Anton Sitkovets
212118048

Assignment 2

EECS 4404

(1) $E(\mu,\Sigma)=\sum_{i=1}^{N}\log p(X|\mu,\Sigma) = \frac{N}{2}\log|2\pi\Sigma^{-1}| - \frac{1}{2}\sum_i^N (x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$

$$\frac{\partial E}{\partial \mu} = -\frac{1}{2}\sum_i^N -2\,\Sigma^{-1}(x_i-\mu)=0$$

$$\sum_{i=1}^{N}\Sigma^{-1}x_i = \sum_i^N \Sigma^{-1}\mu$$

$$\hat{\mu}_{MLE} = \frac{1}{N}\sum_i^N x_i$$

To find Σ_{MLE} rewrite the log likelihood

$$= \frac{N}{2}\log|2\pi\Sigma^{-1}| - \frac{1}{2}\sum_i \mathrm{tr}\left[(x_i-\mu)(x_i-\mu)^T\Sigma^{-1}\right]$$

$$= \frac{N}{2}\log|2\pi\Sigma^{-1}| - \frac{1}{2}\mathrm{trace}\left[S_\mu\Sigma^{-1}\right]$$

where $S_\mu = \sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$

$$\frac{\partial E}{\partial \Sigma} = \frac{N}{2}\Sigma^{-T} - \frac{1}{2}S_\mu^T = 0 \qquad * \; \Sigma^{-T} = \Sigma^{-1} = \Sigma$$

$$\frac{N}{2}\Sigma = \frac{1}{2}S_\mu^T$$

$$\Sigma_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)(x_i-\mu)^T$$

$D\times 1$ matrix $\qquad (D\times 1 \text{ matrix})^T = 1\times D$

Here Σ_{MLE} is a $D\times D$ matrix
within the matrix each element i,j is the covariance ~~of~~ between the i^{th} and j^{th} element of the random vector.
Since we have the Naive Bayes assumption, and we know that the variables are independent from each other, the covariance matrix will be a ~~dig~~ diagonal matrix, since non diagonal values in the matrix will have 0 values because they are independent from each other.

(2) $X \in R^D$ $[D \times 1]$

$Y \in R^M$ $[M \times 1]$ $\qquad \bar{\Sigma}' = \sigma^2 I$

$W \in R^{M \times D}$ $[M \times D]$

a) $\Sigma' \in R^{M \times M}$ $[M \times M]$

$$p(y|x,W) = \mathcal{N}(y|Wx, \Sigma')$$

$$E(w) = \prod_{i=1}^{N} (2\pi)^{-\frac{D}{2}} |\Sigma'|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(y^{(i)} - Wx^{(i)})^T \Sigma'^{-1} (y^{(i)} - Wx^{(i)})\right)$$

$$= -\log \mathcal{N}(y|Wx, \Sigma') = -\sum_{i=1}^{N} \left[-\frac{1}{2}\log 2\pi|\Sigma'| - \frac{1}{2}(y^{(i)} - Wx^{(i)})^T \Sigma'^{-1} (y^{(i)} - Wx^{(i)})\right]$$

$$= \sum_{i=1}^{N} \left[\frac{1}{2}\log 2\pi|\Sigma'| + \frac{1}{2}(y^{(i)} - Wx^{(i)})^T \Sigma'^{-1} (y^{(i)} - Wx^{(i)})\right]$$

$$* \quad \frac{\partial\left[(y-Wx)^T \Sigma'^{-1} (y-Wx)\right]}{\partial W} = -2\Sigma'^{-1}(y-Wx)x^T$$

$$\Rightarrow \frac{\partial E(w)}{\partial W} = \sum_{i=1}^{N} -\Sigma'^{-1}(y^{(i)} - Wx^{(i)})x^{(i)T} = 0$$

$$\sum_{i=1}^{N} -y^{(i)}x^{(i)T} + Wx^{(i)}x^{(i)T} = 0$$

$$-\sum_{i=1}^{N} y^{(i)}x^{(i)T} + \sum_{i=1}^{N} Wx^{(i)}x^{(i)T} = 0$$

$$W \cdot \underbrace{\sum_{i=1}^{N} x^{(i)}x^{(i)T}}_{XX^T} = \underbrace{\sum_{i=1}^{N} y^{(i)} x^{(i)T}}_{YX^T}$$

substitute:

$$WXX^T = YX^T$$

$$W = YX^T(XX^T)^{-1}$$

$M \times 1$, $1 \times D$, $D \times 1$, $1 \times D$; $D \times D$; $1 \times D$; $M \times D$

