DS215: Assignment 3

Aneesh Panchal

06-18-01-10-12-24-1-25223

Sol^n 1 K means clustering is defined as,

$\min \sum_{i=1}^{n} \sum_{r=1}^{K} z_{ir} \|x_i - \mu_r\|^2$ where $x_i \sim$ data pts, $\mu_r \sim$ cluster centroids

$z_{ir} = \begin{cases} 1 & \text{if } r = \text{argmin}_j \|x_i - \mu_j\|^2 \\ 0 & \text{else} \end{cases}$

Update step : $\mu_r = \dfrac{\sum^n z_{ir} x_i}{\sum^n z_{ir}}$

Aneesh 25223

Gaussian Mixture Model is,

$p(x|y=i) \sim N(\mu_i, \sigma^2 I)$

$p(x) = \sum_i p(x|y=i) P(y=i)$

we want to learn : $\theta = [\mu_1, \mu_2, \dots, \mu_K]$

Expectation step,, (assume we have given θ^{t-1})

$R_{i,j}^{t-1} = P(y_j = i | x_j, \theta^{t-1}) \propto P(x_j | y_j = i, \theta^{t-1}) P(y_j = i)$

$\propto \exp\left\{\frac{-1}{2\sigma^2} \|x_j - \mu_i^{t-1}\|^2\right\} \underbrace{\pi_i^{t-1}}_{P(y_j=i)}$ ——— estimated known values

$\underset{(say)}{Q(\theta^t|\theta^{t-1})} = \sum_{j=1}^{n} \sum_{i=1}^{K} \underbrace{P(y_j = i | x_j, \theta^{t-1})}_{R_{i,j}^{t-1}} \log P(x_j, y_j = i | \theta^t)$

↳ already calculated in expectation step

which can be rewritten as,

$Q(\theta^t|\theta^{t-1}) = \sum_{j=1}^{n} \sum_{i=1}^{K} R_{i,j}^{t-1} [\log P(x_j | y_j = i, \theta^t) + \log P(y_j = i | \theta^t)]$

Maximization step,,

$\frac{\partial}{\partial \mu_i^t} Q(\mu_i^t | \theta^{t-1}) = \sum_{j=1}^{n} R_{i,j}^{t-1} (x_n - \mu_i^t) = 0$, because $Q(\mu_i^t|\theta^{t-1}) \propto \sum_{j=1}^{n} \sum_{i=1}^{K} R_{i,j}^{t-1} \left(\frac{-1}{2\sigma^2} \|x_j - \mu_i^t\|^2\right)$

$\mu_i^t = \sum_{j=1}^{n} w_j x_j$ where $w_j = \dfrac{P(y_j = i | x_j, \theta^{t-1})}{\sum_{m=1}^{n} P(y_m = i | x_m, \theta^{t-1})}$ which is equivalent to z_{ir} of the K means clustering distrib^n $N(\mu_i, \sigma^2 I)$ becomes peaky at μ_i & hence

Now, as $\sigma \to 0$, density [redacted] $P(y_j = i \mid x_j, \theta^{t-1})$ turns into indicator which is similar to z_{in} defined in K means clustering. (matches almost equally)

Hence, as $\sigma \to 0$, EM algorithm to estimate GMM parameter coincides as with K means clustering.

Sol^n 2. a students got grade A with $P(A) = 1/2$

b students got grade B with $P(B) = \mu$

c students got grade C with $P(C) = 2\mu$

d students got grade D with $P(D) = 1/2 - 3\mu$

(a) If all values are given, ML estimate is given by,

Let A, B, C, D follow multinomial distribution,

$$p = \frac{(a+b+c+d)!}{a!\,b!\,c!\,d!}\left(\frac{1}{2}\right)^a (\mu)^b (2\mu)^c \left(\frac{1}{2} - 3\mu\right)^d$$

$$\ln p = k - a\log 2 + b\log\mu + c\log 2\mu + d\log(1/2 - 3\mu)$$

we have to maximize log likelihood,

$$\frac{\partial \ln p}{\partial \mu} = \frac{b}{\mu} + \frac{c}{\mu} - \frac{3d}{\frac{1}{2} - 3\mu} = 0$$

