Clustering

- 前言

clustering 是 unsupervised learning
可用來分類 incomplete data.

complete data	incomplete data
target / label	only x is given
regression $y = w^T x + \varepsilon$	EM Algorithmn
classification $\frac{1}{1+e^{-x^T w}}$	KNN
	GMM (Gaussian Mixure Model)

Hierarchy cluster

- 概念：① measure pairwise distance/similarity between data
  ② keep grouping the closest

前言

- complete data

1. one-coin tossing problem

$X = \{H, H, T, T, H\} \sim \text{Bernoulli}\left(\begin{smallmatrix}H \\ T\end{smallmatrix} \middle| P\right)$

$P_{MLE} = \dfrac{\sum_{i=1}^{n} x_i}{N}$

2. two-coin tossing problem

有兩種銅板 C_0, C_1，各正面机率為 P_0, P_1

$X = \{1, 0, 1, 1, 0, 0, 1, 1, 0, 0\}$

$Z = \{C_0\ C_1\ C_0\ C_0\ C_1\ C_1\ C_0\ C_0\ C_1\ C_1\}$

欲求 ① λ: chance that C_0 is used
② likelihood : $P(x \mid \underbrace{P_0, P_1}_{\theta}, \underbrace{Z}_{\text{label}})$

$$= \prod_{i=1}^{n} \left[\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right]^{1-z_i} \left[(1-\lambda) P_1^{x_i}(1-P_1)^{1-x_i}\right]^{z_i}$$

$\boxed{\text{以第一筆 } X, Z \text{ 為例}, x_i = 1, z_i = 0 \text{ 則 likelihood } = [\lambda P_0^1 (1-P_0)^0]^1}$

MLE找參數：$J = \log(\text{likelihood})$

$$J=\sum_{i=1}^{n}(1-z_i)\log\left[\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right]+\sum_{i=1}^{n}z_i\log\left[(1-\lambda)P_1^{x_i}(1-P_1)^{1-x_i}\right]$$

① 令 $\frac{\delta J}{\delta\lambda}=\sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda}-\sum_{i=1}^{n}z_i\frac{1}{1-\lambda}=0$

$\Rightarrow \sum_{i=1}^{n}(1-z_i)\frac{1}{\lambda}=\sum_{i=1}^{n}z_i\frac{1}{1-\lambda}$

$\Rightarrow (1-\lambda)\sum_{i=1}^{n}(1-z_i)=\lambda\sum_{i=1}^{n}z_i$

$\Rightarrow \sum_{i=1}^{n}(1-z_i)=\lambda\sum_{i=1}^{n}(1-z_i)+z_i=n$

$\therefore \lambda=\frac{\sum_{i=1}^{n}(1-z_i)}{n}=\frac{\text{次數}(C_0)}{n}$

② 令 $\frac{\delta J}{\delta P_0}=\sum_{i=1}^{n}(1-z_i)\left(\frac{x_i}{P_0}-\frac{1-x_i}{1-P_0}\right)=0$

$\Rightarrow \sum(1-z_i)\frac{x_i}{P_0}=\sum(1-z_i)\frac{1-x_i}{1-P_0}$

$\Rightarrow (1-P_0)\sum(1-z_i)x_i=P_0\sum(1-z_i)(1-x_i)$

$\Rightarrow \sum(1-z_i)x_i=P_0\sum(1-z_i)\left[(1-x_i)+x_i\right]$

$\Rightarrow \sum(1-z_i)x_i=P_0\sum(1-z_i)$

$\therefore P_0=\frac{\sum_{i=1}^{n}(1-z_i)x_i}{\sum_{i=1}^{n}(1-z_i)}=\frac{\text{次數}(C_0\text{且正})}{\text{次數}(C_0)}$

③ P_1 同理

- incomplete data：沒有 label z_i，故需 λ

$X=\{1,0,1,1,0,0,1,0,0,0\}$

目標學 w_i, where $w_i=\frac{P(z_i=C_0,x_i|\theta)}{P(z_i=C_0,x_i|\theta)+P(z_i=C_1,x_i|\theta)}$

其中已觀察到 $\overset{\theta}{\lambda,P_0,P_1}$，$\begin{cases}P(z_i=C_0,x_i|\theta)=\lambda P_0^{x_i}(1-P_0)^{1-x_i}\\ P(z_i=C_1,x_i|\theta)=(1-\lambda)P_1^{x_i}(1-P_1)^{1-x_i}\end{cases}$

