Final exam 12/28, close book

12/21 review, 习题课?

Lecture 12.

EM Algorithm.

- Maximal likelihood estimation with unknown parameters

Example: Let w be Bernoulli r.v., $w = \begin{cases} 1 & \text{w.p. } \delta \\ 0 & \text{w.p. } 1-\delta \end{cases}$, y (binary r.v.) be a noise observation of w.

* $P_{Y|W}(y|w) = \begin{cases} \varepsilon, & y \neq w \\ 1-\varepsilon, & y = w \end{cases}$

Let $\underline{w} = [w_1, \dots, w_n]^T$ be n iid samples of w.

$\underline{y} = [y_1, \dots, y_n]^T$

we do not observe $\underline{w}$, but observe $\underline{y}$.

Our goal is to estimate δ and ε.

ML: $P_Y(\underline{y}; \varepsilon, \delta) = \prod_{i=1}^{n} P_{Y_i}(y_i; \varepsilon, \delta) = \prod_{i=1}^{n} [P_{Y_i|W_i}(y_i | w_i=0; \varepsilon, \delta) \cdot P_{W_i}(0; \varepsilon, \delta) + P_{Y_i|W_i}(y_i | w_i=1; \varepsilon, \delta) \cdot P_{W_i}(1; \varepsilon, \delta)]$

(hidden/latent r.v.; Bayes rule.)

$(\varepsilon^*, \delta^*) = \arg\max_{\varepsilon, \delta} P_Y(\underline{y}; \varepsilon, \delta)$

- No close form solution for the optimal ε, δ.
- Algorithm to compute the optimal parameters with hidden variables $\Rightarrow$ EM Alg.

General setup:

- Assume complete data z, generated by $P_z(\cdot\,; x)$ (x: parameter we want to estimate.)
- Can only observe $y = g(z)$, observed data, g: deterministic function.
- In our example: $z = [\underline{\omega}, \underline{y}]$, $y = \underline{y}$, $x = [\varepsilon, \delta]$

The goal is to find $x^* = \arg\max_x P_Y(y;x) = \arg\max_x \log P_Y(y;x)$

Note that $\underbrace{P_Z(z;x)}_{P_{Z,Y}(z,y;x)} = \sum_y P_{Z|Y}(z|y;x)\cdot P_Y(y;x) = P_{Z|Y}(z|g(z);x)\cdot P_Y(g(z);x)$

$\Rightarrow$ let $y = g(z)$: $\log P_Y(y;x) = \log P_Z(z;x) - \log P_{Z|Y}(z|y;x)$

Do expectation of both sides over $P_{Z|Y}(z|y;x')$:

$$\text{LHS} = \sum_{z'} P_{Z|Y}(z'|y;x')\log P_Y(y;x) = \Big(\sum_{z'} P_{Z|Y}(z'|y;x')\Big)\cdot \log P_Y(y;x) = \log P_Y(y;x)$$

$$\text{RHS} = \mathbb{E}\left[\log P_Z(z;x)\,\middle|\, Y=y,\, x=x'\right] - \mathbb{E}\left[\log P_{Z|Y}(z|y;x)\,\middle|\, Y=y,\, X=x'\right]$$

$$= \underbrace{\sum_{z'} P_{Z|Y}(z'|y;x')\log P_Z(z';x)}_{\triangleq U(x,x')} - \underbrace{\left(-\sum_{z'} P_{Z|Y}(z'|y;x')\log P_{Z|Y}(z|y;x)\right)}_{\triangleq V(x,x')}$$

$$\Rightarrow \log P_Y(y;x) = U(x,x') + V(x,x'),\quad \forall x'$$

Lemma: $V(x,x') \geq V(x',x')$

Proof: $V(x,x') - V(x',x') = \sum_z P_{Z|Y}(z|y;x')\log\frac{P_{Z|Y}(z|y;x')}{P_{Z|Y}(z|y;x)} = D(P_{Z|Y;x'}\,\|\,P_{Z|Y;x}) \geq 0$ (K-L divergence) $\square$

$\Rightarrow$ If we can find $x \neq x'$ such that $U(x,x') \geq U(x',x')$, then

$$\log P_Y(y;x) = U(x,x') + V(x,x') \geq U(x',x') + V(x',x') = \log P_Y(y;x')$$

EM Algorithm:

(i) Initialization: Choose a $\hat{x}^{(0)}$

(ii) Repeat until convergence:

E-step: given the previous estimation $\hat{x}^{(n)}$, compute.

