CS-667 Advanced Machine Learning

Nazar Khan

PUCIT

The EM Algorithm

The EM Algorithm

- We have seen that K-means and GMMs are examples of latent variable models.
- Specifically for GMMs, we have seen an incremental algorithm for learning the parameters via ML.
- That algorithm is actually an instance of a powerful framework called *Expectation-Maximisation (EM)*.
- EM is used for *solving latent variable problems via ML*.
- We will now present a more general explanation of the EM algorithm.

- Maximum likelihood is equivalent to maximising the log-likelihood $\ln p(\mathbf{X}|\boldsymbol{\theta})$.
- Using the sum-rule

$$\ln p(\mathbf{X}|\boldsymbol{\theta}) = \ln\left(\sum_{\mathbf{Z}} p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})\right)$$

- Maximisation is no longer straight-forward since ln is 'blocked'by the summation.
- So we take another approach.

- We will denote $\{\mathbf{X}, \mathbf{Z}\}$ as the *complete* dataset.
- We will denote $\{\mathbf{X}\}$ as the *incomplete* dataset.
- The goal now is to maximise the complete-data log-likelihood function $p(\mathbf{X}, \mathbf{Z}|\theta)$.
- But for that we need to know the values of $\mathbf{Z}$ which are unobserved. What *can* be computed about $\mathbf{Z}$, however, is the posterior $p(\mathbf{Z}|\mathbf{X}, \theta)$.
- So instead of the *uncomputable, actual value* of log-likelihood, the *next best computable number* would be its *expected-value under the posterior* $p(\mathbf{Z}|\mathbf{X}, \theta)$.

- This yields the *E-step* of the EM algorithm.

$$\mathbb{E}_{\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}}[\ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})] = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X},\boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X},\mathbf{Z}|\boldsymbol{\theta})$$

- Since we are eventually interested in optimal parameters $\boldsymbol{\theta}^*$ we treat this expectation as a function of $\boldsymbol{\theta}$ and denote it by $\mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$.
- The *M-step* corresponds to maximising this expectation

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta},\boldsymbol{\theta}^{\text{old}})$$

- In short, EM replaces the log-likelihood by the *expected log-likelihood* and maximises it.
- Each EM cycle either moves toward or stays at a local maximum of $\ln p(\mathbf{X}|\boldsymbol{\theta})$.

The General EM Algorithm

Goal is to maximise likelihood $p(\mathbf{X}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ by introducing joint distribution $p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$ involving latent variables $\mathbf{Z}$.

1. Choose initial $\boldsymbol{\theta}^{\text{old}}$
2. **E-step**: Evaluate $p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}})$
3. **M-step**: Obtain new estimate $\boldsymbol{\theta}^{\text{new}}$ by maximising the expectation $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$

   $$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$$

   where $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta})$.
4. Check for convergence of either log-likelihood or parameters. If not converged, then

   $$\boldsymbol{\theta}^{\text{old}} \leftarrow \boldsymbol{\theta}^{\text{new}} \tag{1}$$

   and return to step 2.

Extensions of EM

- EM for MAP estimation via prior $p(\boldsymbol{\theta})$ amounts to modifying the M-step only.

$$\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) + \ln p(\boldsymbol{\theta})$$

- For problems with a 'difficult/intractable' M-step, maximisation can be replaced by a step that just increases $\mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$. This is known as the *Generalised EM* algorithm.

Proof of Convergence of EM

- Notice that $p(X|\theta) = p(X|\theta)\frac{p(Z|X,\theta)}{p(Z|X,\theta)} = \frac{p(X,Z|\theta)}{p(Z|X,\theta)}$.
- Recall that $\sum_{\mathbf{x}} q(\mathbf{x}) = 1$ for any distribution q over any random variable $\mathbf{x}$.
- Also recall that Kullback-Leibler divergence between probability distributions p and q is computed as

$$KL(p||q) = -\sum_{\mathbf{x}} p(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})}$$

which is non-symmetric

$$KL(q||p) = -\sum_{\mathbf{x}} q(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})}$$

and always non-negative.

Proof of Convergence of EM

This allows us to write the incomplete data log-likelihood as

$$\ln p(X|\theta) = \ln p(X|\theta) \underbrace{\sum_Z q(Z)}_{1}$$

$$= \sum_Z \ln p(X|\theta) q(Z) = \sum_Z \ln \frac{p(X,Z|\theta)}{p(Z|X,\theta)} q(Z)$$

$$= \sum_Z q(Z) \ln p(X,Z|\theta) - q(Z) \ln p(Z|X,\theta)$$

$$= \sum_Z q(Z) \ln p(X,Z|\theta) - q(Z) \ln q(Z) - q(Z) \ln p(Z|X,\theta) + q(Z) \ln q(Z)$$

$$= \underbrace{\sum_Z q(Z) \ln \frac{p(X,Z|\theta)}{q(Z)}}_{\mathcal{L}(q,\theta)} \underbrace{- \sum_Z q(Z) \ln \frac{p(Z|X,\theta)}{q(Z)}}_{KL(q||p) \geq 0}$$

Proof of Convergence of EM

- First term is a function of θ and a functional of q.
- Second term is the KL-divergence between $q(Z)$ and posterior $p(Z|X,\theta)$.
- Since $KL(q||p)$ is always ≥ 0

$$\ln p(X|\theta) = \mathcal{L}(q,\theta) + KL(q||p) \tag{2}$$
$$\implies \mathcal{L}(q,\theta) \leq \ln p(X|\theta)$$

- Therefore $\mathcal{L}(q,\theta)$ is a lower bound on the value of the incomplete data log-likelihood $\ln p(X|\theta)$.
- If we choose q or θ that increase the value of $\mathcal{L}(q,\theta)$, then the value of $\ln p(X|\theta)$ will also increase.

Proof of Convergence of EM

- *E-step:* Maximize $\mathcal{L}(q, \theta)$ with respect to q.
  - Since $\mathcal{L}(q, \theta)$ cannot exceed $\ln p(X|\theta)$, it's maximum value is $\ln p(X|\theta)$.
  - This occurs when $KL(q||p) = 0$.
  - This occurs when $q(Z) = p(Z|X, \theta)$. So that is q^*.
- *M-step:* Maximize $\mathcal{L}(q, \theta)$ with respect to θ.
- Since both the E-step and the M-step either increase or retain the lower-bound $\mathcal{L}(q, \theta)$, they either increase or retain the log-likelihood $\ln p(X|\theta)$ as well.
- Furthermore, since $KL(q||p) \geq 0$, Equation 2 implies that $\ln p(X|\theta)$ increases even more than the increase in the lower-bound.
- *Since each EM iteration either increases or retains the complete data log-likelihood, the algorithm is guaranteed to converge to a local maximum*.