

25
SOICT

YEARS ANNIVERSARY

ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG



ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Nhập môn Học máy và Khai phá dữ liệu (IT3190)

Nội dung môn học

- Lecture 1: Giới thiệu về Học máy và khai phá dữ liệu
- Lecture 2: Thu thập và tiền xử lý dữ liệu
- Lecture 3: Hồi quy tuyến tính (Linear regression)
- Lecture 4+5: Phân cụm
- Lecture 6: Phân loại và Đánh giá hiệu năng
- Lecture 7: dựa trên láng giềng gần nhất (KNN)
- Lecture 8: Cây quyết định và Rừng ngẫu nhiên
- **Lecture 9: Học dựa trên xác suất**
- Lecture 10: Mạng nơron (Neural networks)
- Lecture 11: Máy vector hỗ trợ (SVM)
- Lecture 12: Khai phá tập mục thường xuyên và các luật kết hợp
- Lecture 13: Thảo luận ứng dụng trong thực tế

Expectation maximization

Expectation maximization

GMM

- Xét việc học GMM, với K phân phối Gaussian, từ dữ liệu huấn luyện $\mathbf{D} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$.
- Hàm mật độ $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \sum_{k=1}^K \phi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
 - $\boldsymbol{\phi} = (\phi_1, \dots, \phi_K)$ chứa cho trọng số của từng phân phối $P(z = k | \boldsymbol{\phi}) = \phi_k$
 - Mỗi Gaussian đa biến có hàm mật độ:

$$\mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_k)}} \exp \left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) \right]$$

- MLE cố gắng cực đại hàm log-likelihood sau:

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = \sum_{i=1}^M \log \sum_{k=1}^K \phi_k \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Không thể tìm được công thức nghiệm cụ thể!
- **Naïve gradient decent** : lặp hai bước sau cho đến khi hội tụ
 - Tối ưu hóa $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$ theo biến $\boldsymbol{\phi}$, khi cố định $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
 - Tối ưu hóa $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$ theo biến $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, khi cố định $\boldsymbol{\phi}$.

GMM và K-means

□ GMM: ta cần biết

- Trong số K Gaussian, phân bố nào sinh ra dữ liệu \mathbf{x}
chỉ số z của phân bố đó
- Tham số của từng phân phối:
 $(\mu_k, \Sigma_k, \phi_k)$

□ K-means:

- Trong số K cụm thì x thuộc về cụm nào?
Chỉ số z của cụm
- Tham số của từng cụm: Tâm cụm

□ Ý tưởng cho GMM

- $P(z|\mathbf{x}, \mu, \Sigma, \phi)$?
(chú ý $\sum_{k=1}^K P(z = k|\mathbf{x}, \mu, \Sigma, \phi) = 1$)
(gán “mềm” vào các cụm)
- Cập nhật tham số cho từng phân bố Gaussian: $(\mu_k, \Sigma_k, \phi_k)$

□ Huấn luyện K-means:

- Bước 1: phân bổ mỗi \mathbf{x} vào cụm gần nhất
(gán nhãn cụm cho từng \mathbf{x})
(cách gán “cứng nhắc”)
- Bước 2: tính toán lại tâm các cụm

GMM: cận dưới

□ Ý tưởng của GMM?

- Bước 1: tính $P(z|x, \mu, \Sigma, \phi)$? (note $\sum_{k=1}^K P(z = k|x, \mu, \Sigma, \phi) = 1$)
- Bước 2: Cập nhật tham số cho các phân bố: $\theta = (\mu, \Sigma, \phi)$

• Xét hàm log-likelihood

$$L(\theta) = \log P(\mathbf{D}|\theta) = \sum_{i=1}^M \log \sum_{k=1}^K \phi_k \mathcal{N}(\mathbf{x}_i | \mu_k, \Sigma_k)$$

