Lecture 12

Clustering and the EM algorithm

X is a sequence of i.i.d. RV

$P\{X_i = 0\} = P\{X_i = 1\} = 1/2$

W is also i.i.d given X

$p(y_i / x_i) \sim N(\mu_{x_i}, \sigma_{x_i}^2)$

γ_0

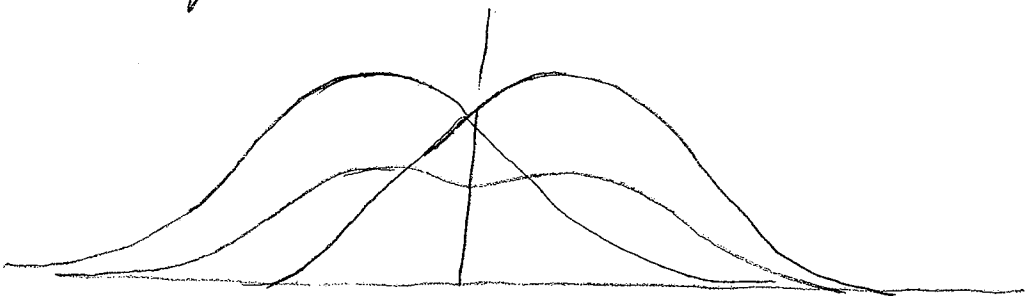
(μ_0, σ_0^2) – parameter of class 0

γ_1

(μ_1, σ_1^2) – parameter of class 1.

Question: How do we estimate cluster parameters directly from y?

Histogram



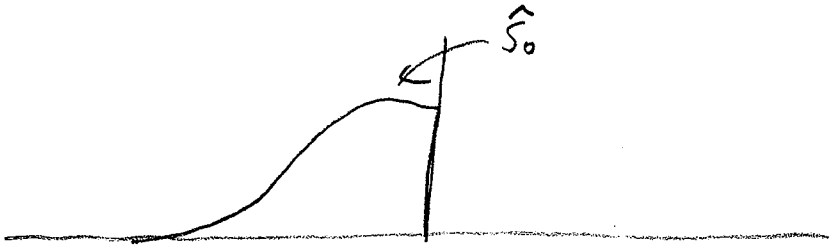
If we know X the problem is easy

$$S_0 = \{i : X_i = 0\}$$
$$S_1 = \{i : X_i = 1\}$$

$$\hat{\mu}_0 = \frac{1}{|S_0|} \sum_{i \in S_0} y_i \qquad \hat{\sigma}_0^2 = \frac{1}{|S_0|} \sum_{i \in S_0} (y_i - \hat{\mu}_0)^2$$

What if we estimate S_0?

$$\hat{S}_0 = \{ i : Y_i \leq 0 \}$$



$$\frac{1}{|\hat{S}_0|} \sum_{i \in \hat{S}_0} Y_i < \frac{1}{|S_1|} \sum_{i \in S_0} Y_i$$

This is called the "incomplete data" problem because we are missing X.

$$\hat{\theta}_{ML} = \arg\max_{\theta} p(y|\theta)$$

$$= \arg\max_{\theta} \left\{ \int p(y|x,\theta) p(x|\theta)\, dx \right\}$$

↑ difficult to do

New approach:

$$\log p(y|\theta') = \int \log p(y|\theta')\, p(x|y,\theta)\, dx$$

$$p(y,x|\theta') = p(x|y\theta')p(y|\theta')$$

$$= \int \log \left\{ \frac{p(y,x|\theta')}{p(x|y,\theta')} \right\} p(x|y,\theta)\, dx$$

