1. Dataset $D = \{(x_i, y_i)\}_{i=1}^{N}$ $f(x) = p(y=c|x) \propto p(x|y=c,\theta)\, p(y=c|\theta)$.

$$\log p_r(y=c|x;\pi,\mu,\Sigma)$$
$$= \sum_{i=1}^{N} \log p_r(y_i=c|x_i;\pi,\mu,\Sigma)$$
$$= \sum_{i=1}^{N} \log p_r(y_i;\pi) + \sum_{i=1}^{N} \log p_r(x_i|y_i;\mu,\Sigma)$$
$$= \sum_{i=1}^{N} \log \pi_{y_i} + \sum_{i=1}^{N} \log \mathcal{N}(x_i; \mu_{y_i}, \Sigma_{y_i})$$
$$\mathcal{LL} = \sum_{i=1}^{N}\sum_{c=1}^{K} \mathbb{I}\{y_i=c\} \log \pi_c + \sum_{i=1}^{N}\sum_{c=1}^{K} \mathbb{I}\{y_i=c\} \log \mathcal{N}(x_i;\mu_c,\Sigma_c)$$

(I) Derivating in terms of π_c and setting it to 0.

$$\mathcal{LL}_{\pi} = \mathcal{LL} + \lambda\left(1 - \sum_{c=1}^{K} \pi_c\right)$$ [the maximization is subject to $\left[\sum_{c=1}^{n} \pi_c = 1\right]$ λ Lagrange multiplier]

$$\frac{\partial \mathcal{LL}_{\pi}}{\partial \pi_c} = 0.$$

$$\sum_{i=1}^{N} \frac{\mathbb{I}\{y_i=c\}}{\pi_c} - \lambda = 0 \Rightarrow \pi_c = \frac{1}{\lambda} N_c \quad \left[N_c = \sum_{i=1}^{N} \mathbb{I}\{y_i=c\}\right]$$

Now, $\sum_{c=1}^{K} \lambda_c = 1$ $$\boxed{\pi_c = \frac{N_c}{N}}$$

$$\sum_{c=1}^{K} \frac{1}{\lambda} N_c = 1.$$

$$N_1 + N_2 + \dots N_K = \lambda \Rightarrow \lambda = N.$$

(II) Derivating in terms of μ_c and setting it to 0.

$$\frac{\partial \mathcal{LL}}{\partial \mu_c} = 0 \Rightarrow \sum_{i=1}^{N} \mathbb{I}\{y_i=c\} \frac{\partial}{\partial \mu_c}\left(-\tfrac{1}{2}(x_i-\mu_c)^T \Sigma_c (x_i-\mu_c)\right) = 0$$

$$\sum_{i=1}^{N} \mathbb{I}\{y_i=c\} \Sigma_c (x_i-\mu_c) = 0$$

$$\boxed{\mu_c = \frac{\sum_{i:\, y_i=c} x_i}{N_c}} \quad \Rightarrow \sum_{i:\, y_i=c} \mu_c = \sum_{i:\, y_i=c} x_i \Rightarrow \mu_c \cdot N_c = \sum_{i:\, y_i=c} x_i$$

(iii) Derivating in terms of Σ_c and setting to 0.

$$\frac{\partial LL}{\partial \Sigma_c} = 0 \qquad \sum_{y_i = c} \Big[\underbrace{\Big(-\tfrac{1}{2}\frac{\partial}{\partial \Sigma_c^{-1}} \log|\Sigma_c|\Big)}_{\downarrow} - \tfrac{1}{2}\frac{\partial}{\partial \Sigma_c^{-1}} \underbrace{\Big((x_i-\mu_c)^T \Sigma_c^{-1}(x_i-\mu_c)\Big)}_{\downarrow} \Big] = 0$$

$$\frac{\partial}{\partial \Sigma_c^{-1}} \log|\Sigma_c^{-1}| = \cancel{\Sigma_c} \qquad \text{Scalar } x^TAx = tr[xx^TA]$$

$$= \Sigma^T = \Sigma_c \qquad \frac{\partial}{\partial A} tr[xx^TA] = xx^T.$$

$$\sum_{y_i = c} \Big[\tfrac{1}{2}\Sigma_c - \tfrac{1}{2}(x_i-\mu_c)(x_i-\mu_c)^T \Big] = 0$$

$$\Rightarrow \quad N_c \Sigma_c = \sum_{y_i = c} (x_i-\mu_c)(x_i-\mu_c)^T \quad \Rightarrow \quad \boxed{\Sigma_c = \frac{\sum_{y_i=c} (x_i-\mu_c)(x_i-\mu_c)^T}{N_c}}$$

b) $\hat{y_0} = f(x_0) = \arg\max_{y_i} \; p(y_i)\, p(x_0|y_i) \qquad \forall\, i = 1 \ldots K.$