- Quá phức tạp nếu trực tiếp sử dụng đạo hàm
- Lưu ý rằng

$$\begin{aligned} \log P(\mathbf{x}|\theta) &= \log \sum_z P(z, \mathbf{x}|\theta) = \log \sum_z P(z|\mathbf{x}, \theta) P(\mathbf{x}|\theta) \\ &= \log \mathbb{E}_{z|\mathbf{x}, \theta} P(\mathbf{x}|\theta) \geq \mathbb{E}_{z|\mathbf{x}, \theta} \log P(\mathbf{x}|\theta) = \sum_z P(z|\mathbf{x}, \theta) \log P(\mathbf{x}|\theta) \end{aligned}$$

BĐT
Jensen

- Tối đa hóa $L(\theta)$ có thể được thực hiện bằng cách tối đa hóa giới hạn dưới $\mathbb{E}_{z|\mathbf{D}, \theta} \log P(\mathbf{D}|\theta)$

GMM: cực đại hoá cận dưới

□ Ý tưởng của GMM?

- Bước 1: tính $P(z|\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$? (note $\sum_{k=1}^K P(z = k|\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi}) = 1$)
- Bước 2: Cập nhật tham số cho từng phân phối Gaussian: $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$

- Quy tắc Bayes: $P(z|\mathbf{x}, \boldsymbol{\theta}) = P(\mathbf{x}|z, \boldsymbol{\theta})P(z|\boldsymbol{\phi})/P(\mathbf{x}) = \phi_z \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z)/C$, trong đó $C = \sum_k \phi_k \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ là hằng số chuẩn hóa.
 - Có nghĩa là người ta có thể tính $P(z|\mathbf{x}, \boldsymbol{\theta})$ nếu biết $\boldsymbol{\theta}$
 - Đặt $T_{ki} = P(z = k|\mathbf{x}_i, \boldsymbol{\theta})$ với mọi $k = \overline{1, K}, i = \overline{1, M}$
- Còn về $\boldsymbol{\phi}$ thì sao?
 - $\phi_z = P(z|\boldsymbol{\phi}) = P(z|\boldsymbol{\theta}) = \int P(z, \mathbf{x}|\boldsymbol{\theta})d\mathbf{x} = \int P(z|\mathbf{x}, \boldsymbol{\theta})P(\mathbf{x}|\boldsymbol{\theta})d\mathbf{x} = \mathbb{E}_{\mathbf{x}}(P(z|\mathbf{x}, \boldsymbol{\theta})) \approx \frac{1}{M} \sum_{\mathbf{x} \in D} P(z|\mathbf{x}, \boldsymbol{\theta}) = \frac{1}{M} \sum_{i=1}^M T_{zi}$
- Khi đó, cận dưới có thể được cực đại hóa theo mỗi phân bố $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$:

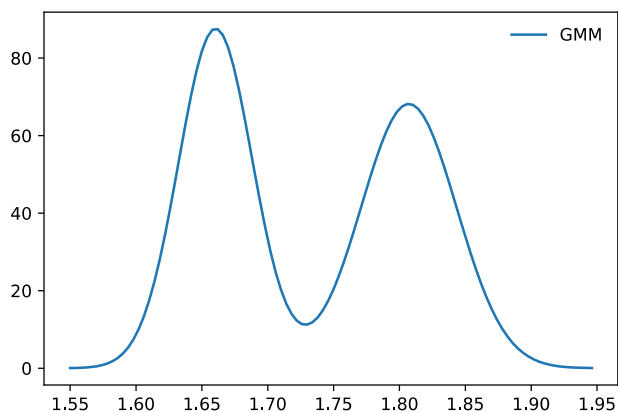
$$\begin{aligned} \mathbb{E}_{z|\mathbf{D}, \boldsymbol{\theta}} \log P(\mathbf{D}|\boldsymbol{\theta}) &= \sum_{\mathbf{x} \in D} \sum_z P(z|\mathbf{x}, \boldsymbol{\theta}) \log P(\mathbf{x}|\boldsymbol{\theta}) \\ &= \sum_{i=1}^M \sum_{k=1}^K T_{ki} \left[-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) - \log \sqrt{\det(2\pi \boldsymbol{\Sigma}_k)} \right] \end{aligned}$$