$w_i=\frac{P(z,x|\theta)}{\sum_z P(z,x|\theta)}=\frac{P(z,x|\theta)}{P(x|\theta)}=\frac{P(x|z,\theta)P(z|\theta)}{P(x|\theta)}$ w_i is posterior $=P(z|x,\theta)$

則 $P(X|\lambda,P_0,P_1,w_i)=\prod_{i=1}^{n}\left[\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right]^{w_i}\cdot\left[(1-\lambda)P_1^{x_i}(1-P_1)^{1-x_i}\right]^{1-w_i}$

$J=\sum w_i\log\left(\lambda P_0^{x_i}(1-P_0)^{1-x_i}\right)+\sum(1-w_i)\log\left[(1-\lambda)P_1^{x_i}(1-P_1)^{1-x_i}\right]$

$\frac{\delta J}{\delta w_i}=\sum w_i\frac{1}{\lambda}-\sum(1-w_i)\frac{1}{1-\lambda}$

$\Rightarrow (1-\lambda)\sum w_i=\lambda\sum(1-w_i)$

$\Rightarrow \sum w_i=\lambda n$

$\Rightarrow \lambda_{updated}=\frac{\sum w_i}{n}$

$\frac{\delta J}{\delta P_0}=\sum w_i\left(\frac{x_i}{P_0}-\frac{1-x_i}{1-P_0}\right)=0$

$\Rightarrow (1-P_0)\sum w_i x_i=P_0\sum w_i(1-x_i)$

$\Rightarrow \sum w_i x_i=P_0\sum w_i$

$\Rightarrow P_0=\frac{\sum w_i x_i}{\sum w_i}$

同理得 $P_0=\frac{\sum w_i x}{\sum w_i}$ $P_1=\frac{\sum(1-w_i)x_i}{\sum(1-w_i)}$ 同理 $P_1=\frac{\sum(1-w_i)x_i}{\sum(1-w_i)}$

EM 範例

$X = \{1, 0, 1, 1, 0, 0, 1, 0, 0, 0\}$

初始 θ : $\lambda^{(0)} = \frac{1}{3}$ $\quad P_0^{(0)} = \frac{1}{3}$ $\quad P_1^{(0)} = \frac{2}{3}$

1 Iter : E step

① 出現 1 時 為 C_0 的 posterior $w_i = 0.2$
為 C_1 的 posterior $(1-w_i) = 0.8$

$$w_i = \frac{P(Z_i = C_0, x_i = 1 \mid \theta)}{P(Z_i = C_0, x_i = 1 \mid \theta) + P(Z_i = C_1, x_i = 1 \mid \theta)}$$

$$= \frac{\lambda P_0}{\lambda P_0 + (1-\lambda) P_1} = \frac{\frac{1}{9}}{\frac{1}{9} + \frac{4}{9}} = 0.2$$

② 出現 0 時 為 C_0 的 posterior $w_i = 0.5$
為 C_1 的 posterior $(1-w_i) = 0.5$

$$w_i = \frac{P(Z_i = C_0, x_i = 0 \mid \theta)}{P(Z_i = C_0, x_i = 0 \mid \theta) + P(Z_i = C_1, x_i = 0 \mid \theta)}$$

$$= \frac{\lambda (1-P_0)}{\lambda (1-P_0) + (1-\lambda)(1-P_1)} = \frac{\frac{2}{9}}{\frac{2}{9} + \frac{2}{9}} = 0.5$$

1 Iter : M step

$$\lambda = \frac{\sum w_i}{n} = \frac{0.2 \times 4 + 0.5 \times 6}{10} = 0.38$$

$$P_0 = \frac{\sum w_i x}{\sum w_i} = \frac{0.2 \times 4}{3.8} = \frac{4}{19}$$

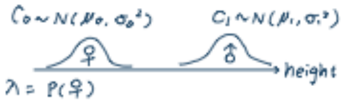
$$P_1 = \frac{\sum (1-w_i) x}{\sum (1-w_i)} = \frac{0.8 \times 4}{0.8 \times 4 + 0.5 \times 6} = \frac{16}{31}$$

