CS 446/ECE 449: Machine Learning

A. G. Schwing

University of Illinois at Urbana-Champaign, 2020

L22: Gaussian Mixture Models

Goals of this lecture

- Understanding Gaussian mixture models
- Getting to know more details about generative modeling
- Learning the relationship between Gaussian mixture models and kMeans

Reading material:

- C. Bishop; Pattern Recognition and Machine Learning; Chapter 9.2
- K. Murphy; Machine Learning: A Probabilistic Perspective; Chapter 11

Recall: Linear regression (discriminative)

$$p(y^{(i)}|x^{(i)}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y^{(i)} - \boldsymbol{w}^\top\phi(x^{(i)}))^2\right)$$

Now: (generative)

$$p(x^{(i)}|\underbrace{\mu,\sigma}_{\theta \text{ or } \boldsymbol{w}}) = \mathcal{N}(x^{(i)}|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x^{(i)} - \mu)^2\right)$$

Important difference: we are now interested in modeling the distribution of the data $x^{(i)}$ and not the class labels $y^{(i)}$. Though it is sometimes ambiguous what you call data or labels.

Given a dataset $\mathcal{D} = \{(x^{(i)})\}$ how to find $\theta = (\mu, \sigma)$ of

$$p(x^{(i)}|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x^{(i)} - \mu)^2\right)$$

Minimize negative log-likelihood

Program:

$$\min_{\mu,\sigma} -\log \prod_{i\in\mathcal{D}} p(x^{(i)}|\mu, \sigma) := \sum_{i\in\mathcal{D}} \frac{1}{2\sigma^2}(x^{(i)} - \mu)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

Program:

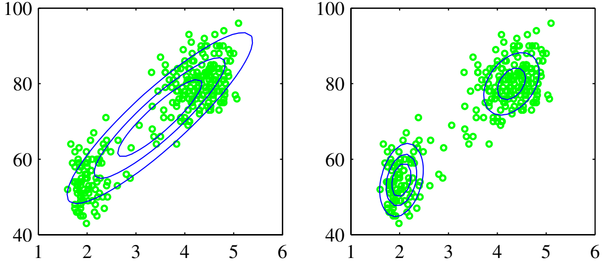
$$\min_{\mu,\sigma} -\log \prod_{i\in\mathcal{D}} p(x^{(i)}|\mu,\sigma) := \sum_{i\in\mathcal{D}} \frac{1}{2\sigma^2}(x^{(i)}-\mu)^2 + \frac{N}{2}\log(2\pi\sigma^2)$$

Optimality condition:

$$\frac{\partial}{\partial\mu}: \quad \frac{1}{\sigma^2}\sum_{i\in\mathcal{D}}(x^{(i)}-\mu) = 0 \quad \Longrightarrow \quad \mu = \frac{1}{N}\sum_{i\in\mathcal{D}} x^{(i)}$$

$$\frac{\partial}{\partial\sigma}: \quad \frac{-1}{\sigma^3}\sum_{i\in\mathcal{D}}(x^{(i)}-\mu)^2 + \frac{N}{\sigma} = 0 \quad \Longrightarrow \quad \sigma^2 = \frac{1}{N}\sum_{i\in\mathcal{D}}(x^{(i)}-\mu)^2$$

Issue: single Gaussian isn't that flexible


100
80
60
40
1
2
3
4
5
6
100
80
60
40
1
2
3
4
5
6

Fix: linear superposition of Gaussians

$$p(x^{(i)}|\underbrace{\pi,\mu,\sigma}_{\text{all components}}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)$$

Constraints:

$$\sum_{k=1}^{K} \pi_k = 1 \qquad \pi_k \geq 0$$

Minimize negative log-likelihood:

$$\min_{\pi,\mu,\sigma} -\log \prod_{i\in\mathcal{D}} p(x^{(i)}|\pi,\mu,\sigma) := -\sum_{i\in\mathcal{D}} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k)$$

How to optimize:

No closed form solution. Gradient descent is possible.

Alternative: auxiliary/latent variable $z_{ik} \in \{0, 1\}$ with $\sum_{k=1}^{K} z_{ik} = 1 \;\forall i$

Marginal for z_{ik}

$$p(z_{ik} = 1) = \pi_k \qquad p(\boldsymbol{z}_i) = \prod_{k=1}^{K} \pi_k^{z_{ik}} \text{ where } \boldsymbol{z}_i = [z_{i1}, \ldots, z_{iK}]^\top$$

Conditional

$$p(x^{(i)}|z_{ik} = 1) = \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)$$

Marginal for $x^{(i)}$

$$\begin{aligned} p(x^{(i)}|\pi, \mu, \sigma) &= \sum_{\boldsymbol{z}_i} p(x^{(i)}|\boldsymbol{z}_i)p(\boldsymbol{z}_i) = \sum_{\boldsymbol{z}_i} \prod_{k=1}^{K} \pi_k^{z_{ik}} \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)^{z_{ik}} \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k) \end{aligned}$$

Posterior:

$$r_{ik} = p(z_{ik} = 1|x^{(i)}) = \frac{p(z_{ik} = 1)p(x^{(i)}|z_{ik} = 1)}{\sum_{\hat{k}=1}^{K} p(z_{i\hat{k}} = 1)p(x^{(i)}|z_{i\hat{k}} = 1)} = \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^{(i)}|\mu_{\hat{k}}, \sigma_{\hat{k}})}$$

With all those definitions at hand, minimize negative log-likelihood:

$$\min_{\pi,\mu,\sigma} -\log \prod_{i\in\mathcal{D}} p(x^{(i)}|\pi,\mu,\sigma) := -\sum_{i\in\mathcal{D}} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x^{(i)}|\mu_k,\sigma_k) \;\; \text{s.t.} \; \sum_{k=1}^{K} \pi_k = 1$$

