(N1) Suppose we have a sample of e-mails of class $y_i \in \{spam; non\text{-}spam\}$, each may contain word j ($x_j^i = 1$) or not ($x_j^i = 0$): $(y_1, x^1), \dots, (y_N, x^N)$.

Then the likelihood function equals:

$$L(y_1, x^1, \dots, y_N, x^N, \theta) = \prod_{i=1}^{N} P(Y = y_i, (X_j^i = x_j^i)_{j=1}^{d} | \theta) = \prod_{i=1}^{N} P(Y = y_i) P(X_j^i = x_j^i, j = \overline{1,d} | \theta, Y = y_i)$$

$$= \prod_{i=1}^{N} p_{y_i} \prod_{j=1}^{d} \theta_{jy_i}^{x_j^i} (1 - \theta_{jy_i})^{1 - x_j^i} \to \max_{\theta_{jy_i}, p_{y_i}, j = \overline{1,d}, y_i \in \{c, \bar{c}\}}$$

$$\Leftrightarrow \max_{\theta_{jy_i}, p_{y_i}} \Big[\underbrace{\sum_{i=1}^{N} \ln p_{y_i}}_{A} + \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{d} \left(x_j^i \ln \theta_{jy_i} + (1 - x_j^i) \ln(1 - \theta_{jy_i}) \right)}_{B} \Big] \quad (*)$$

Since A does not depend on θ_{jy_i} and B does not depend on p_{y_i}, (*) is equivalent to maximizing A and B separately.

For A:

$$\sum_{i=1}^{N} \ln p_{y_i} = \sum_{i=1}^{N} \left[\ln p_c \cdot \mathbb{1}\{y_i = c\} + \ln p_{\bar{c}} \mathbb{1}\{y_i = \bar{c}\} \right] = \ln p_c \sum_{i=1}^{N} \mathbb{1}\{y_i = c\} +$$

$$+ \ln(1 - p_c) \sum_{i=1}^{N} \mathbb{1}\{y_i = \bar{c}\} = n_c \ln p_c + (N - n_c) \ln(1 - p_c),$$ where n_c - total # of class c e-mails

Then, we have the following optimization problem:

$$\max_{p_c \geq 0} \left[n_c \ln p_c + (N - n_c) \ln(1 - p_c) \right]$$

F.O.C.

$$\frac{n_c}{\hat{p}_c} - \frac{N - n_c}{1 - \hat{p}_c} = 0 \Leftrightarrow \hat{p}_c = \frac{n_c}{N} \text{ and } \hat{p}_{\bar{c}} = \frac{N - n_c}{N}.$$

For B: for each $j = \overline{1,d}$

$$\sum_{i=1}^{N} \left(x_j^i \ln \theta_{jy_i} + (1 - x_j^i) \ln(1 - \theta_{jy_i}) \right) = \sum_{i=1}^{N} \left(x_j^i \mathbb{1}\{y_i = c\} \ln \theta_{jc} + x_j^i \mathbb{1}\{y_i = \bar{c}\} \ln \theta_{j\bar{c}} + \right.$$

$$\left. + (1 - x_j^i) \mathbb{1}\{y_i = c\} \ln(1 - \theta_{jc}) + (1 - x_j^i) \mathbb{1}\{y_i = \bar{c}\} \ln(1 - \theta_{j\bar{c}}) \right)$$

$$\Rightarrow \min_{\theta_{jc}, \theta_{j\bar{c}} \geq 0} B \Leftrightarrow \min_{\theta_{jc} \geq 0} B(\theta_{jc}) + \min_{\theta_{j\bar{c}} \geq 0} B(\theta_{j\bar{c}}) \quad \text{for } j = \overline{1,d}$$

F.O.C.:

$$\sum_{i=1}^{N} \left[x_j^i \mathbb{1}\{y_i = c\} \frac{1}{\hat{\theta}_{jc}} - (1 - x_j^i) \mathbb{1}\{y_i = c\} \frac{1}{1 - \hat{\theta}_{jc}} \right] = 0 \Leftrightarrow$$

Since $x_j^i = 1$, if word j present in i's email,

$\sum_{i=1}^{N} x_j^i \mathbb{1}\{y_i = c\} = n_{jc}$ - # e-mails of class c word j appeared in

$$\Leftrightarrow \frac{n_{jc}}{\hat{\theta}_{jc}} = \frac{n_c - n_{jc}}{1 - \hat{\theta}_{jc}} \Leftrightarrow \hat{\theta}_{jc} = \frac{n_{jc}}{n_c} \quad \text{for } j = \overline{1,d}$$

Equivalently $\hat{\theta}_{j\bar{c}} = \frac{n_{j\bar{c}}}{n_{\bar{c}}}$, $j = \overline{1,d}$

