; \mu_k, \sigma_k^2\right)\right) + \lambda(\sum_k \pi_k - 1) &= \sum_i \gamma_{ki} \frac{\partial}{\partial \pi_k} \log(\pi_k) + \frac{\partial}{\partial \pi_k} \lambda(\sum_k \pi_k - 1) \\
&= \frac{\sum_i \gamma_{ki}}{\pi_k} + \lambda = 0
\end{aligned}$$

Giving $\pi_k = -\frac{\sum_i \gamma_{ki}}{\lambda}$.

Solving for λ gives:

$$\frac{\partial}{\partial \lambda} \sum_i \sum_k \gamma_{ki} \log\left(\pi_k \mathcal{N}\left(x^{(i)}; \mu_k, \sigma_k^2\right)\right) + \lambda(\sum_k \pi_k - 1) = \sum_k \pi_k - 1 = 0$$

Combining the two gives:

$$\lambda = -\sum_i \sum_k \gamma_{ki}$$

which we recognize as N, the total number of points. Thus $\hat{\pi}_k$ is $\frac{\sum_i \gamma_{ki}}{N}$.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Training 1

1/1 point (graded)

In the first M-step, which Gaussian will shift to the left more (relatively)?

- [ ] Gaussian 1
- [x] Gaussian 2 ✔

Solution:

Intuitively, Gaussian 2 is influenced most by the points $x^{(0)}$,$x^{(1)}$, and so it will move to the left. Gaussian 1 will be more influenced by the points at $x^{(2)}$, $x^{(3)}$ and $x^{(4)}$ and so it will not move very much to the left. If we computed the actual values, we would see that the updated means for the two Gaussians are approximately $\mu_1 = 5.1317$ and $\mu_2 = 1.4710$.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Training 2

1/1 point (graded)
In the first M-step, which Gaussian's variance will increase more (relatively)?

- [ ] Gaussian 1
- [x] Gaussian 2 ✔

Solution:

Intuitively, the variance of Gaussian 2 spreads out to cover points $x^{(0)}$ and $x^{(1)}$ which it is most influenced by. The 3 points which most influence Gaussian 1 are concentrated around its mean, we would not expect the variance to increase. Numerically, σ_1 decreases to approximately 0.7846 while σ_2 increases to 2.6395.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Training 3

0/1 point (graded)
After convergence, which variance will be larger?

- [ ] σ_1^2 ✔
- [x] σ_2^2 ✘

Solution:

Gaussian 1 will be centered around the cluster of 3 points on the right, while Gaussian 2 will be centered around the 2 points on the left. Gaussian 1 will have larger variance because of the larger spread of the right cluster.

I'm getting

```
mu=  [5.03812884 0.69307411], sigma2=  [0.6199125  5.32681503]
```

Not sure what I've done wrong. Could anyone have a look at my code?

Submit You have used 1 of 1 attempt

I got the different results : pi = [0.6 0.4] mu= [4.99999998 -0.5], sigma= [0.81649664 0.5]

Pay attention to your sigma, don't use sigma squared as sigma in your repeating calculation. When I made this mistake, I got the same results as yours.

posted 11 days ago by m_zhang0521

Answers are displayed within the problem