2.6 Method of moments for latent-variable models

Motivation:

1. Rather tricky: we need to both estimate param. and infer the latent variables.

2. Latent variables: maximizing the marginal likelihood leads to a non-convex optimization. In practice, Expectation Maximization is often used to optimize these objective functions, but EM is only guaranteed to converge to a local optimum.

Goal:

explore a technique for param. estimation based on methods of moments.

[Example 1] Naive Bayes mixture model.

Let k be the number of document clusters.

Let b be ________ words in the vocabulary.

Let L be the length of a document.

* Model parameter $\theta = (\pi, B)$:

- $\pi \in \Delta_k$: prior distribution over clusters.
- $B = (\beta_1, \cdots, \beta_k) \in (\Delta_b)^k$: for each cluster h, $\beta_h \in \Delta_b$ is a distribution over words for cluster h.

Let Θ denotes the set of all possible θ.

* The probability model $p_\theta(h, x)$ is defined as follows:

- Sample the cluster: $h \sim \text{Multinomial}(\pi)$
- Sample the words in document independently:

$$x = (x_1, \cdots, x_L) \mid h \sim \text{Multinomial}(\beta_h)$$

Question:

Given n documents $\{x^{(1)}, \dots, x^{(n)}\}$ drawn i.i.d. from p_{θ^*}, return an estimate $\hat{\theta} = (\hat{\pi}, \hat{B})$ of $\theta^* = (\pi^*, B^*)$.

① Maximum (marginal) likelihood estimator:

$$\hat{\theta} = \arg\max_{\theta \in \Theta_1} \sum_{i=1}^{n} -\log \sum_{h=1}^{k} p_\theta(h, x^{(i)})$$

Optimization: EM

(i) E-step: for each example i, compute the posterior:

$$q_i(h) = p_\theta(h^{(i)} = h \mid x^{(i)})$$

(ii) M-step: optimise the expected log-likehood:

$$\max_\theta \sum_{i=1}^{n} \sum_{h=1}^{k} q_i(h) \log p_\theta(h, x^{(i)}).$$

② Method of moments

(i) define a moment mapping M

(ii) plug in the empirical moment $\hat{m}$ and get estimate $\hat{\theta}$ via the inverted mapping.

a. moment mapping

Let $\phi(x) \in \mathbb{R}^d$ be an observation function. Define the moment mapping as:

$$M(\theta) := \mathbb{E}_{x \sim p_\theta}[\phi(x)].$$

We say a mixture model is identifiable if $|M^{-1}(m)| = k!$ for all $m \in M(\Theta_1)$.

b. Plug in

(i) Define the empirical moments:

$$\hat{m} := \frac{1}{n} \sum_{i=1}^{n} \phi(x^{(i)})$$

(ii) Yield the method of moments estimator:

$$\hat{\theta} := M^{-1}(\hat{m})$$

c. Asymptotic analysis.

(i) By Central Limit Theorem:

$$\sqrt{n}(\hat{m} - m^*) \xrightarrow{d} N(0, \mathrm{Cov}_{x \sim p^*}[\phi(x)])$$

(ii) Assume that M^{-1} is continuous around m^*, by delta method

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \nabla M^{-1}(m^*) \mathrm{Cov}_{x \sim p^*}[\phi(x)] \nabla M^{-1}(m^*)^T)$$

Note: Method of moments if only useful if ϕ is well s.t.

(i) M is invertible.

(ii) M^{-1} is computationally tractable.

Now we compute $\hat{\theta}$ for Example 1.

Preliminaries:

(i) Assume each document has $L \geq 3$ words.

(ii) Assume $b \geq k$

(iii) each word x_j is represented into a one-hot vector $\in \mathbb{R}^b$.

Start with first-order moments:

$$M_1 := E[x_1] = \sum_{h=1}^{k} \pi_h \beta_h = B\pi$$

M_1 is a vector of marginal word probabilities.

We can write the second-order moments:

$$M_2 := E[x_1 x_2^T] = \sum_{h=1}^{k} \pi_h \beta_h \beta_h^T = B \operatorname{diag}(\pi) B^T$$

$M_2 \in R^{d \times d}$ is a matrix of co-occurrence word probabilities.

$M_2(u,v)$ is the probability of seeing u and v (marginally)

And we need a third-order moments:

$$M_3(\eta) := E[x_1 x_2^T (x_3^T \eta)] = \sum_{h=1}^{k} \pi_h \beta_h \beta_h^T (\beta_h^T \eta)$$

$$= B \operatorname{diag}(\pi) \operatorname{diag}(B^T \eta) B^T$$

[Lemma]

Suppose $X = BDB^T$, $Y = BEB^T$ where

(i) D, E are diagonal matrices s.t. $\{D_{ii}/E_{ii}\}_{i=1}^{k}$ are all non-zero and distinct.

(ii) $B \in \mathbb{R}^{b \times k}$ has full column rank.

The we can recover B

Pf: (i) Assume B is invertible, then X, Y are invertible.

$$Y X^{-1} = BEB^T B^{-T} D^{-1} B^{-1} = BED^{-1}B^{-1} \quad (ED^{-1} \text{ is diagonal})$$

The RHS has the form of an eigendecomposition, so the eigenvectors of $Y X^{-1}$ are exactly the columns of B up to permutation and scaling. B is full ranked since $ED^{-1}{}_{ii}$ is distinct for each $i = 1, \cdots, k$.

(ii) Now, suppose X, Y are not invertible.

Let $U \in \mathbb{R}^{b \times k}$ be any orthonormal basis of the column space of B, we have: $\tilde{B} := U^T B \in \mathbb{R}^{k \times k}$ is invertible.

Besides, we have

$$U^T X U = \tilde{B} D \tilde{B}^T, \quad U^T Y U = \tilde{B} E \tilde{B}^T,$$

which back to (i), and we can recover $\tilde{B}$.

Then $B = U\tilde{B}$. $\square$

We apply the Lemma with $X = M_2$, $Y = M_3(\eta)$

$D = \operatorname{diag}(\pi)$ and $E = \operatorname{diag}(\pi)\operatorname{diag}(B^T \eta)$

Once we recover B, then we can recover π via

$$\pi = B^{\dagger} M_1 \quad (B^{\dagger} \text{ is the pseudoinverse})$$