COMS W4903: Machine Learning for Data Science

Homework 2

Jc4609

Problem 1:

Problem1:

(a) Let $f(\pi) = \prod_{i=1}^{n} \pi^{y_i}(1-\pi)^{1-y_i}$, suppose the number of $y_i=1$ is K then, $f(\pi) = \pi^K \cdot (1-\pi)^{n-K}$.

$\frac{\partial f(\pi)}{\partial \pi} = 0 \Rightarrow \pi = \frac{K}{n}$

(b) suppose ~~there are~~ the number of

$y=1$ and $x_{i1}=1$ is L

$y=1$ and $x_{i1}=0$ is K

$y=0$ and $x_{i1}=1$ is m

$y=0$ and $x_{i1}=0$ is n

then $f(\theta_1^{(1)}, \theta_0^{(1)}) = \theta_1^{(1)L} (1-\theta_1^{(1)})^K \theta_0^{(1)m} (1-\theta_0^{(1)})^n$

$\begin{cases} \frac{\partial f}{\partial \theta_1^{(1)}} = 0 \\ \frac{\partial f}{\partial \theta_0^{(1)}} = 0 \end{cases} \Rightarrow \theta_1^{(1)} = \frac{L}{L+K}$

$\theta_0^{(1)} = \frac{m}{m+n}$

(c) ~~$f(\theta_1, \theta_2)$ suppose $\sum_{i=1}^{n} x_{i,2} = K$ then~~

suppose the number of $y_i=1$ is K

$f(\theta_1^{(2)}, \theta_0^{(2)}) = \theta_1^{(2)K} \theta_0^{(2)n-K} \prod_{i=1}^{n} x_{i2}^{-(\theta_{y_i}^{(2)}+1)}$

$g(\theta_1^{(2)}, \theta_0^{(2)}) = \ln f(\theta_1^{(2)}, \theta_0^{(2)}) = K \ln \theta_1^{(2)} + (n-K) \ln \theta_0^{(2)} - \sum_{i=1}^{n} (\theta_{y_i}^{(2)}+1) \ln x_i$

$$\begin{cases} \frac{\partial g}{\partial \theta_1^{(2)}} = 0 \\ \frac{\partial g}{\partial \theta_0^{(2)}} = 0 \end{cases} \Rightarrow \theta_1^{(2)} = \frac{K}{\sum_{i \in A} \ln x_{i2}}, \quad A = \{ i \mid y_i = 1 \}$$

$$\theta_0^{(2)} = \frac{n-K}{\sum_{i \in B} \ln x_{i2}} \qquad B = \{ i \mid y_i = 0 \}.$$

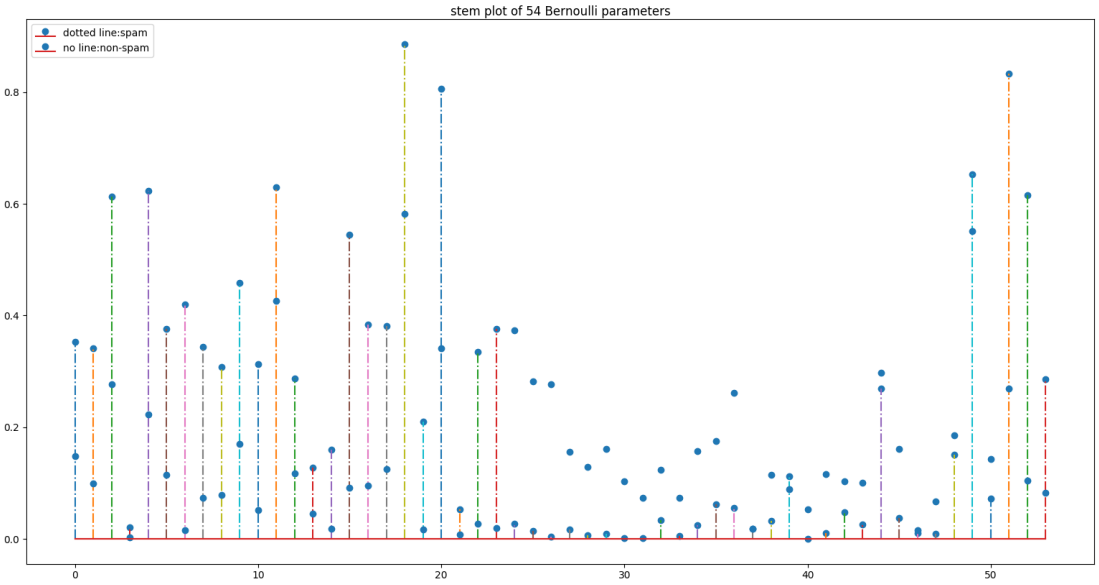
Problem 2:

(a)

Y_test / Y_prediction	1	0
1	32	2
0	4	54

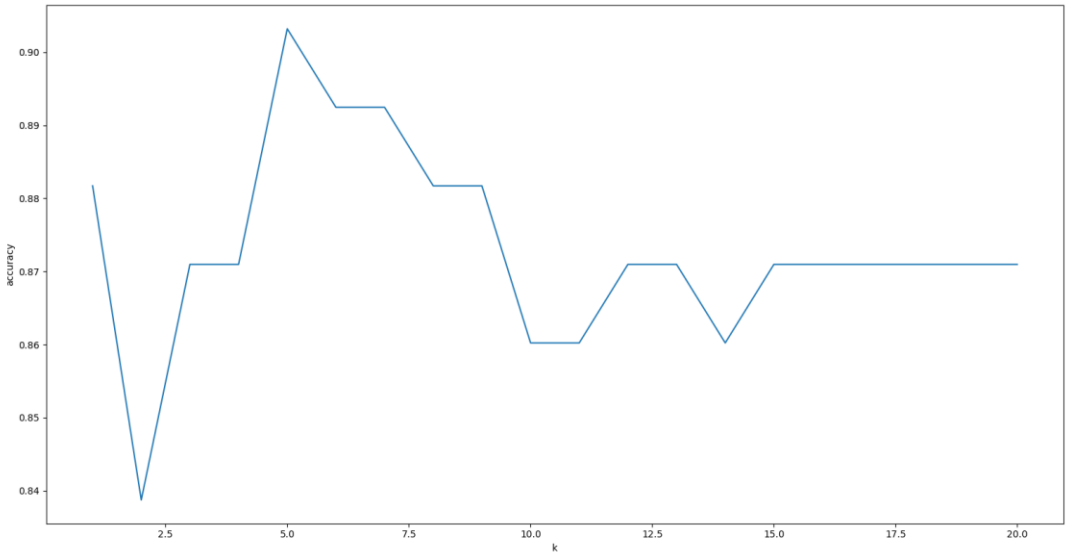
Accuracy = 0.924731182796

(b)