(N3) To find the distance (smallest) between a point x_0 and a hyperplane we need to solve the following optimization problem:

$$\begin{cases} \min_{\theta} \|x_0 - \theta\| \\ \text{s.t. } \beta^T \theta + \beta_0 = 0 \end{cases}$$, θ - is a point on a hyperplane $\beta^T x + \beta_0 = 0$

$$\Leftrightarrow \begin{cases} \min_{\theta} \frac{1}{2} \|x_0 - \theta\|^2 \\ \text{s.t. } \beta^T \theta + \beta_0 = 0 \end{cases} \quad (**)$$

$$L(\theta, \lambda) = \frac{1}{2} \|x_0 - \theta\|^2 + \lambda(\beta^T \theta + \beta_0) = \frac{1}{2}(x_0 - \theta)^T (x_0 - \theta) +$$

$$+ \lambda(\beta^T \theta + \beta_0) = \frac{1}{2}(\theta^T \theta - 2\theta^T x_0 + x_0^T x_0) + \lambda(\theta^T \beta + \beta_0)$$

$(**) \Leftrightarrow \min\limits_{\theta,\lambda} L(\theta,\lambda)$

FOC:

$$\begin{cases}\theta - x_0 + \lambda\beta = 0 \\ \theta^T\beta + \beta_0 = 0\end{cases} \Leftrightarrow \begin{cases}\theta = x_0 - \lambda\beta \\ x_0^T\beta - \lambda\beta^T\beta + \beta_0 = 0\end{cases} \Leftrightarrow \begin{cases}\theta = x_0 - \lambda\beta \\ \lambda = \dfrac{\beta^T x_0 + \beta_0}{\beta^T\beta}\end{cases} \Leftrightarrow$$

$$\Leftrightarrow \quad \theta = x_0 - \frac{\beta^T x_0 + \beta_0}{\beta^T\beta}\beta$$

Then the distance is the norm of a vector, that connects a point x_0 and a hyperplane $-(x_0-\theta)$

$$\|x_0 - \theta\| = \frac{\|\beta^T x_0 + \beta_0\|\,\|\beta\|}{\|\beta\|^2} = \frac{\|\beta^T x_0 - \beta^T\theta\|}{\|\beta\|} = \frac{\|\beta^T(x_0-\theta)\|}{\|\beta\|}.$$

N3 $\quad X_j | Y = y_k \sim N(\mu_{jk}, \sigma_{jk}^2)$, $k = \overline{1,K}$

$$L((x^i, y_i)_{i=1}^N, (\mu_{jk}, \sigma_{jk}^2), j = \overline{1,d}, k = \overline{1,K}) = \prod_{i=1}^N P(Y = y_i, X = x^i | \mu_{jy_i}, \sigma_{jy_i}^2) =$$

$$= \prod_{i=1}^N p_{y_i} \prod_{j=1}^d \frac{1}{\sqrt{2\pi\sigma_{jy_i}^2}} e^{-\frac{1}{2\sigma_{jy_i}^2}(x_j^i - \mu_{jy_i})^2} \to \max_{\mu_{jy_i}, \sigma_{jy_i}^2}, \; j = \overline{1,d}, \; y_i \in \{1, .., K\}$$

$$\Leftrightarrow \underbrace{\sum_{i=1}^N \ln p_{y_i}}_{\text{doesn't depend on } \mu_{jy_i}, \sigma_{jy_i}^2} + \sum_{i=1}^N \sum_{j=1}^d \left(-\ln\sqrt{2\pi} - \frac{1}{2}\ln\sigma_{jy_i}^2 - \frac{1}{2\sigma_{jy_i}^2}(x_j^i - \mu_{jy_i})^2\right) \to \max_{\mu_{jy_i}, \sigma_{jy_i}^2}$$

$$\Leftrightarrow \text{for } j = \overline{1,d} \quad \max_{\mu_{jy_i}, \sigma_{jy_i}^2} \underbrace{\sum_{i=1}^N \left(-\ln\sigma_{jy_i}^2 - \frac{1}{\sigma_{jy_i}^2}(x_j^i - \mu_{jy_i})^2\right)}_{A}$$

$$A = \sum_{i=1}^N \left(-\sum_{k=1}^K \ln\sigma_{jk}^2 \, \mathbb{1}\{y_i = k\} - \sum_{k=1}^K \frac{1}{\sigma_{jk}^2}(x_j^i - \mu_{jk})^2 \, \mathbb{1}\{y_i = k\}\right)$$

FOC: for $k = \overline{1,K}$

$$\begin{cases} -\sum\limits_{i=1}^N \dfrac{1}{\hat\sigma_{jk}^2} \mathbb{1}\{y_i = k\} + \sum\limits_{i=1}^N \dfrac{1}{\hat\sigma_{jk}^4}(x_j^i - \hat\mu_{jk})^2 \mathbb{1}\{y_i = k\} = 0 \\ \sum\limits_{i=1}^N \dfrac{2}{\hat\sigma_{jk}^2} \cdot (x_j^i - \hat\mu_{jk}) \mathbb{1}\{y_i = k\} = 0 \end{cases} \Leftrightarrow$$

$$\Leftrightarrow \begin{cases} \hat\mu_{jk} = \dfrac{\sum_{i=1}^N x_j^i \mathbb{1}\{y_i = k\}}{\sum_{i=1}^N \mathbb{1}\{y_i = k\}} \\ \hat\sigma_{jk} = \dfrac{\sum_{i=1}^N (x_j^i - \hat\mu_{jk})^2 \mathbb{1}\{y_i = k\}}{\sum_{i=1}^N \mathbb{1}\{y_i = k\}} \end{cases}$$