GMM: Thuật toán EM

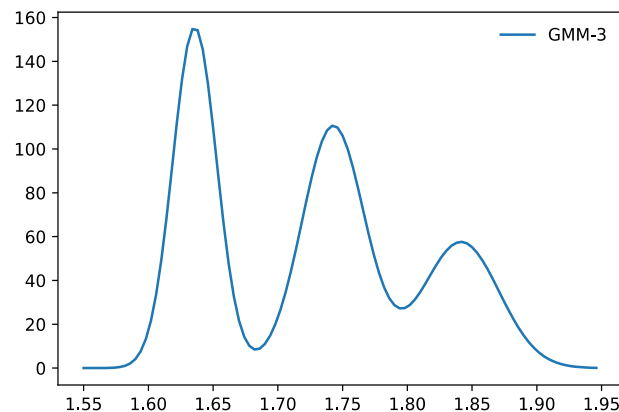
- **Đầu vào:** dữ liệu huấn luyện $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}$, $K > 0$
- **Đầu ra:** tham số mô hình $(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\phi})$
- Khởi tạo $(\boldsymbol{\mu}^{(0)}, \boldsymbol{\Sigma}^{(0)}, \boldsymbol{\phi}^{(0)})$ một cách ngẫu nhiên
 - $\boldsymbol{\phi}^{(0)}$ phải không âm và tổng bằng 1.
- Tại lần lặp t :
 - *Bước E:* tính $T_{ki} = P(z = k | \mathbf{x}_i, \boldsymbol{\theta}^{(t)}) = \phi_k^{(t)} \mathcal{N}(\mathbf{x}_i | \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}) / C$ cho mọi k, i
 - *Bước M:* cập nhật cho mọi k
$$\phi_k^{(t+1)} = \frac{1}{M} \sum_{i=1}^M T_{ki}; \boldsymbol{\mu}_k^{(t+1)} = \frac{1}{M \phi_k} \sum_{i=1}^M T_{ki} \mathbf{x}_i;$$
$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{1}{M \phi_k} \sum_{i=1}^M T_{ki} (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)}) (\mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1)})^T$$
- Nếu không hội tụ, chuyển đến bước lặp $t + 1$.

GMM: Ví dụ 1

- Chúng ta mong muốn mô hình hóa chiều cao của một người
 - Chúng ta đã có dữ liệu thu thập từ 10 người ở Hà Nội và 10 Người ở Sysney: $\mathbf{D} = \{1.60, 1.70, 1.65, 1.63, 1.75, 1.71, 1.68, 1.72, 1.77, 1.62, 1.75, 1.80, 1.85, 1.65, 1.91, 1.78, 1.88, 1.79, 1.82, 1.81\}$



