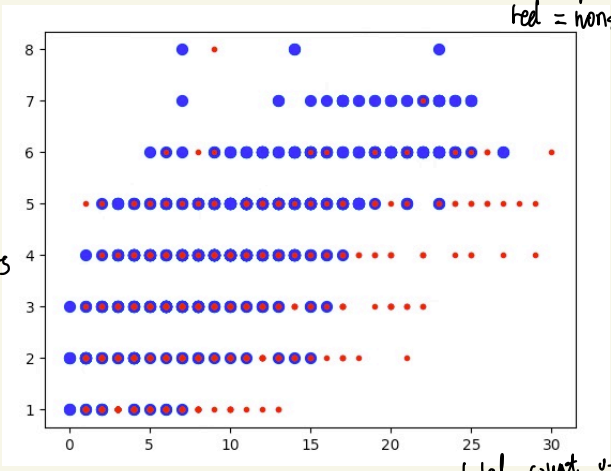
```
std: minimum average error rate is 0.08939641109298532, corresponding lambda is 0.15142857142857144
std: error rate for entire training set: 0.07732463295269165
std: error rate for test set: 0.10546875
log: minimum average error rate is 0.06329526916802612, corresponding lambda is 0.9797959183673469
log: error rate for entire training set: 0.05089722675367048
log: error rate for test set: 0.05989583333333337
binary: minimum average error rate is 0.08058727569331159, corresponding lambda is 0.8383673469387755
binary: error rate for entire training set: 0.06329526916802608
binary: error rate for test set: 0.072265625
```

	λ	average CV error	full training error	test set error
standardization	0.15	8.9%	7.7%	10.5%
log	0.98	6.3%	5.1%	6.0%
binary	0.84	8.1%	6.3%	7.2%

iii)

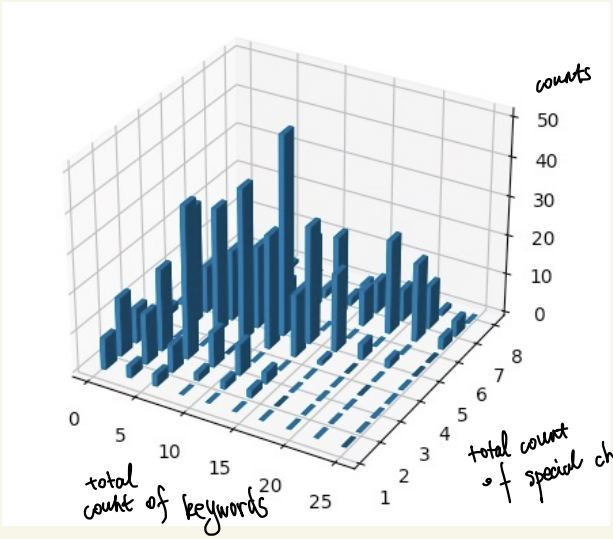
blue = spam
red = nonspam.



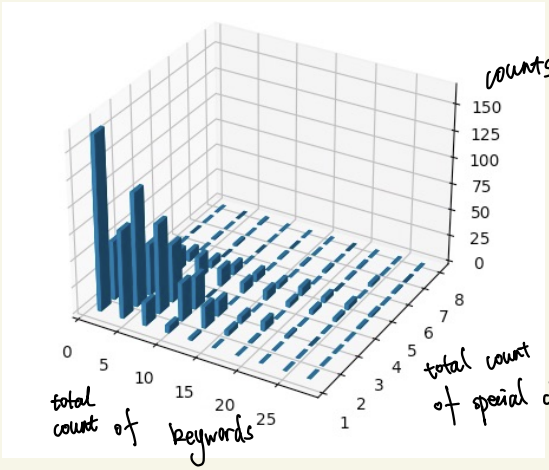
total count of special characters

total count of keywords.

(ii)



(iii)



(iv) non-spam emails have less keywords and special characters

8.5 (b)

$$g = \frac{d}{dw} f(w) = \sum_i (\mu_i - y_i) x_i = X^T(\mu - y)$$

$$\mu = \theta(u) = \frac{e^u}{1+e^u} = \text{sign}(w^T x)$$

log likelihood: $\log \prod_{i=1}^{N} \mu_i^{[y_i=1]} (1-\mu_i)^{[y_i=0]}$

$$= \sum_{i=1}^{N} \log \mu_i^{y_i} (1-\mu_i)^{1-y_i}$$

$$\nabla_w \sum_{i=1}^{N} \log \mu_i^{y_i} (1-\mu_i)^{1-y_i} = \nabla_w \sum_{i=1}^{N} \log \mu_i^{y_i} + \log(1-\mu_i)^{1-y_i} \qquad -\frac{d\mu_i}{dw}$$

$$= \nabla_w \sum_{i=1}^{N} y_i \log \mu_i + (1-y_i)\log(1-\mu_i) = \sum_{i=1}^{N} \frac{y_i}{\mu_i}\frac{d\mu_i}{dw} + \frac{1-y_i}{1-\mu_i}\frac{d(1-\mu_i)}{dw}$$

$$= \sum_{i=1}^{N} \frac{y_i}{\mu_i}\frac{d\mu_i}{dw} - \frac{1-y_i}{1-\mu_i}\frac{d\mu_i}{dw} = \sum_{i=1}^{N} \left(\frac{y_i}{\mu_i} - \frac{1-y_i}{1-\mu_i}\right)\frac{d\mu_i}{dw}$$

$$= \sum_{i=1}^{N} \frac{y_i - \mu_i}{\mu_i(1-\mu_i)} \frac{d\mu_i}{dw}$$

$$= \sum_{i=1}^{N} \frac{y_i - \mu_i}{\mu_i(1-\mu_i)} x_i \mu_i(1-\mu_i)$$

$$= \sum_{i=1}^{N} (y_i - \mu_i) x_i$$

$$= X^T(y-\mu)$$

Eq 8.5 is the gradient of negative log-likelihood.

$$\mu_i = \frac{e^{w^T x_i}}{1+e^{w^T x_i}} = \frac{1}{1+e^{-w^T x_i}}$$

$$\frac{d\mu_i}{dw} = e^{w^T x_i} \cdot (1+e^{w^T x_i})^{-1}$$

$$= x_i e^{w^T x_i} (1+e^{w^T x_i})^{-1} + e^{w^T x_i} \cdot -(1+e^{w^T x_i})^{-2} \cdot x_i e^{w^T x_i}$$

$$= \frac{x_i e^{w^T x_i}}{1+e^{w^T x_i}} - \frac{x_i e^{w^T x_i}{}^2}{(1+e^{w^T x_i})^2}$$

$$= x_i \mu_i (1-\mu_i)$$

c. $H = X^T S X$.

$S \triangleq \text{diag}(\mu_1(1-\mu_1) \cdots \mu_n(1-\mu_n))$ $0 < \mu_i < 1$

S is positive definite

$S = \sqrt{S}^T \sqrt{S}$

$H = X^T \sqrt{S}^T \sqrt{S} X = (\sqrt{S} X)^2 \geq 0$

$\log \mu_i^{y_i} \cdot (1-\mu_i)^{1-y_i}$

when H is PD, the negative log likelihood has a minimum value.

3.

(a) There are 2^D hypotheses

(b) $P\left[\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon\right] \leq 2M e^{-2\varepsilon^2 N}$

M : # of possible hypotheses $= 2^D$

N : # of data points

$$P\left[\left|E_{in}(g) - E_{out}(g)\right| > \varepsilon\right] \leq 2 \cdot 2^D e^{-2\varepsilon^2 N}$$