↑ expectation.

$$U(x, \hat{x}^{(n)}) = \mathbb{E}\left[\log P_Z(z; x) \mid Y = y;\ x = \hat{x}^{(n)}\right]$$

M-step: Find $\hat{x}^{(n+1)}$ maximizing $U(\cdot\,; \hat{x}^{(n)})$

↑ maximization

$$\hat{x}^{(n+1)} = \arg\max_x U(x; \hat{x}^{(n)})$$

$$\Rightarrow U(\hat{x}^{(n+1)}, \hat{x}^{(n)}) \geq U(\hat{x}^{(n)}, \hat{x}^{(n)})$$

- We can get a sequence of $\hat{x}^{(0)}, \hat{x}^{(1)}, \ldots$ such that

$$P_Y(y; \hat{x}^{(0)}) \leq P_Y(y; \hat{x}^{(1)}) \leq \cdots\cdots$$

since $P_Y(y; x) \leq 1$, it is a non-decreasing bound sequence

$\Rightarrow$ it must converges.

- EM Algorithm converges to a stationary point of the likelihood function, i.e.

Let x^* be the convergent point, then $\frac{\partial}{\partial x} P_Y(y; x)\Big|_{x=x^*} = 0$

EM Algorithm for mixture model.

Mixture model:

Assume data generated by the following process:

1. sample $l_i \in \{1, \dots, k\}$, $i = 1, \dots, m$ and $l_i \overset{iid}{\sim}$ multinomial $(\underline{\phi})$, $\underline{\phi} = [\phi_1 \dots \phi_k]$

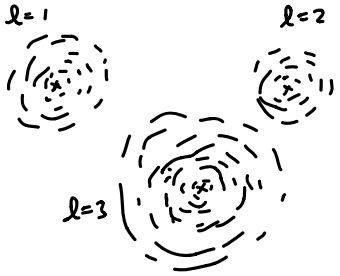
$$p(l_i = j) = \phi_j, \quad \forall j = 1 \dots k, \; i = 1, \dots, m \qquad \sum_{i=1}^{k} \phi_i = 1$$

2. Sample observation y_i from some distribution $p(l_i, y_i)$

$$p(l_i, y_i) = p(l_i) \cdot p(y_i | l_i)$$

generating the y_i from l_i

- Mixture Gaussian model: $p(y_i | l_i = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$



- In our notation, $x = [\underline{\phi}, \underline{\mu}, \underline{\Sigma}]$: $\begin{cases} \underline{\phi} = [\phi_1, \dots, \phi_k] \\ \underline{\mu} = [\mu_1, \dots, \mu_k] \\ \underline{\Sigma} = [\Sigma_1, \dots, \Sigma_k] \end{cases}$

$$z = [l_1, \dots, l_m, y_1, \dots, y_m]$$

$$y = [y_1, \dots, y_m]$$

The EM Algorithm: $(\hat{\underline{\phi}}^{(n)}, \hat{\underline{\mu}}^{(n)}, \hat{\underline{\Sigma}}^{(n)})$

E-step: $U(x, \hat{x}^{(n)}) = \mathbb{E}_{z|Y}(\cdot | y; \hat{x}^{(n)}) [\log p_z(z; x) \mid Y = y; \hat{x}^{(n)}]$

$$= \sum_{i=1}^{m} \sum_{l_i=1}^{k} p(l_i | y_i; \hat{x}^{(n)}) \log p(l_i, y_i; x)$$

M-step: $\hat{x}^{(n+1)} = \underset{x}{\operatorname{argmax}} \; U(x; \hat{x}^{(n)})$