2 Iter : Use $\theta(\lambda, P_0, P_1)$ of last iter to perform EM steps till convergence.

§ Gaussian Mixture Model

- 概念

Data 為 gaussian 分佈 e.g., 男女身高



- complete data ：对 likelihood 作 MLE

likelihood $= p(X \mid \lambda, \mu_0, \mu_1, \sigma_0, \sigma_1)$

$$= \prod_{i=1}^{n} \left[\lambda \frac{1}{\sqrt{2\pi\sigma_0^2}} e^{\frac{-(x_i-\mu_0)^2}{2\sigma_0^2}}\right]^{1-z_i} \left[(1-\lambda)\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}}\right]^{z_i}$$

$$\Rightarrow J = \sum (1-z_i)\left[\ln\lambda - \ln\sqrt{2\pi\sigma_0^2} - \frac{(x_i-\mu_0)^2}{2\sigma_0^2}\right] + \sum z_i \left[\ln(1-\lambda) - \ln\sqrt{2\pi\sigma_1^2} - \frac{(x_i-\mu_1)^2}{2\sigma_1^2}\right]$$

$$\frac{\delta J}{\delta \lambda} = \sum (1-z_i)\frac{1}{\lambda} - \sum z_i \frac{1}{1-\lambda} = 0$$

$\Rightarrow (1-\lambda)\sum(1-z_i) = \lambda \sum z_i$

$\Rightarrow \sum(1-z_i) = \lambda n$

$\Rightarrow \lambda_{MLE} = \frac{\sum(1-z_i)}{n}$

$$\frac{\delta J}{\delta \mu_0} = \sum \frac{(1-z_i)}{\cancel{2\sigma_0^2}} \cancel{2}(x_i-\mu_0) = 0$$

$\Rightarrow \sum(1-z_i)x_i = \sum(1-z_i)\mu_0$

$\Rightarrow \mu_{0_{MLE}} = \frac{\sum(1-z_i)x_i}{\sum(1-z_i)}$

$\Rightarrow \mu_{1_{MLE}} = \frac{\sum z_i x_i}{\sum z_i}$

$$\frac{\delta J}{\delta \sigma_0} = \sum(1-z_i)\left[\frac{-2\cancel{\sqrt{2\pi}}}{\sqrt{\cancel{2\pi}}\sigma_0^2} + \frac{(x_i-\mu_0)^2}{\sigma_0^3}\right] = 0$$

$\Rightarrow \sum(1-z_i)\left[\frac{-2}{\sigma_0} + \frac{(x_i-\mu_0)^2}{\sigma_0^3}\right] = 0$

$\Rightarrow \sum(1-z_i)\frac{2\sigma_0^2}{\cancel{\sigma_0}} = \sum(1-z_i)\frac{(x_i-\mu_0)^2}{\cancel{\sigma_0^3}}$

$\Rightarrow \sigma_{0_{MLE}}^2 = \frac{\sum(1-z_i)(x_i-\mu_0)^2}{2\sum(1-z_i)}$

$\Rightarrow \sigma_{1_{MLE}}^2 = \frac{\sum z_i(x_i-\mu_1)^2}{2\sum z_i}$

若 likelihood 為 multivariate gaussian

likelihood $= \prod_i \underline{N(x_i \mid \mu_0, \Sigma_0)}^{1-z_i} \cdot N(x_i \mid \mu_1, \Sigma_1)^{z_i}$

$= \lambda_0 \cdot \frac{1}{\sqrt{2\pi}^k |\Sigma|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$

$$J = \sum(1-z_i)\left[-\frac{1}{2}\log|\Sigma_0| - \frac{1}{2}(x-\mu_0)^T\Sigma_0^{-1}(x-\mu_0) + \log\lambda_0\right]$$
$$+ \sum(z_i)\left[-\frac{1}{2}\log|\Sigma_1| - \frac{1}{2}(x-\mu_1)^T\Sigma_1^{-1}(x-\mu_1) + \log\lambda_1\right]$$

$\lambda_0 = \frac{\sum(1-z_i)}{n}$

$$\frac{\delta J}{\delta \mu_0} = \sum(1-z_i)\cancel{\Sigma_0^{-1}}(x-\mu_0) = 0$$

$\Rightarrow \mu_0 = \frac{\sum(1-z_i)x}{\sum(1-z_i)}$

$$\frac{\delta J}{\delta \Sigma_0} = \sum(1-z_i)\left[-\cancel{\frac{1}{2}}\frac{1}{|\Sigma_0|} + \cancel{\frac{1}{2}}(x-\mu_0)^T(x-\mu_0)\Sigma_0^{-2}\right] = 0$$

