Machine Learning for Data Science (CS4786)
Lecture 13

Probabilistic Modelling, MLE Vs MAP Vs Bayesian,
Latent Variables

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Announcement

- Assignment P1 posted, due on thursday midnight.
- Office hours with TA are not for discussing P1
  (any question about P1: email me or as private question in Piazza)
- My office hours for tomorrow cancelled due to visit day, instead will have office hours on thursday, 3-4pm

I

Clustering

- For arbitrary set of points, we can have either
  - Scale invariance
  - Consistency

OR

  - Universality/Richness
- Assume structure or prior information on the set of points
- Assume we have set Θ of possible models and data is generated from one of these $\theta \in \Theta$:

$$(x_t, c_t) \sim P_\theta(|(x_1, c_1), \ldots, (x_{t-1}, c_{t-1}))$$

Examples

- Apple doesn't fall far from its tree model:
  - Each θ consists of position of initial trees $\mu_1, \ldots, \mu_K \in \mathbb{R}^2$ and mixture distribution $\pi = (\pi_1, \ldots, \pi_K)$ where π_i is the probability with which we get tree of fruit i
  - At time t we generate a new tree as follows:
    - $c_t \sim \pi$
    - $\text{Parent}_t \sim$ pick a parent tree uniformly from one of the c_t trees
    - $x_t \sim N(x_{\text{Parent}_t}, \Sigma)$
- Gaussian Mixture Model
  - Each θ consists of mixture distribution $\pi = (\pi_1, \ldots, \pi_K)$, means $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ and covariance matrices $\Sigma_1, \ldots, \Sigma_K$
  - At time t we generate a new tree as follows:

$$c_t \sim \pi, \qquad x_t \sim N(\mu_{c_t}, \Sigma_{c_t})$$

Probabilistic Models

More generally:

- Θ consists of set of possible parameters
- We have a distribution P_θ over the data induced by each $\theta \in \Theta$
- Data is generated by one of the $\theta \in \Theta$
- Learning: Estimate value or distribution for $\theta^* \in \Theta$ given data

Maximum Likelihood Principal

Pick $\theta \in \Theta$ that maximizes probability of observation

Reasoning:

- One of the models in Θ is the correct one
- Given data we pick the one that best explains the observed data
- Equivalently pick the maximum likelihood estimator,

$$\theta_{MLE} = \text{argmax}_{\theta \in \Theta} \log P_\theta(x_1, \ldots, x_n)$$

Often referred to as frequentist view

Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

Reasoning:

- Models are abstractions that capture our belief
- We update our belief based on observed data
- Given data we pick the model that we believe the most
- Pick θ that maximizes $\log P(\theta|x_1, \ldots, x_n)$

I want to say : Often referred to as Bayesian view

There are Bayesian and there Bayesians

Maximum A Posteriori

Pick $\theta \in \Theta$ that is most likely given data

Maximize a posteriori probability of model given data

$$\begin{aligned}
\theta_{MAP} &= \mathrm{argmax}_{\theta \in \Theta} P(\theta | x_1, \ldots, x_n) \\
&= \mathrm{argmax}_{\theta \in \Theta} \frac{P(x_1, \ldots, x_n | \theta) P(\theta)}{\sum_{\theta \in \Theta} P(x_1, \ldots, x_n | \theta) P(\theta)} \\
&= \mathrm{argmax}_{\theta \in \Theta} \frac{P(x_1, \ldots, x_n | \theta) P(\theta)}{P(x_1, \ldots, x_n)} \\
&= \mathrm{argmax}_{\theta \in \Theta} \underbrace{P(x_1, \ldots, x_n | \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}} \\
&= \mathrm{argmax}_{\theta \in \Theta} \log P(x_1, \ldots, x_n | \theta) + \log P(\theta)
\end{aligned}$$

EXAMPLE: GAUSSIAN MIXTURE MODEL

MLE: $\theta = (\mu_1, \ldots, \mu_K), \pi$

$$P_\theta(x_1, \ldots, x_n) = \prod_{t=1}^{n} \left(\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{(2 * 3.1415)^2 |\Sigma_i|}} \exp\left(-(x_t - \mu_i)^\top \Sigma_i (x_t - \mu_i)\right) \right)$$