solving which we get, $\mu = \dfrac{b+c}{6(b+c+d)}$

Aneesh
25223

(b) we are given $h = a+b$ & $\hat{\mu}$, then E step follows,

$$a = E[A \mid \mu] = \frac{1/2}{1/2 + \hat{\mu}}\,h$$

$$b = E[B \mid \mu] = \frac{\hat{\mu}}{1/2 + \hat{\mu}}\,h$$

Now we are at M step, we are given $\hat{a}, \hat{b}$ calculated at E step,

ML estimate of μ can be taken from part (a) as an iterate, then we have

$$\mu^{(t+1)} = \frac{\hat{b}^{(t)} + c}{6(\hat{b}^{(t)} + c + d)}$$

Now, iteration will start with $\mu^{(0)}$ & calculate $\hat{a}$ & $\hat{b}$ using E-step

find $\mu^{(t+1)}$ using thus found $\hat{a}$ & $\hat{b}$ using M-step

repeat until convergence.

Sol^n 3.(a) Given number of data points, $n = 30$

we have to find best fit line via linear regression.

Let best fit line be $y = mx + c$ & use MSE as error we have,

$$E(m, c) = \frac{1}{n} \sum_{i=1}^{n} (m x_i + c - y_i)^2$$

Now we have to minimize $E(m,c)$,

$$\frac{\partial E}{\partial m} = \frac{2}{n} \sum_i x_i (m x_i + c - y_i) = 2\left[m \frac{\sum x_i^2}{n} + c \frac{\sum x_i}{n} - \frac{\sum x_i y_i}{n}\right] \text{ ——— (i)}$$

$$\frac{\partial E}{\partial c} = \frac{2}{n} \sum_i (m x_i + c - y_i) = 2\left[m \frac{\sum x_i}{n} + c \frac{\sum 1}{n} - \frac{\sum y_i}{n}\right] \text{ ——— (ii)}$$

As we know,

$$Cov(x,y) = \frac{\sum x_i y_i}{n} - \bar{x}\bar{y} = \rho_{xy} \sigma_x \sigma_y \quad \& \quad Var(x) = \frac{\sum x_i^2}{n} - \bar{x}^2 = \sigma_x^2$$

given values are, $\bar{x} = 4$, $\bar{y} = 3$, $\sigma_x^2 = 2.25$, $\sigma_y^2 = 0.25$, $\rho_{xy} = 0.7$

Substituting values in (i) & (ii) we get,

$$2[m(\sigma_x^2 + \bar{x}^2) + c\bar{x} - (\rho_{xy}\sigma_x\sigma_y + \bar{x}\bar{y})] = 0 \text{ ——— (iii)}$$

$$2[m\bar{x} + c - \bar{y}] = 0 \text{ ——— (iv)}$$

$$\begin{bmatrix} \sigma_x^2 + \bar{x}^2 & \bar{x} \\ \bar{x} & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} \rho_{xy}\sigma_x\sigma_y + \bar{x}\bar{y} \\ \bar{y} \end{bmatrix}$$

Aneesh
25223

$$\begin{bmatrix} 18.25 & 4 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} m \\ c \end{bmatrix} = \begin{bmatrix} 12.525 \\ 3 \end{bmatrix}$$

Solving which we get, $m = 7/30$ & $c = 31/15$

ie best fit line is, $y = \frac{7}{30} x + \frac{31}{15}$

(b) Substituting value $\bar{x} = 4$ in best fit line we get,

$$y = \frac{7}{\cancel{30}_{15}} \times \cancel{4}^{2} + \frac{31}{15} = \frac{\cancel{45}^{3}}{\cancel{15}_{1}} = 3 = \bar{y}$$

hence, best fit line pass through $(\bar{x}, \bar{y})$

it is also obvious from eq^n (iv) that best fit line pass through $(\bar{x}, \bar{y})$.