Stationary point: (per cluster weight $N_k = \sum_{i\in\mathcal{D}} r_{ik}$)

$$\frac{\partial}{\partial \mu_k}: \quad -\sum_{i\in\mathcal{D}} r_{ik}\left(-\frac{1}{2\sigma^2}(x^{(i)}-\mu_k)\right) = 0 \quad \Longrightarrow \quad \mu_k = \frac{1}{N_k}\sum_{i\in\mathcal{D}} r_{ik} x^{(i)}$$

$$\frac{\partial}{\partial \sigma_k}: \sum_{i\in\mathcal{D}} r_{ik}\left(\frac{1}{\sigma} - \frac{1}{\sigma^3}(x^{(i)}-\mu_k)^2\right) = 0 \Longrightarrow \sigma_k^2 = \frac{1}{N_k}\sum_{i\in\mathcal{D}} r_{ik}(x^{(i)}-\mu_k)^2$$

$\frac{\partial}{\partial \pi_k}$: with Lagrange multiplier: $\sum_{i\in\mathcal{D}} \frac{\mathcal{N}(x^{(i)}|\mu_k,\sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^{(i)}|\mu_{\hat{k}},\sigma_{\hat{k}})} + \lambda = 0$

multiplication with π_k and summation over k: $\lambda = -N$

multiplication with π_k and rearranging: $\pi_k = \frac{N_k}{N}$

Not a closed form solution

Gaussian Mixture Model Algorithm:

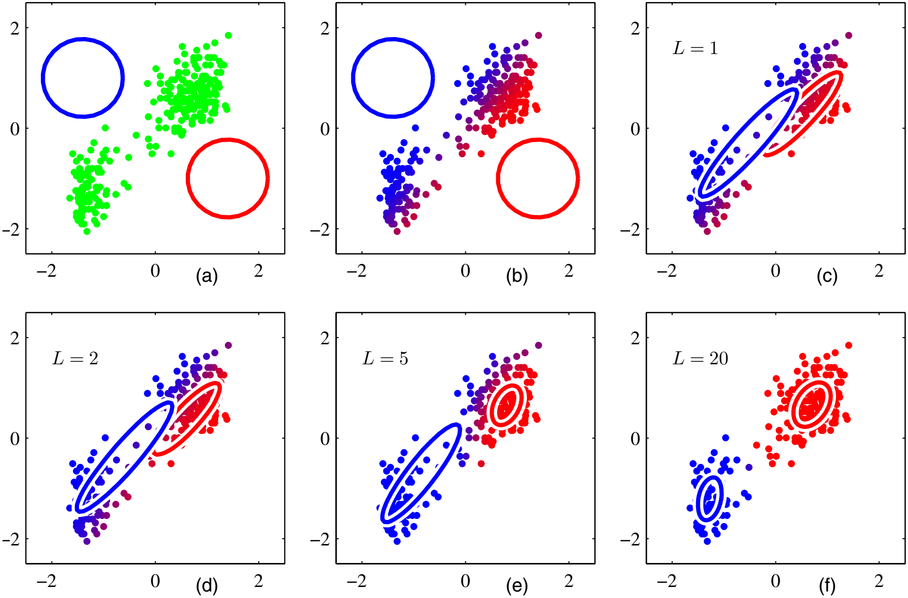
- Initialize μ, σ, π
- Iterate:
  - E-Step: Update

analytical form of subproblem,
converge much faster.
better than gradient descent

$$r_{ik} = \frac{\pi_k \mathcal{N}(x^{(i)}|\mu_k, \sigma_k)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \mathcal{N}(x^{(i)}|\mu_{\hat{k}}, \sigma_{\hat{k}})}$$

  - M-Step: Update

$$\begin{aligned} \mu_k &= \frac{1}{N_k} \sum_{i \in \mathcal{D}} r_{ik} x^{(i)} \\ \sigma_k^2 &= \frac{1}{N_k} \sum_{i \in \mathcal{D}} r_{ik} (x^{(i)} - \mu_k)^2 \\ \pi_k &= \frac{N_k}{N} \end{aligned}$$


2
0
−2
−2
0
(a)
2
(b)
L = 1
(c)
L = 2
(d)
L = 5
(e)
L = 20
(f)

Similarity to kMeans:

- r_{ik} is an assignment of sample i to cluster k, albeit a soft assignment
- μ_k are the cluster centers

Can we make this similarity formal?

Fix $\sigma_k^2 = \epsilon \; \forall k$

Responsibilities:

$$r_{ik} = \frac{\pi_k \exp(-\frac{1}{2\epsilon}(x^{(i)} - \mu_k)^2)}{\sum_{\hat{k}=1}^{K} \pi_{\hat{k}} \exp(-\frac{1}{2\epsilon}(x^{(i)} - \mu_{\hat{k}})^2)}$$

What happens for $\epsilon \to 0$?

- In the denominator the term for which $(x^{(i)} - \mu_{\hat{k}})^2$ is smallest goes to zero slowest
- All responsibilities will go to zero except the one for which $(x^{(i)} - \mu_{\hat{k}})^2$ is smallest, which will go to unity
- Responsibilities are hard assignments
- Cost function can be shown to be identical in the limit

Quiz:

- What is the maximum likelihood solution of fitting the mean and variance of a Gaussian?
- Why do we consider mixtures of Gaussians?
- How do we find the means, variances and responsibilities of the Gaussian mixture model?

Important topics of this lecture

- Generative modeling intuition
- Gaussian mixture model
- Relationship between Gaussian mixture model and kMeans

What's next

- Generalizing the Gaussian mixture model concept