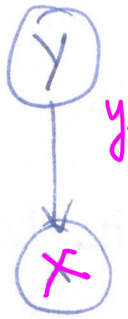
We are uncertain about the hidden variables. Y

We cannot observe them.

We capture this uncertainty by defining a posterior distribution over $Y = y_i$ given every training example $X = x_i$.



$P[Y = y]$

$P[Y = y_i] =$ prior

y_i is the sampled value

$P[Y = y_i \mid X = x_i]$: posterior

$P[Y = y_i]$

$X = x_i$

$P[X = x_i \mid Y = y_i]$: likelihood

Generating an observation by sampling the likelihood function

The matrix Q captures the posterior distribution

$G(Q, \underline{\theta})$

$$Q_{i,y} = P[Y = y_i \mid X = x_i]$$

responsibility that the y_i^{th} Gaussian takes for the feature x_i

The surrogate function $G(Q, \underline{\theta})$ is formulated as

$$G(Q, \underline{\theta}) = \overbrace{F(Q, \underline{\theta})}^{\mathrm{I}} - \underbrace{\sum_{i=1}^{m}}_{\text{for every row } i} \underbrace{\overbrace{\sum_{y=1}^{k} Q_{i,y} \log Q_{i,y}}^{\mathrm{II}}}_{\text{Entropy of the cat distr}}$$

$$\mathrm{I} \quad \underbrace{F(Q, \underline{\theta})}_{\text{Expectation . . .}} = \sum_{i=1}^{m} \sum_{y_i=1}^{k} \underbrace{Q_{i,y_i}}_{\text{Posterior distr over } Y = y_i \text{ for } X = x_i} \underbrace{\log\left(P_{\theta}[\underbrace{X = x_i}_{\text{obs}}, \underbrace{Y = y_i}_{\text{latent}}]\right)}_{\text{likelihood of complete data : observed + hidden random variables}}$$

Maxim

$\therefore$ the function $F(Q, \underline{\theta})$ gives the expectation of the log likelihood of the complete data with respect to the posterior distribution over the latent (hidden) variable Y.

Data

Given Array

$a_1\ a_2\ a_3 \cdots a_d$

Distribution

$p_1\ p_2 \cdots p_d$

$\sum_i p_i = 1$

Expectation of the array w.r.t. the distribution

$$\sum_{i=1}^{d} p_i a_i$$

Complete data log likelihood

$$\begin{array}{c|c} & y=1 \quad y=2 \quad y=3 \quad \cdots \quad y_i \\ \hline x_i \downarrow & \begin{bmatrix} O & O & O & O \\ & \log\left(P_\theta[X=x_i, Y=y_i]\right) & & \\ & & & \end{bmatrix}_{m\times d} \end{array}$$

Posterior over the latent variable Y

$$\begin{array}{c|c} & y \longrightarrow \\ \hline x_i \downarrow & \begin{bmatrix} P[Y=y_i \mid X=x_i] \\ \theta_{iy} \quad \theta_{i\cdot} \\ \end{bmatrix}_{m\times d} \end{array}$$

Mean of a data array $\equiv$ Expectation w.r.t Uniform distr

$$\sum_i \frac{1}{d} a_i = \frac{1}{d}\sum a_i$$

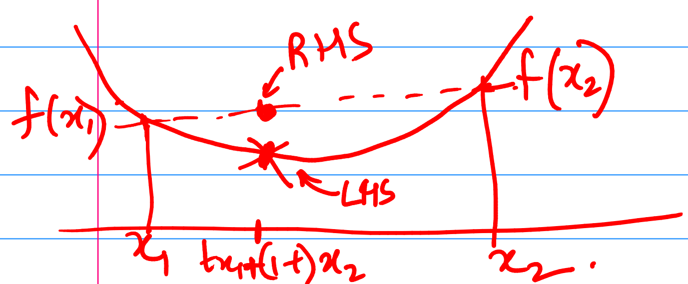
$$Q = \begin{array}{c|c} & 1 \quad 2 \quad y \longrightarrow 3 \quad 4 \quad \cdot \\ \hline x_i \downarrow & \begin{bmatrix} x_1: & p_1 & p_2 & p_3 & = & = \\ x_2: & = & = & = & = & = \\ x & & & & & \end{bmatrix} \end{array} \quad \begin{array}{l} \Sigma = 1. \\ \Sigma = 1. \end{array}$$

Entropy

$$\sum_i p_i \log \frac{1}{p_i}$$

Jensen's Inequality

$$\underbrace{f\left(t x_1 + (1-t)x_2\right)}_{LHS} \leq \underbrace{t f(x_1) + (1-t) f(x_2)}_{RHS}$$