$$= \int \log p(y,x|\theta')\, p(x|y,\theta)\, dx$$

$$- \int \log p(x|y,\theta')\, p(x|y,\theta)\, dx$$

Define

$$Q(\theta',\theta) = \int \log p(y,x|\theta')\, p(x|y,\theta)\, dx$$

$$H(\theta',\theta) = -\int \log p(x|y,\theta')\, p(x|y,\theta)\, dx$$

$$\log p(y|\theta') = Q(\theta',\theta) + H(\theta',\theta)$$

Theorem $\forall \theta$

$$\theta = \underset{\theta'}{\operatorname{argmin}}\ H(\theta', \theta)$$

proof

W.L.O.G. use $p(x|\theta)$ in place of $p(x|y, \theta)$

$$0 = \log\left\{\int p(x|\theta')\,dx\right\}$$

$$= \log\left\{\int \frac{p(x|\theta')}{p(x|\theta)}\, p(x|\theta)\,dx\right\}$$

(Jensen's inequality)

$$\geq \int \log\left\{\frac{p(x|\theta')}{p(x|\theta)}\right\} p(x|\theta)\,dx$$

$$= \int \log p(x|\theta')\, p(x|\theta)\,dx$$

$$- \int \log p(x|\theta)\, p(x|\theta)\,dx$$

$$0 \geq -H(\theta', \theta) + H(\theta, \theta)$$

$$H(\theta', \theta) \geq H(\theta, \theta)$$

Intuition

$$\theta = \arg\min_{\theta'} H(\theta', \theta)$$

$$= \arg\min_{\theta'} -\int \log p(x|\theta')\, p(x|\theta)\, dx$$

$$= \arg\min_{\theta'} -E_\theta\left[\log p(x|\theta')\right]$$

$$= \arg\min_{\theta'} \left\{ \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \log p(x_i|\theta') \right\}$$

where x_i are i.i.d. r.v.s with distribution $p(x|\theta)$

$$= \lim_{N\to\infty} \arg\min_{\theta'} \left\{ \log p(x_1, \ldots x_N|\theta') \right\}$$

$\Rightarrow$ ML estimate of θ is consistent

Theorem $Q(\theta', \theta) > Q(\theta, \theta) \Rightarrow p(y|\theta') > p(y|\theta)$

proof

$$\log p(y|\theta') = Q(\theta', \theta) + H(\theta', \theta)$$

$$> Q(\theta, \theta) + H(\theta', \theta)$$

$$\geq Q(\theta, \theta) + H(\theta, \theta)$$

$$= \log p(y|\theta)$$

New algorithm:

Expectation Maximization (EM)

Baum Welch

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta, \theta^{(k)})$$

$$= \arg\max_{\theta} \int \log p(y, x|\theta)\, p(x|y, \theta^{(k)})\, dx$$

maximization — expectation

Example

$\{x_i\}_{i=1}^{N}$ i.i.d with $P\{x_i=0\} = P\{x_i=1\} = 1/2$

$\{y_i\}_{i=1}^{N}$ conditionally i.i.d, with

$p(y_i | x_i) \sim N(\mu_{x_i}, \underbrace{\sigma^2_{x_i}}_{\gamma_{x_i}})$

$\gamma_0 = (\mu_0, \gamma_0) \rightarrow$ class 0

$\gamma_1 = (\mu_1, \gamma_1) \rightarrow$ class 1

$\log p(y, x | \theta) = \log p(y | x, \theta) + \log p(x | \theta)$

↑ not a function of θ.

$$\log p(y, x \mid \theta) = \sum_{i=1}^{N} \log p(y_i \mid \gamma_0)\,\delta(x_i) + \sum_{i=1}^{N} \log p(y_i \mid \gamma_1)\,\delta(x_i - 1)) - N \log 2$$

$$Q(\theta', \theta) = \int \log p(y, x \mid \theta')\, p(x \mid y, \theta)\, dx$$

$$= E\left[\log p(Y, X \mid \theta') \mid Y = y, \theta\right]$$

$$= \sum_{i=1}^{N} \log p(y_i \mid \gamma_0)\, E\left[\delta(x_i) \mid Y = y, \theta\right] + \sum_{i=1}^{N} \log p(y_i \mid \gamma_1)\, E\left[\delta(x_i - 1) \mid Y = y, \theta\right] - N \log 2$$

$$E\left[\delta(x_i) \mid Y = y, \theta\right] = p\{X_i = 0 \mid Y = y, \theta\} \quad p(x_i = 0 \mid y, \theta)$$

$$p(x_i \mid y, \theta) = \frac{p(y_i \mid x_i, \theta)\, p(x_i)}{\sum_{\tilde{x}_i} p(y_i \mid \tilde{x}_i, \theta)\, p(\tilde{x}_i)}$$

