Gaussian Mixture Models

Instructor: John Thickstun

Discussion Board: Available on Canvas

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

Logistics

- Lectures will be recorded and posted internally starting today.
- A better discussion board (Ed) will be up in the next couple days.
- TA: Sami Davies will be helping out with course infrastructure and moderating questions during lecture

Recap: Overfitting

- Given finite samples $x_1, \ldots, x_n \sim p$ from a continuous distribution $p(x)$.

- Estimate the probability of each element of $\mathcal{X}$ using the MLE?

$$\hat{p}(x) = \begin{cases} \frac{1}{n} & \text{if } x \in \{x_1, \ldots, x_n\}, \\ 0 & \text{otherwise.} \end{cases}$$

- Regularization: restrict our estimator to a parametric family

Maximum Likelihood Estimation

- Given a parametric family of probability distributions $\{p_\theta : \theta \in \Theta\}$

- Choose $\theta \in \Theta$ to maximize the likelihood of observations $x_1, \ldots, x_n$:

$$\sup_\theta \mathop{\mathbb{E}}_{x \sim p} \log p_\theta(x) \approx \sup_\theta \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i).$$

- KL-divergence minimization:

$$\mathop{\mathbb{E}}_{x \sim p} - \log p_\theta(x) = \mathop{\mathbb{E}}_{x \sim p} - \log \frac{p_\theta(x)}{p(x)} p(x) = H(p) + D(p \,\|\, p_\theta) \geq H(p).$$

Generalization

- Measure performance via likelihood of a test set $x_1^{\text{test}}, \ldots, x_m^{\text{test}}$:

$$\frac{1}{m}\sum_{i=1}^{m} \log p_\theta(x_i^{\text{test}}).$$

- Deep learning for classification: send training error to zero

- Deep learning for generative modeling: send log-likelihood to zero?

$$\mathop{\mathbb{E}}_{x\sim p} - \log p_\theta(x) \geq H(p), \qquad p_\theta(x) \approx \begin{cases} \frac{1}{n} & \text{if } x \in \{x_1, \ldots, x_n\}, \\ 0 & \text{otherwise.} \end{cases}$$

Gaussian Mixture Models

- K-Gaussian mixture model over data $x \in \mathbb{R}^d$.
- Each data point x belongs to a latent cluster $z \in \{1, \ldots, K\}$.
- Each cluster is drawn from a Gaussian with mean μ_k and variance Σ_k:

$$
\begin{aligned}
&1. \quad z \sim \text{Categorical}_\pi(K), && \pi \in \Delta^{K-1}, \\
&2. \quad x \sim \mathcal{N}(\mu_z, \Sigma_z), && \mu \in \mathbb{R}^{K \times d}, \Sigma \in \mathbb{R}^{K \times d \times d}.
\end{aligned}
$$

Gaussian Mixture Models

- Generative model:

$$
\begin{aligned}
&1. \quad z \sim \text{Categorical}_\pi(K), && \pi \in \Delta^{K-1}, \\
&2. \quad x \sim \mathcal{N}(\mu_z, \Sigma_z), && \mu \in \mathbb{R}^{K\times d}, \Sigma \in \mathbb{R}^{K\times d\times d}.
\end{aligned}
$$

- Likelihood:

$$
\begin{aligned}
p(x) &= \int_{\mathcal{Z}} p(x, z)\, dz = \int_{\mathcal{Z}} p(x|z)p(z)\, dz \\
&= \sum_{k=1}^{K} \pi_k p(x|z=k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k).
\end{aligned}
$$

Maximize the Likelihood

- Generative model:

$$1. \quad z \sim \text{Categorical}_\pi(K), \qquad \pi \in \Delta^{K-1},$$
$$2. \quad x \sim \mathcal{N}(\mu_z, \Sigma_z), \qquad \mu \in \mathbb{R}^{K\times d}, \Sigma \in \mathbb{R}^{K\times d\times d}.$$

- The maximum likelihood estimator (parameters in red):

$$\sup_\theta \frac{1}{n}\sum_{i=1}^n \log p_\theta(x_i) = \sup_\theta \frac{1}{n}\sum_{i=1}^n \log \sum_{k=1}^K \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k).$$

- No analytical solution; non-convex optimization problem

Gradient Ascent

- The maximum likelihood estimator (parameters in red):

$$\sup_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i) = \sup_{\theta} \frac{1}{n} \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k).$$

- Gradient ascent: initialize with random parameters and iteratively apply

$$\theta^{(i)} = \theta^{(i-1)} + \eta \nabla_\theta \sum_{i=1}^{n} \log p_\theta(x_i).$$

