Comp Lab IV)

naivebayes

(i) $\hat{\theta}_{ML}(c_1^{c_{1,\dots,K}}, y_{1,\dots n}^{1,\dots K})$

$= \operatorname{argmax} \prod_{m=1}^{K} \hat{P}_{C,Y_1,\dots Y_n}(c^m, y_1^m, \dots, n; \theta)$

$= \operatorname{argmax} \begin{cases} s & c^m = \text{spam} \\ 1-s & c^m = \text{ham} \end{cases} \begin{cases} q_i^{y_i^m}(1-q_i)^{1-y_i^m} & c^m = \text{spam} \\ p_i^{y_i^m}(1-p_i^m)^{1-y_i^m} & c^m = \text{ham} \end{cases}$

Logistic Regression classifier

$$\phi(\theta; c, y_1, \dots y_n) = -\sum_{j=1}^{K} \left[1(c^j = \text{spam}) \ln \sigma(\theta_0 + \sum_{i=1}^{n} \theta_i y_i^j) + 1(c^j = \text{ham}) \ln(1 - \sigma(\theta_0 + \sum_{i=1}^{n} \theta_i y_i^j)) \right]$$

$u = \theta_0 + \sum_{i=1}^{n} \theta_i y_i^{(j)}$

$$\rightarrow \frac{\partial \phi}{\partial \theta_0} = -\sum_{j=1}^{K} \left[1(c^j = \text{spam}) \frac{\sigma(u)\sigma(-u)}{\sigma(u)} - 1(c^j = \text{ham}) \frac{\sigma(u)\sigma(-u)}{1-\sigma(u)} \right]$$

$$\frac{\partial \phi}{\partial \theta_i} = -\sum_{j=1}^{K} y_i^{(j)} \left[1(c^j = \text{spam}) \frac{\sigma(u)\sigma(-u)}{\sigma(u)} - 1(c^j = \text{ham}) \frac{\sigma(u)\sigma(-u)}{1-\sigma(u)} \right]$$

e) The second classifier did 41/49 spam and 47/59 ham.

The first classifier did 47/49 spam and 33/51 ham.

→ The RS performs better that the Bayesian.

f) In the new function, we can define "extract_features_word_weight" where the function would account for the weighted frequency of words in files.

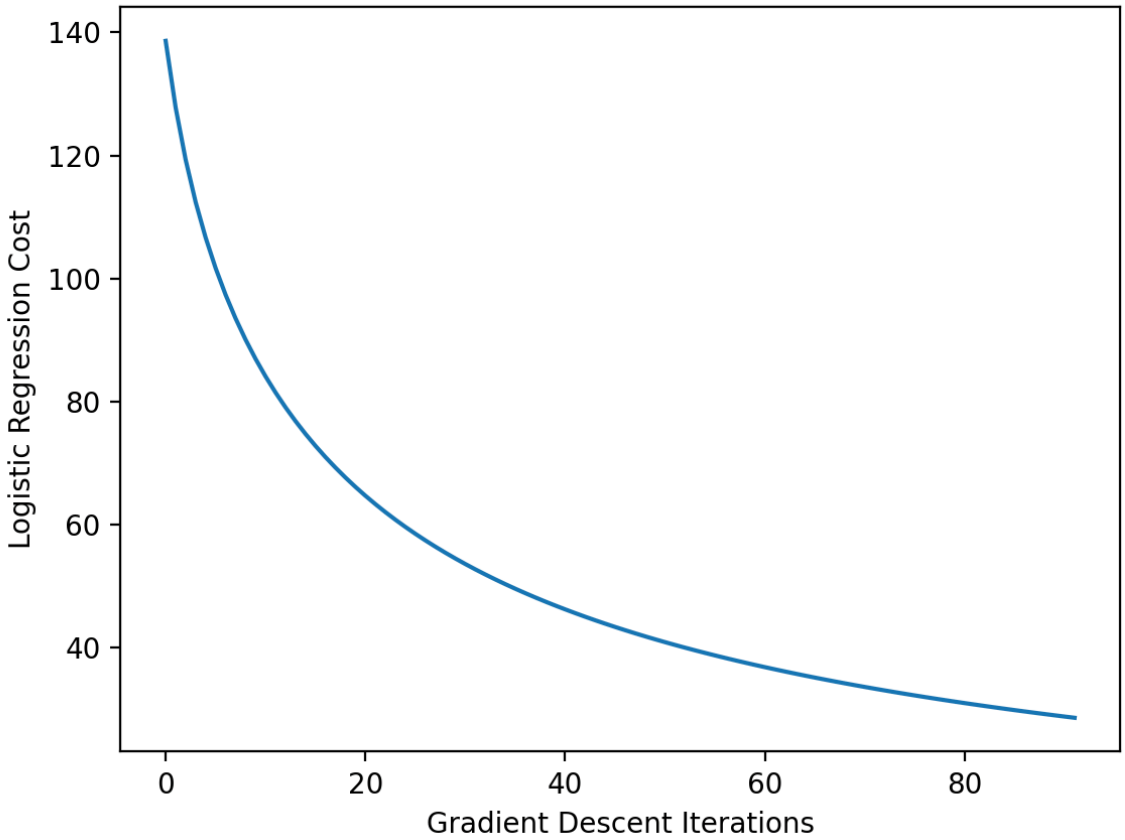
weighted freq = # word; occures in doc x W_{word}

where W_{word} = # total occurance over all docs.

→ with this metric

28/49 spam and 46/51 ham

It's not performing as well because it's giving more weight to words that appeared less.


140
120
100
80
60
40
Logistic Regression Cost
0
20
40
60
80
Gradient Descent Iterations