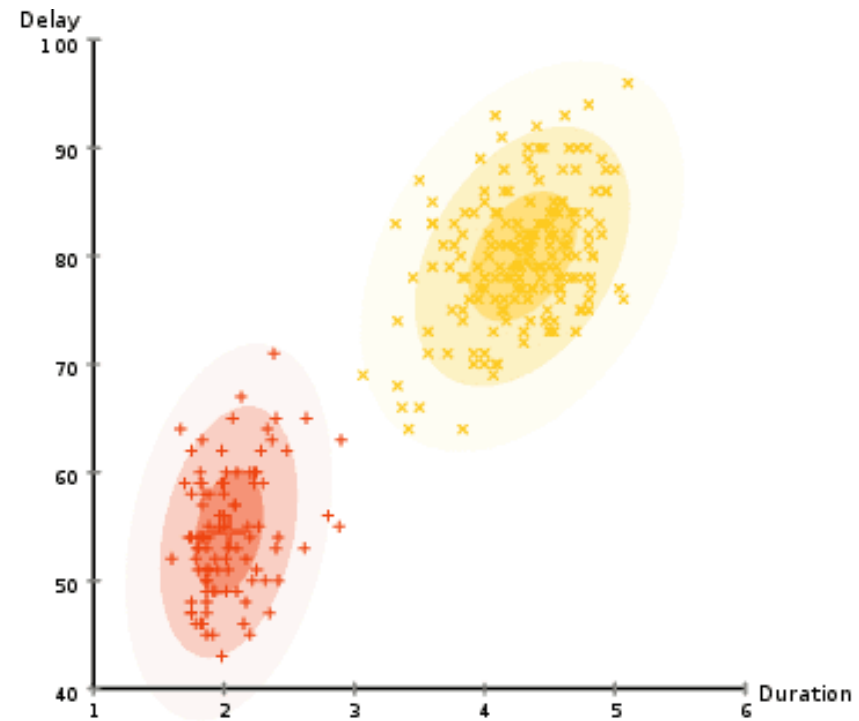
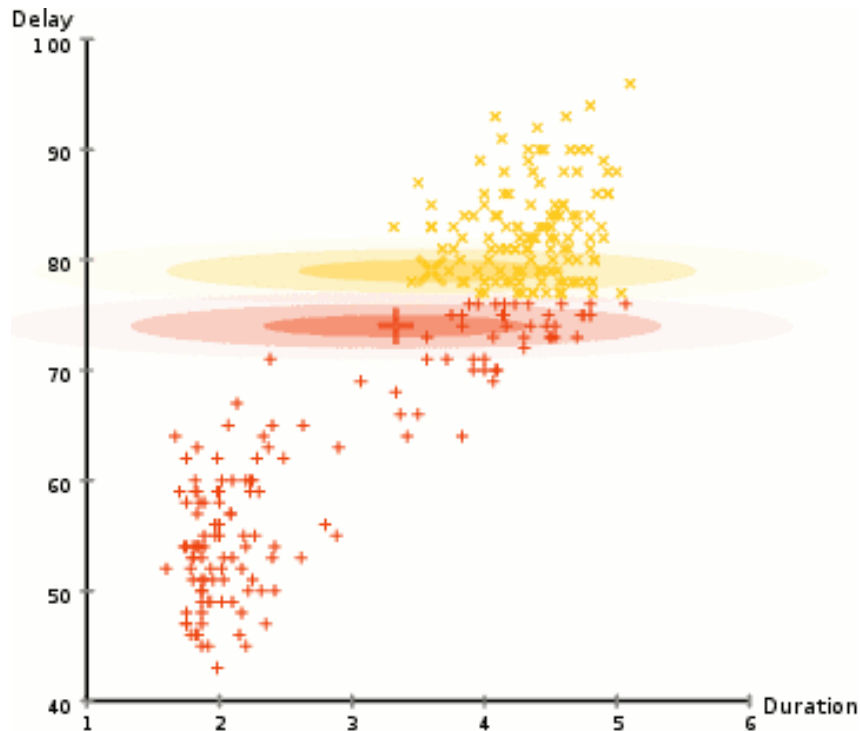
GMM với
2 phân bố con



GMM với
3 phân bố con

GMM: Ví dụ 2

- GMM được huấn luyện trên bộ dữ liệu hai chiều



GMM: so sánh với K-means

□ K-means:

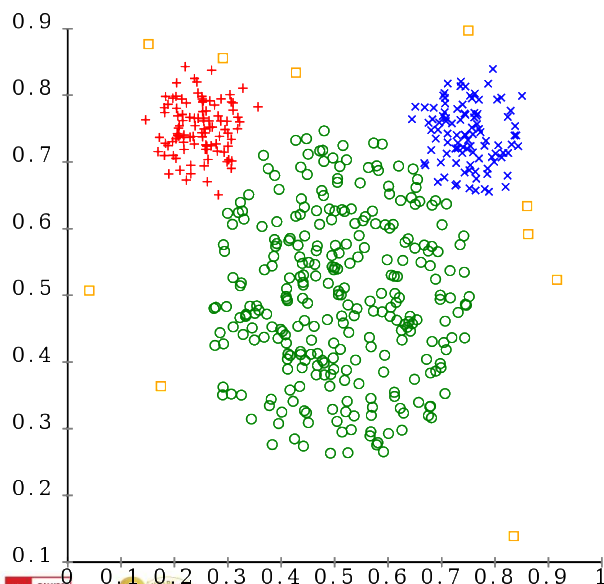
- Bước 1: phân bổ cứng nhắc
- Bước 2: tìm tâm cụm
→ hình dạng các cụm là giống nhau?