- Very important to start with a random initialization

Stochastic Gradient Ascent

- Gradient ascent:

$$\theta^{(i)} = \theta^{(i-1)} + \eta \nabla_\theta \sum_{i=1}^{n} \log p_\theta(x_i). \qquad \mathop{\mathbb{E}}_{x \sim p} \log p_\theta(x) \approx \frac{1}{n} \sum_{i=1}^{n} \log p_\theta(x_i).$$

- Stochastic gradient ascent (SGD):

$$\theta^{(i)} = \theta^{(i-1)} + \eta \nabla_\theta \log p_\theta(x_{i \pmod{\mathrm{n}}}). \qquad \mathop{\mathbb{E}}_{x \sim p} \log p_\theta(x) \approx \log p_\theta(x_i).$$

- Gradient ascent update: O(n)

- SGD update: O(1)

5-Minute Break

Recap: Evaluating the Likelihood

- To run SGD, we needed to evaluate the marginal probability $p_\theta(x)$:

$$p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x, z)\, dz = \int_{\mathcal{Z}} p_\theta(x|z) p_\theta(z)\, dz.$$

- For GMM's, this is tractable (the integral is a simple analytical sum).

- What would we do if the integral weren't tractable?

The Evidence Lower-Bound

- Approximate the marginal with importance sampling:

$$\begin{aligned}\log p_\theta(x) &= \log \mathop{\mathbb{E}}_{z\sim q(\cdot|x)}\left[\frac{p_\theta(x,z)}{q(z|x)}\right]\\ &= \mathop{\mathbb{E}}_{z\sim q(\cdot|x)}\left[\log\frac{p_\theta(x,z)}{q(z|x)}\right] + D(q(z|x)\,\|\,p_\theta(z|x))\\ &\geq \mathop{\mathbb{E}}_{z\sim q(\cdot|x)}\left[\log\frac{p_\theta(x,z)}{q(z|x)}\right].\end{aligned}$$

- The distribution $q(z|x)$ is called a proposal distribution

- Bound is tight when $q(z|x) = p(z|x)$ (when the proposal is the posterior)

The Evidence Lower-Bound

- We can also derive the Evidence Lower-Bound by Jensen's inequality:

$$\log p_\theta(x) = \log \mathop{\mathbb{E}}_{z \sim q(\cdot|x)} \left[\frac{p_\theta(x,z)}{q(z|x)} \right] \geq \mathop{\mathbb{E}}_{z \sim q(\cdot|x)} \left[\log \frac{p_\theta(x,z)}{q(z|x)} \right].$$

- Machine Learning community calls this lower bound the ELBO

- Another way to look at the ELBO:

$$\mathop{\mathbb{E}}_{z \sim q(\cdot|x)} \left[\log \frac{p_\theta(x,z)}{q(z|x)} \right] = \mathop{\mathbb{E}}_{z \sim q(\cdot|x)} \left[\log p_\theta(x|z) \right] - D(q(z|x) \parallel p(z)).$$

Expectation Maximization

- Jointly optimize the ELBO over θ and q:

$$\hat{\theta}_{\text{mle}} = \arg\max_{\theta} \max_{q} \mathop{\mathbb{E}}_{\substack{x \sim p \\ z \sim q(\cdot|x)}} \left[\log \frac{p_\theta(x, z)}{q(z|x)} \right].$$

- Alternating Optimization (Expectation Maximization):

1. Fix θ and optimize the proposal distribution q (E-step).

2. Fix the proposal distribution q and optimize θ (M-step).

EM for GMMs

- Sometimes the maximizer of the E-step has an analytic solution.

- For GMM's, the E-step is:

$$p_\theta(z|x) = \frac{p_\theta(x|z)p_\theta(z)}{p_\theta(x)} = \frac{\pi_z \mathcal{N}(x; \mu_z, \Sigma_z)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}.$$

- GMM's also have an analytical M-step:

$$\pi'_k = \frac{1}{n}\sum_{i=1}^{n} p_\theta(z_k|x_i), \qquad \mu'_k = \frac{1}{n\pi'_k}\sum_{i=1}^{n} p_\theta(z_k|x_i)x_i, \qquad \Sigma'_k = \frac{1}{n\pi'_k}\sum_{i=1}^{n} p_\theta(z_k|x_i)(x_i - \mu'_k)^{\otimes 2}.$$

Preview of Upcoming Topics

- On Wednesday we'll start talking about sequence models
- Next week: deep dive into neural parameterization of sequence models
- Homework 1 will be posted by the end of the week (due October 26th)
- The language modeling on HW1 will rely on next week's lectures