If f is a convex function

likelihood of x_i. $Y = y_i$ $x_i \sim P(X \mid Y = y_i)$. $(x_i \; y_i)$

Posterior x_i $P(Y = y \mid X = x_i)$

The $G(Q, \theta)$ can be simplified as:

$$Q = \begin{array}{c|c} & y \rightarrow \\ i \downarrow & \begin{bmatrix} \circ \circ \circ \circ \circ \\ Q_{i,y} \end{bmatrix} \end{array}$$

$$G(Q, \underline{\theta}) = \underbrace{\sum_{i=1}^{m} \sum_{y_i=1}^{k} Q_{i,y_i} \log P_{\underline{\theta}}[X = x_i, Y = y_i] - \sum_{i=1}^{m} \sum_{y_i=1}^{k} Q_{i,y_i} \log Q_{i,y_i}}_{F(Q,\underline{\theta})}$$

$Y = y_i$ $(Y = y)$

entropy posterior over latent variables

x_i

$y_i \equiv y$

$P[y \mid X = x_i]$

Since it is implied that the random variable Y takes the value y_i in the context of example x_i we write y instead of y_i

$$\log \frac{A}{B} = \log A - \log B$$

$$G(Q, \underline{\theta}) = \sum_{i=1}^{m} \sum_{y=1}^{k} Q_{i,y} \log \frac{P_{\underline{\theta}}[X = x_i, Y = y]}{Q_{i,y}}$$

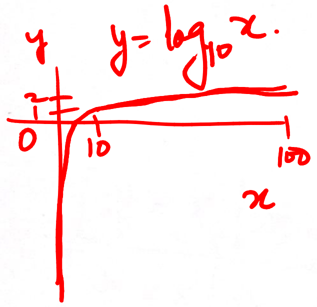
$Q_{i,y}$: responsibility taken by the y^{th} category for generating x_i

The upper bound of $\underline{G(Q, \underline{\theta})}$ is $\underline{L(\underline{\theta})}$.

$$G(Q, \underline{\theta}) = \sum_{i=1}^{m} \sum_{y=1}^{k} Q_{i,y} \underbrace{\log}_{f} \frac{P_{\underline{\theta}}[X = x_i, Y = y_i]}{Q_{i,y}}$$

$$\leq \sum_{i=1}^{m} \underbrace{\log}_{f} \left(\sum_{y=1}^{k} \cancel{Q_{i,y}} \frac{P_{\underline{\theta}}[X = x_i, Y = y_i]}{\cancel{Q_{i,y}}} \right)$$

Jensen's inequality



$$= \sum_{i=1}^{m} \log \left(\underbrace{\sum_{y=1}^{k} P_{\underline{\theta}}[X = x_i, Y = y_i]} \right)$$

Summing out the latent vars

$$= \sum_{i=1}^{m} \log \underline{P_{\underline{\theta}}[X = x_i]}$$

observed data log likelihood.

$$= \underline{L(\underline{\theta})}$$

$$\therefore \boxed{G(Q, \underline{\theta}) \leq L(\underline{\theta})}$$

The maximization procedure involves alternating between two steps

t: step $\quad t = 0, 1, 2 \ldots$

equivalent steps

E-step I $\quad Q^{(t+1)} = \arg\max_{Q} G\left(Q, \theta^{(t)}\right) \quad\Big|\quad$ Compute $Q^{(t+1)}$ with elements

closed form $\quad Q^{(t+1)}_{i,y} = P_{\theta^{(t)}}\left[Y = y_i \,\middle|\, X = x_i\right]$

II M-step $\quad \theta^{(t+1)} = \arg\max_{\theta} G\left(Q^{(t+1)}, \theta\right) \quad\Big|\quad \theta^{(t+1)} = \arg\max_{\theta} F\left(Q^{t+1}, \theta\right)$

EM algorithm

$Y \rightarrow X$

We know the property of G function that $G(Q, \theta) \leq L(\theta)$

that means $\quad G\left(Q, \theta^{(t)}\right) \leq L\left(\theta^{(t)}\right)$

Now, if we substitute a Q matrix with elements

$$Q_{i,y} = P_{\theta}\left[Y = y_i \,\middle|\, X = x_i\right]$$

in $G(Q, \theta)$

$$G(Q, \theta) = \sum_{i=1}^{m} \sum_{y_i=1}^{k} \left(Q_{i,y} \log P_{\theta}\left[X = x_i, Y = y_i\right] - Q_{i,y} \log Q_{i,y} \right)$$

$y_i \equiv y$.

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} Q_{i,y} \log \frac{P_{\theta}\{X = x_i, Y = y_i]}{Q_{i,y}}$$

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} P_{\theta}\left[Y = y_i \,\middle|\, X = x_i\right] \log \frac{P_{\theta}\left[X = x_i, X = y_i\right]}{P_{\theta}\left[Y = y_i \,\middle|\, X = x_i\right]}$$

$$= \sum_{i=1}^{m} \sum_{y_i=1}^{k} P_{\theta}\left[Y = y_i \,\middle|\, X = x_i\right] \log P_{\theta}\left[X = x_i\right]$$

$$= \sum_{i=1}^{m} \log P_{\theta}\left[X = x_i\right] \sum_{y_i=1}^{k} P_{\theta}\left[Y = y_i \,\middle|\, X = x_i\right]$$

$$= \sum_{i=1}^{m} \log P_{\theta}\left[X = x_i\right] = L(\theta).$$