□ GMM:

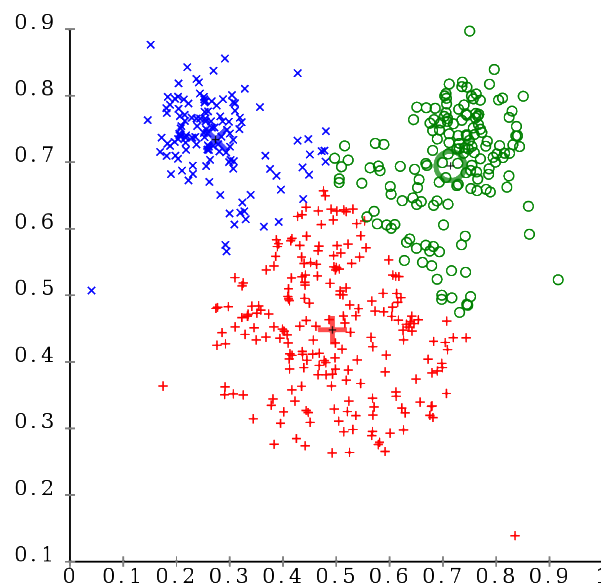
- Gán “mềm” dữ liệu vào các cụm
- Tham số $(\mu_k, \Sigma_k, \phi_k)$
→ hình dáng các cụm có thể khác nhau

Different cluster analysis results on "mouse" data set:

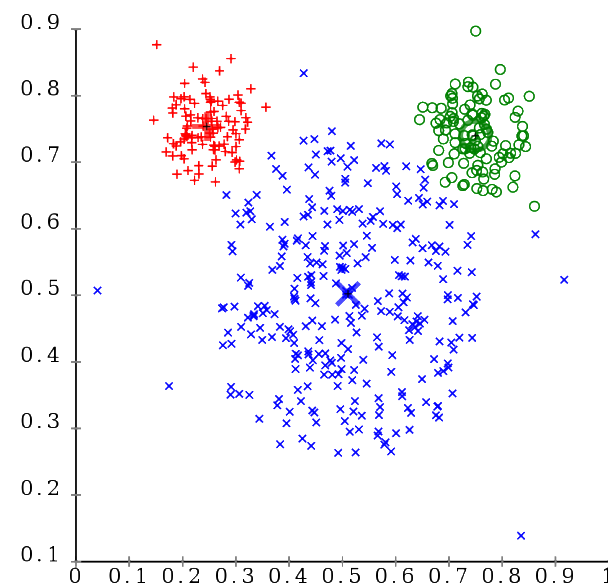
Original Data



k-Means Clustering



EM Clustering



Mô hình tổng quát

- Chúng ta có thể áp dụng thuật toán EM trong các trường hợp tổng quát hơn.
- Xét một mô hình $B(x, z; \theta)$ với biến quan sát được x , biến ẩn z và được tham số hóa bởi θ
 - x phụ thuộc vào z và θ , trong khi z có thể phụ thuộc vào θ
 - Mô hình hỗn hợp: mỗi điểm dữ liệu được quan sát có một biến ẩn tương ứng, chỉ định component tạo ra điểm dữ liệu
- Việc học là tìm một mô hình cụ thể, từ họ mô hình được tham số hóa bởi θ , mà giúp hàm log-likelihood trên dữ liệu huấn luyện D đạt cực đại:

$$\theta^* = \operatorname{argmax}_{\theta} \log P(D|\theta)$$

- Giả sử D bao gồm các mẫu độc lập, hàm log-likelihood $\mathbb{E}_{z|D, \theta} \log P(D|\theta)$ có thể tính toán dễ dàng
 - Do có một biến ẩn, MLE có thể không có công thức nghiệm tường minh

Thuật toán EM

- Thuật toán Expectation maximization (EM) được giới thiệu vào năm 1977 bởi Arthur Dempster, Nan Laird và Donald Rubin.