Hence we get W as a $M \times D$ matrix and this means that the ~~only~~ rows i of W are only dependent on the corresponding ith row from Y.

b) In the first part (a) we used $\Sigma' = \sigma^2 I = \begin{bmatrix} \sigma^2 & & \\ & \sigma^2 & \\ & & \ddots \\ & & & \sigma^2 \end{bmatrix}$ while here we use $\Sigma' = \begin{bmatrix} \sigma_1^2 & & \\ & \sigma_2^2 & \\ & & \ddots \\ & & & \sigma_M^2 \end{bmatrix}$. From the derivation above, for W we do not use Σ' in the derivation. Hence, W is independent on the value of Σ' and the derivation of W will remain the same and continues to show that the estimate of row i of W only uses the values of ith dimension of output vectors y.

$$\mu_N = \frac{1}{N}\sum_1^N x_i \qquad \Sigma_N = \frac{1}{N}\sum_1^N (x_i-\mu_N)(x_i-\mu_N)^T$$

(3) a) $\mu_N = \mu_{N-1} + \frac{1}{N}(X_N - \mu_{N-1})$

$$= \frac{1}{N-1}\sum_{i=1}^{N-1} x_i + \frac{1}{N}\left(X_N - \frac{1}{N-1}\sum_1^{N-1} x_i\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N-1} x_i + \frac{X_N}{N} - \frac{1}{N(N-1)}\sum_1^{N-1} x_i$$

$$= \frac{N}{N(N-1)}\sum_{i=1}^{N-1} x_i + \frac{(N-1)X_N}{N(N-1)} - \frac{\sum_1^{N-1} x_i}{N(N-1)} = \frac{N}{N(N-1)}\sum_{i=1}^{N-1} x_i - \frac{\sum_1^{N-1} x_i}{N(N-1)} + \frac{(N-1)X_N}{N(N-1)}$$

$$= \frac{(N-1)\sum_1^{N-1} x_i}{N(N-1)} + \frac{(N-1)}{N(N-1)}X_N = \frac{\cancel{(N-1)}\left[\sum_1^{N-1} x_i + X_N\right]}{N\cancel{(N-1)}} = \frac{1}{N}\sum_1^N x_i$$

b)

$$\cancel{\Sigma_{N-1} = E(X_{N-1}X_{N-1}^T) - \frac{N^2}{(N-1)^2}\left[E(X_N X_N^T) - \Sigma_N\right]}$$

$$\cancel{+ \frac{N}{(N-1)^2} X^N \mu_N^T + \frac{N}{(N-1)^2}\mu_N X}$$

$$\cancel{\mu_N \mu_N^T = E(X_N X_N^T) - \Sigma_N}$$

$$\lambda_N = E(XX^T) = \alpha \sum_{i=1}^N x_i x_i^T \qquad \text{where } \alpha = \frac{1}{N}$$

$$= \alpha\sum_1^{N-1} x_i x_i^T + \alpha X_N X_N^T$$

$$= \alpha(N-1)\lambda_{N-1} + \alpha X_N X_N^T = \boxed{(1-\alpha)\lambda_{N-1} + \alpha X_N X_N^T}$$

$$\Sigma_N = \lambda_N - \mu_N\mu_N^T = \alpha(N-1)\lambda_{N-1} + \alpha X_N X_N^T - \mu_N\mu_N^T$$

From part a use recursive def'n of μ_N

$$\mu_N = \mu_{N-1} + \alpha(X_N - \mu_{N-1}) = \mu_{N-1} + \alpha X_N - \alpha\mu_{N-1}$$

$$= \boxed{(1-\alpha)\mu_{N-1} + \alpha X_N}$$

$$= \lambda_N - \mu_N\mu_N^T = \lambda_N - [(1-\alpha)\mu_{N-1} + \alpha X_N][(1-\alpha)\mu_{N-1} + \alpha X_N]^T$$

$$= \lambda_N - [(1-\alpha)^2\mu_{N-1}\mu_{N-1}^T + \alpha(1-\alpha)[\mu_{N-1}X_N^T + X_N\mu_{N-1}^T] + \alpha^2 X_N X_N^T]$$

$$= \lambda_N - (1-\alpha)^2\mu_{N-1}\mu_{N-1}^T - \alpha(1-\alpha)[\mu_{N-1}X_N^T + X_N\mu_{N-1}^T] - \alpha^2 X_N X_N^T$$