The Mixture Gaussian :

$$P(l_i = j \mid y_i ; \hat{x}^{(n)}) = \frac{P(y_i \mid l = j ; \hat{x}^{(n)}) \cdot P(l_i = j ; \hat{x}^{(n)})}{P(y_i ; \hat{x}^{(n)})} \triangleq w_{ij}$$

($P(y_i \mid l = j ; \hat{x}^{(n)}) \sim \mathcal{N}(\hat{\mu}_j^{(n)}, \hat{\Sigma}_j^{(n)})$, $P(l_i = j ; \hat{x}^{(n)}) = \hat{\phi}_j^{(n)}$)

$$P(y_i ; \hat{x}^{(n)}) = \sum_{j=1}^{k} P_{l,r}(y_i, l_i = j ; \hat{x}^{(n)})$$

$$P_{l,r}(y_i, l_i = j ; \hat{x}^{(n)}) = P(y_i \mid l = j ; \hat{x}^{(n)}) \cdot P(l_i = j ; \hat{x}^{(n)})$$

$\Rightarrow$ Compute $U(x ; \hat{x}^{(n)}) \quad \forall x$

find $x^* = \arg\max U(x ; \hat{x}^{(n)})$

M-step : updating parameters

$$\hat{\phi}_j^{(n+1)} = \frac{1}{m} \sum_{i=1}^{m} w_{ij}$$

$$\hat{\mu}_j^{(n+1)} = \frac{\sum_{i=1}^{m} w_{ij} \cdot y_i}{\sum_{i=1}^{m} w_{ij}}$$

$$\hat{\Sigma}_j^{(n+1)} = \frac{\sum_{i=1}^{m} w_{ij} \cdot (y_i - \hat{\mu}_j^{(n)})(y_i - \hat{\mu}_j^{(n)})^T}{\sum_{i=1}^{m} w_{ij}}$$

Minimal Mean square Error Estimator

X : prior distribution $P_X(x)$, $X \to \boxed{P_{Y|X}} \to Y$

$\hat{X}_{MMSE}(Y) \triangleq \underset{\hat{x}}{\operatorname{argmin}} \mathbb{E}[(X-\hat{x}(Y))^2]$ ← mean square error.

Thm: $\hat{X}_{MMSE}(Y) = \mathbb{E}[X|Y]$

Proof: for any $f(Y)$, $\mathbb{E}[(X-f(Y))^2] = \mathbb{E}[((X-\mathbb{E}[X|Y])+(\mathbb{E}[X|Y]-f(Y)))^2]$

$= \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + 2\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-f(Y))] + \mathbb{E}[(\mathbb{E}[X|Y]-f(Y))^2]$

$2\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-f(Y))] = 0$

$\mathbb{E}[(X-\mathbb{E}[X|Y])(\mathbb{E}[X|Y]-f(Y))] = \mathbb{E}[X\cdot\mathbb{E}[X|Y]] - \mathbb{E}[(\mathbb{E}[X|Y])^2] - \mathbb{E}[Xf(Y)] + \mathbb{E}[\mathbb{E}[X|Y]\cdot f(Y)]$

$\mathbb{E}[X\cdot\mathbb{E}[X|Y]] = \sum_{x',y} x' P_{X|Y}(x'|y)\cdot P_Y(y)\cdot\left(\sum_x x\cdot P_{X|Y}(x|y)\right)$

$= \sum_y P_Y(y)\cdot\left(\sum_x x\cdot P_{X|Y}(x|y)\right)^2 = \mathbb{E}[(\mathbb{E}[X|Y])^2]$

$\mathbb{E}[Xf(Y)] = \mathbb{E}[\mathbb{E}[X\cdot f(Y)|Y]] = \mathbb{E}[\mathbb{E}[X|Y]\cdot f(Y)]$

$= \mathbb{E}[X\cdot f(Y)]$

$= 0$

$\mathbb{E}[(X-f(Y))^2] = \mathbb{E}[(X-\mathbb{E}[X|Y])^2] + \mathbb{E}[(\mathbb{E}[X|Y]-f(Y))^2]$