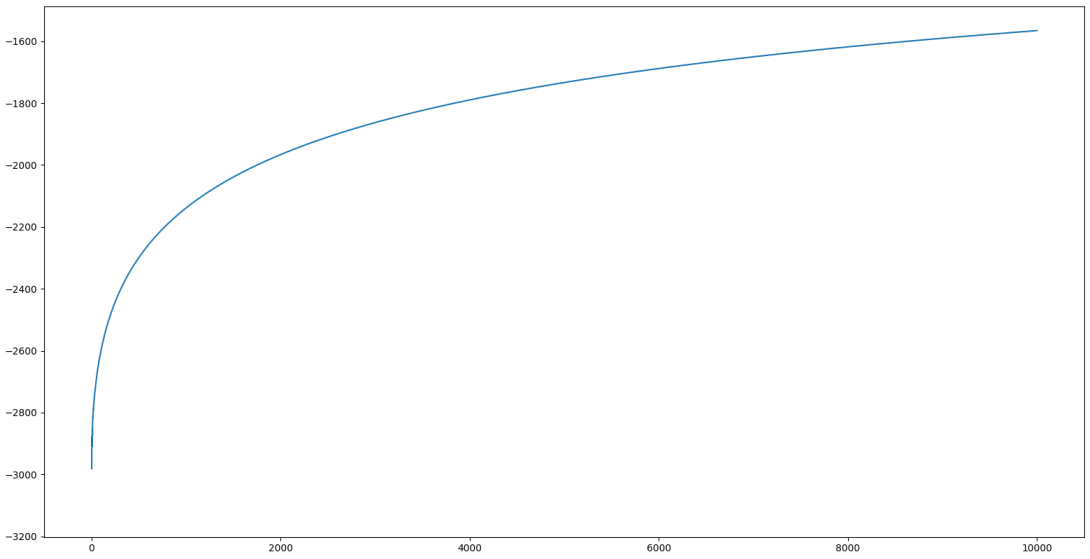
The point with dotted line represent the Bernoulli parameter of spam emails. The point with no line represent the Bernoulli parameter of non-spam emails.
We can know from the file “spambase.names” that the 16th parameter represent the word frequency of ‘free’ and 52th value represent the word frequency of ‘!’.
Since the Index of my plot start from zero, so the 16th and 52th parameter in my plot is actually the 15th and 51th. We can see that the 16th and 52th parameter of spam emails is about 0.55 and 0.8, while the 16th and 52th parameter of non-spam emails is about 0.1 and 0.25. Thus the conclusion is that the frequency of word ‘free’ and ‘!’ in spam email is much higher than that in non-spam email.

(c)



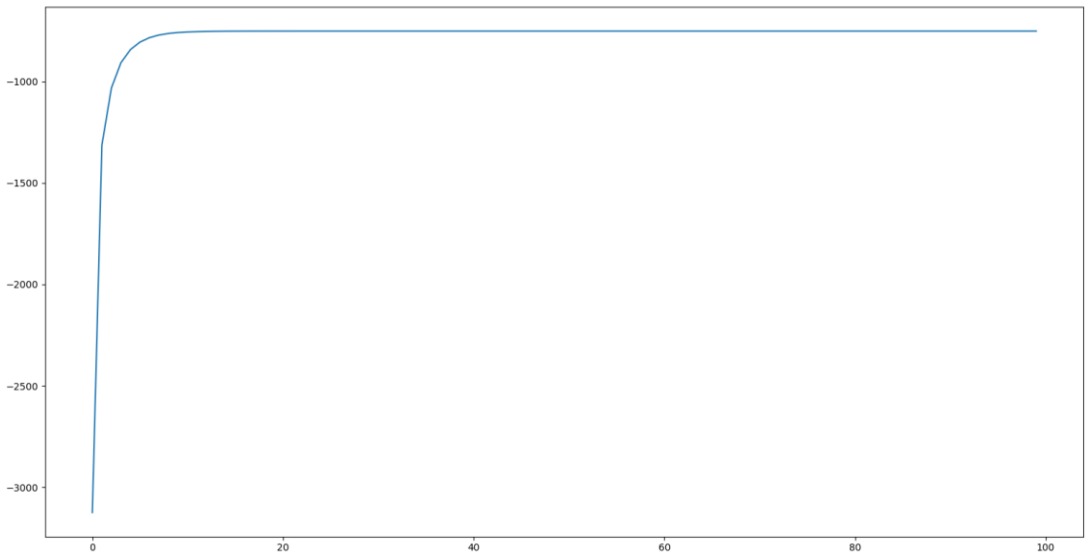
(d)

At first, I tried to do the steepest ascent algorithm on the data without normalizing, the result looks strange. It seems that the step size of every iteration is too big, so the objective training function turns to infinite at the second step. Then I normalize the data and it produce the following result:



The accuracy is 0.903225806452

(e)



the accuarcy is 0.913978494624.