↑ easy to compute

(E-step) $\triangleq f(x_i \mid y_i, \theta)$

Lecture 29

Since

$$\log p(y_i \mid \theta) = -\frac{1}{2\sigma^2}(y_i - \mu)^2 - \frac{1}{2}\log(2\pi\sigma^2)$$

$$Q(\theta', \theta) = \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma_0^2}(y_i - \mu_0')^2 - \frac{1}{2}\log(2\pi\gamma_0') \right\} f(0 \mid y_i, \theta)$$

$$+ \sum_{i=1}^{N} \left\{ -\frac{1}{2\sigma_1^2}(y_i - \mu_1')^2 - \frac{1}{2}\log(2\pi\gamma_1') \right\} f(1 \mid y_i, \theta)$$

$$- N \log 2$$

1st term only a function of (μ_0, σ_0^2)
2nd term only a function of (μ_1, σ_1^2)

$$\hat{\theta} = \underset{\theta'}{\operatorname{argmax}}\, Q(\theta', \theta)$$

Answer:

$$\hat{N}_0 = \sum_{i=1}^{N} f(0 \mid y_i, \theta) \rightarrow \text{average \# of 0 points}$$

$$\hat{\mu}_0 = \frac{1}{\hat{N}_0} \sum_{i=1}^{N} y_i f(0 \mid y_i, \theta) \rightarrow \text{weighted mean}$$

$$\hat{\gamma}_0 = \frac{1}{\hat{N}_0} \sum_{i=1}^{N} (y_i - \hat{\mu}_0)^2 f(0 \mid y_i, \theta)$$

$\rightarrow$ weighted variance

$$\hat{\nu}_1 = \sum_{i=1}^{N} f(1/y_i, \theta)$$

$$\hat{\mu}_1 = \frac{1}{\hat{\nu}_1} \sum_{i=1}^{N} y_i f(1/y_i \theta)$$

$$\hat{\gamma}_1 = \frac{1}{\hat{\nu}_1} \sum_{i=1}^{N} (y_i - \hat{\mu}_1)^2 f(1/y_i \theta)$$

Repeat this procedure until you converge!

Convergence of EM algorithm

Assume

1) The ML estimate
$$\hat{\theta} = \arg\max_{\theta \in \Omega} p(y|\theta) \text{ exists.}$$

2) $Q(\theta', \theta)$ and $H(\theta', \theta)$ are well defined, continuous and differentiable functions of θ, and θ'

3) θ finite dimensional

4) $\hat{\theta}$ falls in an open set contained in Ω

Theorem If $\theta^* = \lim_{k \to \infty} \theta^{(k)}$ exists,

then $$\nabla_\theta p(y|\theta)\Big|_{\theta=\theta^*} = 0$$

proof

1) $$\nabla_\theta Q(\theta, \theta^*)\Big|_{\theta=\theta^*} = 0$$
{ steady state condition

2) $$\nabla_\theta H(\theta, \theta^*)\Big|_{\theta=\theta^*} = 0$$
{ result of theorem

3) $$\nabla_\theta \log p(y|\theta)\Big|_{\theta=\theta^*}$$
$$= \nabla_\theta Q(\theta, \theta^*)\Big|_{\theta=\theta^*} + \nabla_\theta H(\theta, \theta^*)\Big|_{\theta=\theta^*}$$
$$= 0 + 0 = 0$$