- EM cực đại hoá cận dưới của hàm log-likelihood:

$$L(\boldsymbol{\theta}; \boldsymbol{D}) = \log P(\boldsymbol{D}|\boldsymbol{\theta}) \geq \mathbb{E}_{\boldsymbol{z}|\boldsymbol{D},\boldsymbol{\theta}} \log P(\boldsymbol{D}|\boldsymbol{\theta}) = \sum_{\boldsymbol{z}} P(\boldsymbol{z}|\boldsymbol{D}, \boldsymbol{\theta}) \log P(\boldsymbol{D}|\boldsymbol{\theta})$$

- Khởi tạo: $\boldsymbol{\theta}^{(0)}$, $t = 0$

- Tại lần lặp t :

- **Bước E:** tính hàm kỳ vọng Q khi cố định giá trị $\boldsymbol{\theta}^{(t)}$ đã biết ở bước trước

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{z}|\boldsymbol{D},\boldsymbol{\theta}^{(t)}} \log P(\boldsymbol{D}|\boldsymbol{\theta}^{(t)})$$

- **Bước M:** tìm điểm $\boldsymbol{\theta}^{(t+1)}$ mà làm cho hàm Q đạt cực đại
$$\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$$

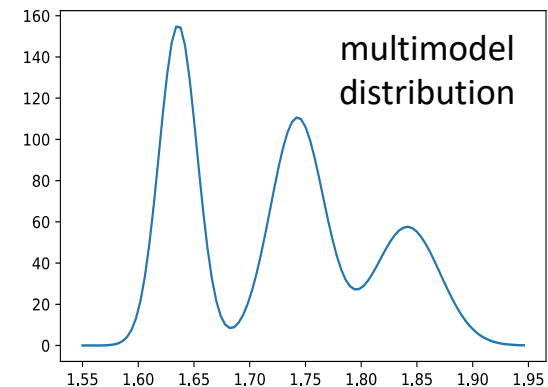
- Nếu không hội tụ, chuyển đến bước lặp $t + 1$.

EM: điều kiện hội tụ

- Các điều kiện khác nhau có thể được sử dụng để kiểm tra sự hội tụ
 - $\mathbb{E}_{z|\mathbf{D},\boldsymbol{\theta}} \log P(\mathbf{D}|\boldsymbol{\theta})$ không thay đổi nhiều giữa hai lần lặp liên tiếp
 - $\boldsymbol{\theta}$ không thay đổi nhiều giữa hai lần lặp lại liên tiếp
- Trong thực tế, đôi khi chúng ta cần giới hạn số lần lặp tối đa

EM: vài tính chất

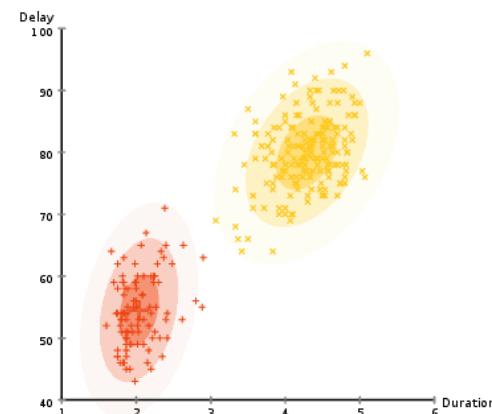
- Thuật toán EM được đảm bảo trả về một điểm dừng của cận dưới $\mathbb{E}_{z|D, \theta} \log P(D|\theta)$
 - Đó có thể là tối đa cục bộ
- Do cực đại hoá cận dưới, EM không nhất thiết phải trả về nghiệm tối ưu
 - Không có lý thuyết để đảm bảo
 - Trong các trường hợp phức tạp (vd, multimodel), khi mà hàm log-likelihood là không lồi
- Thuật toán Baum-Welch là trường hợp đặc biệt của EM cho các mô hình Markov ẩn