$\geq \mathbb{E}[(X-\mathbb{E}[X|Y])^2]$

Example: $Y = X + Z$ (signal + noise) $\quad X \sim \mathcal{N}(0, \sigma_X^2)$, $Z \sim \mathcal{N}(0, \sigma_Z^2)$

PDF $f_{X|Y}(x|y) \sim \mathcal{N}\left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}y, \frac{\sigma_X^2\sigma_Z^2}{\sigma_X^2+\sigma_Z^2}\right)$

$\hat{X}_{MMSE}(y) = \frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}\cdot y$, $MSE = \frac{\sigma_X^2\cdot\sigma_Z^2}{\sigma_X^2+\sigma_Z^2} = \left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_Z^2}\right)\cdot\sigma_Z^2$

Remark: If cannot observe y, random guess $\Rightarrow MSE = \sigma_Z^2$

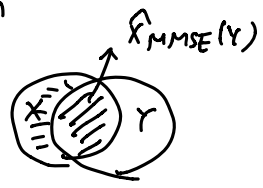
Orthogonality principle:

Thm: suppose $\hat{X}_{MMSE}$ is the MMSE estimator of X given Y, then

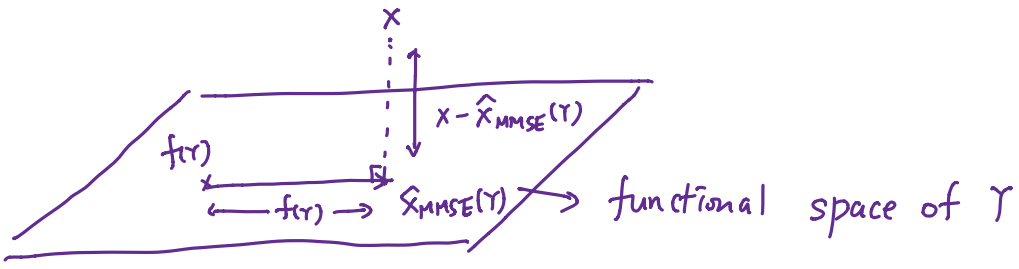
$$\mathbb{E}\left[(X - \hat{X}_{MMSE}(Y)) \cdot f(Y) \right] = 0 \quad , \forall\, f(Y)$$

($X - \hat{X}_{MMSE}(Y)$: estimation error; $f(Y)$: observation)

$\Rightarrow$ Estimation error and observation are uncorrelated.



Proof: $\mathbb{E}[(X - \mathbb{E}[X|Y]) \cdot f(Y)] = \mathbb{E}[X \cdot f(Y)] - \mathbb{E}[\mathbb{E}[X|Y] \cdot f(Y)] = 0$ □



Corollary: If $\hat{X}_{MMSE}(Y) = a \cdot Y$, then $a = \dfrac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}$

Proof: $\mathbb{E}[(X - a\,Y) \cdot Y] = 0 \Rightarrow a \cdot \mathbb{E}[Y^2] = \mathbb{E}[XY]$ □

($X - aY$: $\hat{X}_{MMSE}$; Y: $f(Y)$)

Linear regression: $X \in \mathbb{R}$, $\underline{Y} \in \mathbb{R}^m$, $P_{X\underline{Y}}$

Take $\hat{X}(\underline{Y}) = A \cdot \underline{Y}$

Goal: $A^* = \underset{A}{\operatorname{argmin}}\ \underset{P_{X\underline{Y}}}{\mathbb{E}}\left[(X - A\underline{Y})^2\right]$

$\Rightarrow A^* = K_{X\underline{Y}} \cdot K_{\underline{Y}\underline{Y}}^{-1}$, $K_{X\underline{Y}} = \underset{P_{X\underline{Y}}}{\mathbb{E}}[X\underline{Y}^T]$, $K_{\underline{Y}\underline{Y}} = \mathbb{E}[\underline{Y}\underline{Y}^T]$

- If X, Y are jointly Gaussian $\Rightarrow \hat{X}_{MMSE}(\underline{Y}) = A^* \cdot \underline{Y}$