$\Rightarrow \sum(1-z_i)\Sigma_0 = \sum(1-z_i)(x-\mu_0)^T(x-\mu_0)$

$\Rightarrow \Sigma_0 = \frac{\sum(1-z_i)(x-\mu_0)^T(x-\mu_0)}{\sum(1-z_i)}$

- incomplete data

remind that w_i is posterior ： $\frac{P(z, x \mid \theta)}{\sum_z P(z, x \mid \theta)} = \frac{P(x \mid z, \theta)P(z \mid \theta)}{P(x \mid \theta)}$

故 $w_{i,c_0} = \frac{\lambda P_0}{\lambda P_0 + (1-\lambda)P_1} = \frac{\lambda e^{\frac{-(x_i-\mu_0)^2}{2\sigma_0^2}}}{\lambda e^{\frac{-(x_i-\mu_0)^2}{2\sigma_0^2}} + (1-\lambda)e^{\frac{-(x_i-\mu_1)^2}{2\sigma_1^2}}}$

$w_{i,c_1} = \frac{(1-\lambda)P_1}{\lambda P_0 + (1-\lambda)P_1}$

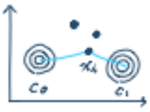
EM algo :

初始化 $\theta(\mu_0, \mu_1, \Sigma_0, \Sigma_1, \lambda_0, \overset{(1-\lambda)}{\lambda_1})$

E step：更新 posterior w_i　　M step：MLE 找参数 θ

ㄆ K-means clustering

- Specific case of GMM



$k=2$

令 C_0, C_1 為 isotropic Gaussian

E step : 計算每點到 C_0, C_1 的距離

近 C_0 則 $w_i=0$，近 C_1 則 $w_i=1$

故將 w_i 改成 z，即可用

complete data 的 EM algo 計算

M step :

likelihood $\xrightarrow{\text{取}\log}$ J $\xrightarrow[\text{MLE}]{\text{偏微}}$ 更新 gaussian 分布之 θ

- K-means vs. GMM

K-means 假設 isotropic gaussian 分佈

所以可以用距離分類

⇒ 得到 z，可改用 complete data

方式作 EM

Appendix : Why EM works

- 推導

$\ln P(X|\theta)$

$= \ln \sum_z P(X, z|\theta)$

$= \ln \sum_z q(z) \frac{P(X,z|\theta)}{q(z)}$

$= \ln \left[\mathop{E}_q \frac{P(X,z|\theta)}{q(z)} \right]$

$\geq \mathop{E}_q \left[\ln \frac{P(X,z|\theta)}{q(z)} \right]$ ∵ 

$= \mathop{E}_q \left[\ln \frac{P(z|X,\theta)P(X|\theta)}{q(z)} \right]$

要 maximum 的　relative entropy (KL divergence)

$= \sum q(z) \ln P(X|\theta) - \sum q(z) \ln \frac{q(z)}{P(z|X,\theta)}$　$= KL(q \| P)$

會 ≤ 0，所以欲最大化 lower bound 要將此項化為 0

- 因 $q(z)$ 必 ≥ 0，故將 $\ln \frac{q(z)}{P(z|X,\theta)}$ 化為 0

故 $\frac{q(z)}{P(z|X,\theta)} = 1 \Rightarrow q(z) = P(z|X,\theta)$

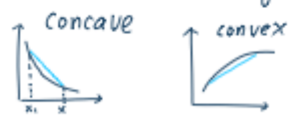
回顧 $w_i = \frac{P(z_i=C_0, x_i|\theta)}{P(z_i=C_0, x_i|\theta) + P(z_i=C_1, x_i|\theta)}$

$= \frac{P(z_i, x|\theta)}{\sum_i P(z_i, x|\theta)}$　$= P(x|\theta)$

$= P(z|x,\theta)$

發現 $q(z)$ 即 w_i

- Jensen's Inequality



$E[f(x)] > f(E(x))$　　$f(E(x)) > E[f(x)]$

- $J = \ln P(X|P_0, P_1, w_i, \lambda)$

$= \sum_{i=1}^{n} w_i [\ln\lambda + x_i \ln P_0 + (1-x_i)\ln(1-P_0)]$

$+ \sum_{i=1}^{n} (1-w_i)[\ln(1-\lambda) + x_i P_i + (1-x_i)\ln(1-P_1)]$

$= \sum_i w_i P(x|P_0, P_1, \lambda)$

$= \sum_i q(z) P(X|\theta)$