EM, mô hình hỗn hợp và phân cụm

- **Mô hình hỗn hợp**: giả sử tập dữ liệu bao gồm K thành phần (phân bố) khác nhau và mỗi quan sát được tạo ra từ một trong các phân bố đó
 - Ví dụ: mô hình hỗn hợp Gaussian, mô hình hỗn hợp Multinomial, mô hình hỗn hợp Bernoulli,...
 - Hàm mật độ hỗn hợp có thể được viết dưới dạng

$$f(x; \theta, \phi) = \sum_{k=1}^K \phi_k f_k(x | \theta_k)$$



trong đó $f_k(x | \theta_k)$ là hàm mật độ của phân phối thứ k

- Một phân bố hỗn hợp tạo ra một cách chia không gian dữ liệu ra thành các vùng khác nhau, mà mỗi vùng tương ứng với 1 thành phần trong hỗn hợp đó
- Do đó, các mô hình hỗn hợp cung cấp các phương pháp để phân cụm
- EM cung cấp một cách tự nhiên để học các mô hình hỗn hợp

EM: hạn chế

- Khi cần dưới $\mathbb{E}_{z|\mathbf{D},\boldsymbol{\theta}} \log P(\mathbf{D}|\boldsymbol{\theta})$ không cho phép tính toán kỳ vọng hoặc tối ưu hóa dễ dàng
 - Mô hình hỗn hợp, mô hình hỗn hợp Bayes
 - Mô hình xác suất phân cấp
 - Mô hình phi tham số
- EM tìm ước lượng điểm, do đó dễ dàng bị kẹt ở mức tối ưu cục bộ
- Trong thực tế, EM nhạy cảm với việc khởi tạo
 - Sử dụng ý tưởng về K-mean++ để khởi tạo?
- Đôi khi EM hội tụ chậm trong thực tế

Chủ đề thêm

- Variational inference
 - Suy diễn cho các mô hình tổng quát hơn
- Deep generative models
 - Mạng nơron + lý thuyết xác suất
- Mạng nơron Bayes
 - Mạng nơron + Suy luận Bayes
- Amortized inference
 - Mạng nơron để thực hiện suy diễn Bayes
 - Học cách suy diễn

Tài liệu tham khảo

- Blei, David M., Alp Kucukelbir, and Jon D. McAuliffe. "Variational inference: A review for statisticians." *Journal of the American Statistical Association* 112, no. 518 (2017): 859-877.
- Blundell, Charles, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. "Weight Uncertainty in Neural Network." In *International Conference on Machine Learning (ICML)*, pp. 1613-1622. 2015.
- Dempster, A.P.; Laird, N.M.; Rubin, D.B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm". *Journal of the Royal Statistical Society, Series B*. 39 (1): 1-38.
- Gal, Yarin, and Zoubin Ghahramani. "Dropout as a bayesian approximation: Representing model uncertainty in deep learning." In *ICML*, pp. 1050-1059. 2016.
- Ghahramani, Zoubin. "Probabilistic machine learning and artificial intelligence." *Nature* 521, no. 7553 (2015): 452-459.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." In *International Conference on Learning Representations (ICLR)*, 2014.
- Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." *Science* 349, no. 6245 (2015): 255-260.
- Tosh, Christopher, and Sanjoy Dasgupta. "The Relative Complexity of Maximum Likelihood Estimation, MAP Estimation, and Sampling." In *COLT, PMLR* 99:2993-3035, 2019.
- Sontag, David, and Daniel Roy, "Complexity of inference in latent dirichlet allocation" in: *Advances in Neural Information Processing System*, 2011.



25 YEARS ANNIVERSARY
SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

**Thank
you for
your
attentions
!**



soict.hust.edu.vn/



fb.com/groups/soict