Theorem If $p(y|\theta)$ is a strictly concave function of θ then and Ω is compact and convex

$$\theta^* = \lim_{n \to \infty} \theta^{(n)} = \theta_{ML}$$

proof

1) $\lim_{n \to \infty} \theta^{(n)}$ exists

2) $\nabla_\theta p(y|\theta)\big|_{\theta=\theta^*} = 0$

See J. Wu, "On the Convergence of the EM Algorithm" vol. 11, no. 1, pp. 95-103, Annals of Statistics, 1983

R. Redner and H. Walker, "Mixture Densities, Maximum Likelihood and The EM Algorithm", SIAM Review vol. 26, no. 2, April 1984.

Lecture 30

Example

as before, but

(μ_0, σ_0^2)
(μ_1, σ_1^2) } are known

$$P\{X_i = 0\} = \pi_0 \qquad P\{X_i = 1\} = 1 - \pi_0$$

$p(y_i \mid x_i)$ known $\quad \theta = \pi_0$

$$p(y \mid \theta) = \prod_{i=1}^{N} \left(\theta p(y \mid 0) + (1-\theta) p(y \mid 1) \right)$$

$$\log p(y \mid \theta) = \sum_{i=1}^{N} \log \left(\theta p(y \mid 0) + (1-\theta) p(y \mid 1) \right)$$

$\log p(y \mid \theta)$ is a concave function of θ.

$\Rightarrow$ EM algorithm converges to θ_{ML}

$$\log p(y, x \mid \theta) = \underbrace{\log p(y \mid x, \theta)}_{\text{does not depend on } \theta} + \log p(x \mid \theta)$$

Let $K = \#\, x_i = 0 = \sum_{i=1}^{N} \delta(x_i)$

$$\log p(x \mid \theta) = \log \left(\theta^K (1-\theta)^{N-K} \right)$$

$$\log p(x|\theta) = K \log\theta + (N-K)\log(1-\theta)$$

$$= K(\log\theta - \log(1-\theta)) + N\log(1-\theta)$$

$$Q(\theta';\theta) = \int \log p(y,x|\theta')\, p(x|y,\theta)\, dx$$

$$= E\left[K(\log\theta' - \log(1-\theta')) + C \,\middle|\, Y=y,\theta\right]$$

$$= E[K|Y=y,\theta]\,(\log\theta' - \log(1-\theta')) + N\log(1-\theta)$$

$$E[K|Y=y,\theta] = E\left[\sum_{i=1}^{N}\delta(x_i)\,|\,Y=y,\theta\right]$$

$$= \sum_{i=1}^{N} E[\delta(x_i)|Y=y,\theta]$$

$$= \sum_{i=1}^{N} p\{x_i=0 \,|\, y_i=y_i,\theta\}$$

$$= \sum_{i=1}^{N}\left\{\frac{p(y_i|x_i=0)\,\theta}{p(y_i|x_i=0)\,\theta + p(y_i|x_i=1)(1-\theta)}\right.$$ ⇒ easy to compute

$$= \bar{K} \leftarrow \text{expected number of } x_i\text{'s} = 0 \text{ given } y \text{ and } \theta$$

$$\underset{\theta'}{\operatorname{argmax}}\, Q(\theta', \theta)$$

$$= \underset{\theta'}{\operatorname{argmax}}\, \bar{K} \left(\log \theta' - \log (1-\theta') \right)$$

$$= \frac{\bar{K}}{N} \leftarrow \text{same as 1st homework}$$

EM algorithm

$$\theta^{(k+1)} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{p(y_i | x_i = 0)\theta}{p(y_i | x_i = 0)\theta + p(y_i | x_i = 1)(1-\theta)} \right\}$$

Exercise

Derive EM algorithm when

$$p(y_i | x_i) \sim N(\mu_{x_i}, \sigma_{x_i}^2)$$

$$P\{x_i = 0\} = \pi_0 \quad P\{x_i = 1\} = 1 - \pi_0$$

$$\theta = (\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$$