$$= (1-\alpha)\lambda_{N-1} + \alpha X_N X_N^T - (1-\alpha)^2\mu_{N-1}\mu_{N-1}^T - \alpha(1-\alpha)[\mu_{N-1}X_N^T + X_N\mu_{N-1}^T] - \alpha^2 X_N X_N^T$$

$$= (1-\alpha)\Big[\underbrace{\lambda_{N-1} - \mu_{N-1}\mu_{N-1}^T}_{\Sigma_{N-1}} + \alpha\underbrace{[X_N X_N^T - \mu_{N-1}X_N^T - X_N\mu_{N-1}^T + \mu_{N-1}\mu_{N-1}^T]}_{(X_N-\mu_{N-1})(X_N-\mu_{N-1})^T}\Big]$$

Therefore: $\Sigma_N = (1-\overset{1/N}{\alpha})[\Sigma_{N-1} + 2(X_N - \mu_{N-1})(X_N - \mu_{N-1})^T]$

Scanned by CamScanner

(4) a) $x \in \mathbb{R}^N$ $p(x) = \mathcal{N}(x \mid \mu_x, \Sigma_x)$

$n \in \mathbb{R}^M$ $p(n) = \mathcal{N}(n \mid 0, \Sigma_n)$

$y = Ax + b + n$ $A \in \mathbb{R}^{M \times N}$ & $b \in \mathbb{R}^M$ are constants

$$\mu_y = E[y] = E[Ax + b + n] = AE[x] + AE[b] + E[n]$$
$$= A\mu_x + b + 0 = A\mu_x + b$$

$$\Sigma_y = E[(y-\mu_y)(y-\mu_y)^T] = E[(Ax+b+n)(Ax+b+n)^T]$$

$$= E\left[\left(\begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} x \\ n \end{bmatrix} + b - \begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} \mu_x \\ \mu_n \end{bmatrix} - b\right)\left(\begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} x \\ n \end{bmatrix} + b - \begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} \mu_x \\ \mu_n \end{bmatrix} - b\right)^T\right]$$

$$= \left[\begin{bmatrix} A & I_m \end{bmatrix}\left(\begin{bmatrix} x - \mu_x \\ n - \mu_n \end{bmatrix}\right)\left(\begin{bmatrix} x - \mu_x \\ n - \mu_n \end{bmatrix}\right)^T\begin{bmatrix} A^T \\ I_m \end{bmatrix}\right]$$

★ $\Sigma_{xn} = 0$ because x and n are independent

$$= \begin{bmatrix} A & I_m \end{bmatrix} E\left[\begin{bmatrix} (x-\mu_x)(x-\mu_x)^T & (x-\mu_x)(n-\mu_n)^T \\ (n-\mu_n)(x-\mu_x)^T & (n-\mu_n)(n-\mu_n)^T \end{bmatrix}\right]\begin{bmatrix} A^T \\ I_m \end{bmatrix}$$

$$= \begin{bmatrix} A & I_n \end{bmatrix}\begin{bmatrix} \Sigma_x & \Sigma_{xn} \\ \Sigma_{xn}^T & \Sigma_n \end{bmatrix}\begin{bmatrix} A^T \\ I_m \end{bmatrix} = \begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} \Sigma_x & 0 \\ 0 & \Sigma_n \end{bmatrix}\begin{bmatrix} A^T \\ I_m \end{bmatrix}$$

$$= \begin{bmatrix} A\Sigma_x + \Sigma_n \end{bmatrix}\begin{bmatrix} A^T \\ I_m \end{bmatrix} = A\Sigma_x A^T + \Sigma_n$$

b) The mean is the same; $\mu_y = A\mu_x + b$

Since $\Sigma_{xn} \neq 0$, because x & n are not independent:

$$\Sigma_y = \begin{bmatrix} A & I_m \end{bmatrix}\begin{bmatrix} \Sigma_x & \Sigma_{xn} \\ \Sigma_{xn}^T & \Sigma_n \end{bmatrix}\begin{bmatrix} A^T \\ I_m \end{bmatrix} = \begin{bmatrix} A\Sigma_x + \Sigma_{xn}^T & A\Sigma_{xn} + \Sigma_n \end{bmatrix}\begin{bmatrix} A^T \\ I_m \end{bmatrix}$$

$$= A\Sigma_x A^T + \Sigma_{xn}^T A^T + A\Sigma_{xn} + \Sigma_n$$