MAP: with prior $\mu_i \sim N(0, \sigma I)$ and uniform prior on π

$$P(\theta | x_{1,\ldots,n}) = \prod_{t=1}^{n} \left(\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{(2 * 3.1415)^2 |\Sigma_i|}} \exp\left(-(x_t - \mu_i)^\top \Sigma_i (x_t - \mu_i)\right) \right)$$

$$\times \prod_{i=1}^{K} \frac{1}{\sqrt{(4 * 3.1415)^2}} \exp\left(-\|\mu_i\|^2\right)$$

What after we pick $\theta^* \in \Theta$?

- θ^* provides us a model/distribution from which data is generated
- In clustering for example, we can compute $P_{\theta^*}(c_t|x_t)$
- Hence we could assign to x_t cluster id c_t that has the largest probability. Inference step.
- *These are rough arguments*

Don't pick any $\theta^* \in \Theta$

- Model is simply an abstraction
- We have a prosteriori distribution over models, why pick one if in the end of the day we only want cluster assignments
- For each point find probability of cluster assignment we get by integrating over a posteriori probability of parameters θ
- We will come back to this later ...

Latent Variables

- We only observe locations of trees, we don't know which tree they are, ie. $c_1, \ldots, c_n$ are not observable
- Unobserved variables are referred to as latent variables
- We only pick θ_{MLE} or θ_{MAP} that maximizes likelihood or a posteriori probability given observation

So why bother with the latent variables?

MLE for GMM

Let us consider the one dimensional case,

$$\log P_\theta(x_{1,\ldots,n}) = \sum_{t=1}^{n} \log\left(\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2 * 3.1415\sigma_i^2}} \exp\left(-(x_t - \mu_i)^2/\sigma_i^2\right)\right)$$

Now consider the partial derivative w.r.t. μ_1, we have:

$$\frac{\partial \log P_\theta(x_{1,\ldots,n})}{\partial \mu_1} = \sum_{t=1}^{n} \frac{\frac{\pi_1}{\sigma_1} \exp\left(-\frac{(x_t-\mu_1)^2}{\sigma_1^2}\right)}{\sum_{i=1}^{K} \frac{\pi_i}{\sigma_i} \exp\left(-\frac{(x_t-\mu_i)^2}{\sigma_i^2}\right)}$$

Even given all other parameters, optimizing w.r.t. just μ_1 is hard!

MLE FOR GMM

Say by some magic you knew cluster assignments, then

$$\begin{aligned}
\log P_\theta((x_t, c_t)_{1,\ldots,n}) &= \sum_{t=1}^{n} \log\left(\frac{\pi_{c_t}}{\sqrt{2 * 3.1415\sigma_{c_t}^2}} \exp\left(-\frac{(x_t - \mu_{c_t})^2}{2\sigma_{c_t}^2}\right)\right) \\
&= \sum_{t=1}^{n} \left(\log(\pi_{c_t}) - \log(2 * 3.1415 * \sigma_{c_t}^2) - \frac{(x_t - \mu_{c_t})^2}{2\sigma_{c_t}^2}\right)
\end{aligned}$$

Now consider the partial derivative w.r.t. μ_i, we have:

$$\begin{aligned}
\frac{\partial \log P_\theta((x_t, c_t)_{1,\ldots,n})}{\partial \mu_i} &= -\frac{\partial}{\partial \mu_i} \sum_{t=1}^{n} \left(\frac{1}{2\sigma_{c_t}^2}(x_t - \mu_{c_t})^2\right) \\
&= -\frac{1}{2\sigma_i^2} \frac{\partial}{\partial \mu_i} \sum_{t:c_t=i} (x_t - \mu_i)^2 \\
&= \frac{1}{\sigma_1^2} \sum_{t:c_t=i} (x_t - \mu_i)
\end{aligned}$$

MLE for GMM

- Optimize for σ_i and π, what do you get?

Towards EM Algorithm

- Say we are interested in either MLE or MAP estimators
- Latent variables can help, but we have a chicken and egg problem

Given all variables maximizing likelihood/a posteriori is easy

Given model parameter, optimizing distribution over the latent variables is easy