Let $y_i = C$,

$$= \arg\max_c \; p(c)\, p(x_0|c)$$

$$= \arg\max_c \; \pi_c \, \mathcal{N}_{x_0}(\mu_c, \Sigma_c)$$

$$= \arg\max_c \Big[\log \pi_c - \tfrac{1}{2}(x_0-\mu_c)^T \Sigma_c^{-1}(x_0-\mu_c) - \tfrac{1}{2}\log \Sigma_c \Big]$$

[taking log and removing constants]

$$= \arg\max_c \Big[\log \pi_c + x_0^T \Sigma_c^{-1}\mu_c - \tfrac{1}{2}\big(x_0^T\Sigma_c^{-1}x_0 + \mu_c^T \Sigma_c^{-1}\mu_c\big) + \cancel{\tfrac{1}{2}} \log|\Sigma_c| \Big]$$

This is QDA.

$$= \arg\max_c \; x_0^T \Sigma_c^{-1}\mu_c - \tfrac{1}{2} x_0^T \Sigma_c^{-1} x_0 + b_c$$

$$\Big[b_c = \log \pi_c - \tfrac{1}{2}\mu_c^T \Sigma_c^{-1}\mu_c - \tfrac{1}{2}\log|\Sigma_c| \Big]$$

2. $$\min_{\theta_1,\ldots,\theta_K} \frac{1}{m}\sum_{i=1}^{m}\left(\log \sum_{c=1}^{K} \exp(\theta_c^T \phi_i) - \theta_{y_i}^T \phi_i\right) + \frac{\lambda}{2}\sum_{c=1}^{K} \|\theta_c\|_1$$

Repeat

(I) Pick a sample i from training data.

(II) Calculate gradient based on i,

For $j = 1, \ldots K$

$$\nabla_{\theta_j} L_i = \frac{\exp(\theta_j^T \phi_i) \cdot \phi_i}{\sum_{c=1}^{K} \exp(\theta_c^T \phi_i)} \quad [\text{where } j \neq y_i]$$

$$= \frac{\exp(\theta_j^T \phi_i) \cdot \phi_i}{\sum_{c=1}^{K} \exp(\theta_c^T \phi_i)} - \phi_i \quad [\text{where } j = y_i].$$

(III) For $j = 1, \ldots K$

$$\theta_j = \theta_j - \delta \nabla_{\theta_j} L_i$$ // Gradient Descent

// Step size $= \delta$.

(IV) For $j = 1, \ldots K$

$$\theta_j = \mathrm{Prox}_{\delta\lambda/2}\, \theta_j$$ // Proximal operator.

$$\mathrm{Prox}_{\delta\lambda/2 \|\cdot\|_1}\, \theta_j = \begin{cases} \theta_j - \delta\lambda/2 & \theta_j > \delta\lambda/2 \\ 0 & |\theta_j| \leq \delta\lambda/2 \\ \theta_j + \delta\lambda/2 & \theta_j < -\delta\lambda/2 \end{cases}$$

Untill Convergence.

3. $\Pr(y_i = c) = \pi_c \quad \Pr(x_i | y_i = c) \sim \text{Multi}(p_c, L_i) = \dfrac{L_i!}{\prod_{j=1}^{d} x_{ij}!} \prod_{j=1}^{d} p_{cj}^{x_{ij}}$

a) $\log \Pr(x, y ; \pi, p)$

$$= \sum_{i=1}^{n} \log \Pr(x_i, y_i ; \pi, p)$$

$$= \sum_{i=1}^{n} \log \Pr(y_i ; \pi) + \sum_{i=1}^{n} \log \Pr(x_i | y_i ; p)$$

$$= \sum_{i=1}^{n} \log \pi_{y_i} + \sum_{i=1}^{n} \sum_{j=1}^{d} x_{ij} \log p_{y_i j} + \text{Const.}$$

$$\mathcal{LL} = \sum_{i=1}^{n} \sum_{c=1}^{K} \psi_{ic} \log \pi_c + \sum_{i=1}^{n} \sum_{c=1}^{K} \sum_{j=1}^{d} x_{ij} \psi_{ic} \log p_{cd} + \text{const.}$$

b) E-step: Calculate expectation $\psi_{ic} = \dfrac{\pi_c \, f(x_i ; p_c)}{\sum_{k'=1}^{K} \pi_{k'} \, f(x_i ; p_{k'})}$

$$f(x_i ; p_c) = \frac{L_i!}{\prod_{j=1}^{d} x_{ij}!} \prod_{j=1}^{d} p_{cj}^{x_{ij}}$$

M-step:

(i) π

$$\frac{\partial \mathcal{LL}_\pi}{\partial \pi_c} = 0$$

$$\mathcal{LL}_\pi = \mathcal{LL} + \underbrace{\lambda_1}_{\text{Lagrange Multiplier.}} \underbrace{\left(1 - \sum_{c=1}^{K} \pi_c\right)}_{\text{Constraint}}$$

$$\sum_{i=1}^{n} \psi_{ic} / \pi_c - \lambda_1 = 0$$

$$\pi_c = 1/\lambda_1 \sum_{i=1}^{n} \psi_{ic}$$

Constraint $\sum_{c=1}^{K} \pi_c = 1 \Rightarrow \sum_{c=1}^{K} 1/\lambda_1 \sum_{i=1}^{n} \psi_{ic} = 1$

$$\lambda_1 = \sum_{i=1}^{n} \sum_{c=1}^{K} \psi_{ic}.$$

$$= \sum_{i=1}^{n} 1 \quad \left[\sum_{c=1}^{K} \psi_{ic} = 1\right].$$

$$= n$$

$$\boxed{\pi_c = 1/n \sum_{i=1}^{n} \psi_{ic}}$$

(ii) p

$$\sum_{j=1}^{d} p_{cj} = 1$$

$$\mathcal{LL}_p = \mathcal{LL} + \lambda_2 \left(1 - \sum_{j=1}^{d} p_{cj}\right)$$

$$\frac{\partial \mathcal{LL}}{\partial p_{cj}} = 0$$

$$\sum_{i=1}^{n} \frac{x_{ij} \psi_{ic}}{p_{cj}} - \lambda_2 = 0$$

$$p_{cj} = 1/\lambda_2 \sum_{i=1}^{n} x_{ij} \psi_{ic}$$

now $\sum_{j=1}^{d} 1/\lambda_2 \sum_{i=1}^{n} x_{ij} \psi_{ic} = 1$

$$\lambda_2 = \sum_{j=1}^{d} \sum_{i=1}^{n} x_{ij} \psi_{ic}$$

$$= \sum_{i=1}^{n} X_i \psi_{ic} \quad [X_i = \text{total length of words}]$$

$$\boxed{p_{cj} = \frac{\sum_{i=1}^{n} x_{ij} \psi_{ic}}{\sum_{i=1}^{n} X_i \psi_{ic}}}$$

4. Classical scaling:

$$\underset{y_1,\dots,y_m}{\text{minimize}} \sum_{i=1}^{m}\sum_{j=1}^{n} (\tilde{x}_i^T \tilde{x}_j - y_i^T y_j)^2 \quad \text{where } \tilde{x}_i = x_i - (1/n)\sum_{j=1}^{n} x_j$$

a) The above formulation can be rewritten in matrix form,

$$\min_{Y^TY} \| \tilde{X}^T\tilde{X} - Y^TY \|$$

The matrix $\tilde{X}^T\tilde{X}$ is symmetric, therefore can be applied Eigen Value decomposition.

From Eckart Young Theorem,

$$\min_{\tilde{A}} \| A - \tilde{A} \|^2 \quad \text{s.t} \quad \text{rank}(\hat{A}) = k < r$$

is $\hat{A} = \sum_{i=1}^{k} \sigma_i u_i v_i^T \qquad \hat{A} = U_1 \Sigma_1 V_1^T$

This is the best approximation of A of rank k.

In our problem we use eigen decomposition instead of SVD.

$$\tilde{Y}^T Y = U_1 \Lambda_1 U_1^T = (\Lambda_1^{1/2} U_1^T)^T \Lambda_1^{1/2} U_1^T$$

$Y = \Lambda_1^{1/2} U_1^T$ | $\Lambda_1 =$ First k eigen values of Λ

$U_1 =$ First k eigen vectors of U

b) $D_{ij} = \| x_i - x_j \|^2 = \| x_i \|^2 + \| x_j \|^2 - 2x_i^T x_j$

$$= b1^T + 1b^T - 2X^TX$$

Let $b = \begin{bmatrix} \|x_1\|^2 \\ \|x_2\|^2 \\ \vdots \\ \|x_m\|^2 \end{bmatrix}$

Now, $X^TX = \frac{1}{2}(b1^T + 1b^T - D)$

~~Now~~ Let H be $(I - \frac{1}{n}11^T)$, so, $\tilde{X} = XH$ (Centered data)

Now, $\tilde{X}^T\tilde{X} = H^TX^TXH$

$$= \frac{1}{2}(Hb1^TH + H1b^TH - \underline{HDH}) \quad \begin{bmatrix} 1^TH = 0 \\ H^T1 = 0 \end{bmatrix}$$

$$= -\frac{1}{2}(I - \frac{1}{n}11^T) D (I - \frac{1}{n}11^T)$$

$$= -\frac{1}{2} B.$$