Sol^n 4. $y_i \sim N(\beta_0 + x_i^T \beta, \sigma^2)$, $i = 1, 2, \ldots, N$ & $\beta \in \mathbb{R}^p$

$\beta_j \sim N(0, \tau^2)$, $j = 1, 2, \ldots, p$ are independent

Punk

Assume σ^2 & τ^2 are known.

from Bayes theorem we have,

$$P(\beta|y) = \frac{P(y|\beta)\,P(\beta)}{P(y)}$$

According to assumption we have,

$$P(y|\beta) = \frac{1}{(2\pi)^{N/2}\sigma^N} \exp\left\{\frac{-1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2\right\}$$

$$P(\beta) = \frac{1}{(2\pi)^{p/2}\tau^p} \exp\left\{\frac{-1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2\right\}$$

Substituting values of $P(y|\beta)$ & $P(\beta)$ we get,

$$\log_e P(\beta|y) = \ln P(\beta|y) = k - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 - \frac{1}{2\tau^2}\sum_{j=1}^{p}\beta_j^2 + c$$

where $k = -\ln((2\pi)^{N/2}\sigma^N) - \ln((2\pi)^{p/2}\tau^p)$ — constant

$c = -\ln(P(y))$ — constant

let $K = -k - c$ we get,

$$-\ln P(\beta|y) = \frac{1}{2\sigma^2}\left[\sum_{i=1}^{N}\left[y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j\right]^2 + \frac{\sigma^2}{\tau^2}\sum_{j=1}^{p}\beta_j^2\right] + K \qquad (\text{we know } \lambda = \sigma^2/\tau^2)$$

$$-\ln P(\beta|y) \propto \left[\sum_{i=1}^{N}\left[y_i - \beta_0 - \sum_j x_{ij}\beta_j\right]^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right]$$

Anesh
25223

Hence, proved.

Sol^n 5. (a) (a) & (b) are appropriate to use in classification

Reasoning: -ve side of $yF(x)$ determines misclassified data

+ve side of $yF(x)$ determines correctly classified data

hence, error (L) value at -ve side must be +ve & nearly 0 at +ve side.

example: let $y_i = 1$ & $F(x)$ ~ predicted value be $1 \Rightarrow yF(x) = 1$ ~ correctly classified

predicted value be $-1 \Rightarrow yF(x) = -1$ ~ misclassified

$y_i = -1$ & $F(x)$ ~ predicted value be $1 \Rightarrow yF(x) = -1$ ~ misclassified

predicted value be $-1 \Rightarrow yF(x) = 1$ ~ correctly classified

(c) is not appropriate because there is very less error (penalty) for extremely misclassified data i.e. very -ve $yF(x)$

(d) & (e) are not appropriate because they penalize correctly classified data as well

Punk

In general, conditions for L to satisfy are,

(i) L should approximate the 0-1 loss ~ 1 for misclassified & 0 for correctly classified

(ii) L should be non-increasing function of $yF(x)$

(b) (b) is more robust to outliers. For outliers, $yF(x)$ is often very -ve. In (a), outliers are heavily penalized. So, resulting classifier is largely affected by outliers. While in (b), penalty is upper bounded by 1. So, (b) is very less likely to be affected by outliers & hence (b) is more robust.

(c) given $F(x) = w_0 + \sum_{j=1}^{d} w_j x_j$ & $L(yF(x)) = (1 + \exp(yF(x)))^{-1}$

for update, we need to min $\sum_i L(y^i F(x^i)) = \sum_i \frac{1}{1 + \exp(y^i (w_0 + \sum_j^d w_j \cdot x_j^i))}$

$$\frac{\partial}{\partial w_0} \sum_i L(y^i F(x^i)) = -\sum_i \frac{y^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

Ansh
25223

$$\frac{\partial}{\partial w_k} \sum_i L(y^i F(x^i)) = -\sum_i \frac{y^i x_k^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

Hence using gradient descent, update rules are as follows,

$$w_0^{t+1} = w_0^t - \alpha \frac{\partial}{\partial w_0} \sum_i L(y^i F(x^i)) = w_0^t + \alpha \sum_i \frac{y^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

& for other weights, $\forall k = 1, 2, \ldots, d$

$$w_k^{t+1} = w_k^t - \alpha \frac{\partial}{\partial w_k} \sum_i L(y^i F(x^i)) = w_k^t + \alpha \sum_i \frac{y^i x_k^i \exp(y^i F(x^i))}{(1 + \exp(y^i F(x^i)))^2}$$

which are required update rules.

NOTE : we can't directly minimize $L(yF(x))$ due to bias term w_0

hence we need to minimize $\sum_i L(y^i F(x^i))$ for w_0 & w_k $\forall k = 1, 2, \ldots, d$

Punk