Problem 1

Name: Xun Xue UNI: xx2241

(a) $\nabla_\pi \left(\sum_{i=1}^{n} \ln P(y_i|\pi) + \sum_{i=1}^{n} \ln P(x_{i1}|\theta_{y_i}^{(1)}) + \sum_{i=1}^{n} \ln P(x_{i2}|\theta_{y_i}^{(2)}) \right) = 0$

$\nabla_\pi \sum_{i=1}^{n} \ln P(y_i|\pi) = 0 \quad \therefore \nabla_\pi \sum_{i=1}^{n} \ln \pi^{1(y_i=1)} (1-\pi)^{1-1(y_i=1)} = 0$

$\nabla_\pi \sum_{i=1}^{n} \left(1(y_i=1) \ln \pi + (1-1(y_i=1)) \ln(1-\pi) \right) = 0$

$\sum_{i=1}^{n} \left(\frac{1(y_i=1)}{\pi} + \frac{1(y_i=1)-1}{1-\pi} \right) = 0$

$\sum_{i=1}^{n} \frac{1(y_i=1)-\pi}{\pi(1-\pi)} = 0$

$\sum_{i=1}^{n} 1(y_i=1) = n\pi$

$\therefore \pi = \frac{1}{n} \sum_{i=1}^{n} 1(y_i=1)$

(Since $y \in \{0,1\}$, the probability of $y=0$ is $1-\pi$, so we don't need to derive π_0, π_1 seperately)

(b) $\nabla_{\hat{\theta}_y^{(1)}} \left(\sum_{i=1}^{n} \ln P(y_i|\pi) + \sum_{i=1}^{n} \ln P(x_{i1}|\theta_{y_i}^{(1)}) + \sum_{i=1}^{n} \ln P(x_{i2}|\theta_{y_i}^{(2)}) \right) = 0$

Since $\theta_y^{(1)}$ can be in class $y=0$ or $y=1$ and the distribution is independent with each other. So we can derive θ_{y_0} and θ_{y_1} independently and then combine the result to θ_y.

when $y=0$ $\nabla_{\theta_{y_0}^{(1)}} \sum_{i=1}^{n(y=0)} \ln(x_{i1}|\theta_{y_0}^{(1)}) = 0$ $\nabla_{\hat{\theta}_{y_0}^{(1)}} \sum_{i=1}^{n(y=0)} (x_{i1} \ln \theta_{y_0}^{(1)} + (1-x_{i1}) \ln(1-\theta_{y_i}^{(1)})) = 0$

$\sum_{i=1}^{n(y=0)} \left(\frac{x_{i1}}{\theta_{y_0}^{(1)}} + \frac{x_{i1}-1}{1-\theta_{y_0}^{(1)}} \right) = 0 \quad \therefore 1(y=1) \sum_{i=1}^{n} x_{i1} - n_0 \hat{\theta}_{y_0}^{(1)} = 0 \quad \hat{\theta}_{y_0}^{(1)} = \frac{1}{n_0} \sum_{i=1}^{n(y=0)} x_{i1} = \frac{1}{n_0} \sum_{i=1}^{n} x_{i1} \cdot 1(y=0)$

when $y=1$ $\nabla_{\theta_{y_1}^{(1)}} \sum_{i=1}^{n(y=1)} \ln(x_{i1}|\theta_{y_1}^{(1)}) = 0$ $\nabla_{\theta_{y_1}^{(1)}} \sum_{i=1}^{n(y=1)} (x_{i1} \ln \theta_{y_1}^{(1)} + (1-x_{i1}) \ln(1-\theta_{y_1}^{(1)})) = 0$

$\sum_{i=1}^{n(y=1)} \left(\frac{x_{i1}}{\theta_{y_1}^{(1)}} + \frac{x_{i1}-1}{1-\theta_{y_1}^{(1)}} \right) = 0 \quad 1(y=1) \sum_{i=1}^{n} x_{i1} - n_0 \hat{\theta}_{y_1}^{(1)} = 0 \quad \hat{\theta}_{y_1}^{(1)} = \frac{1}{n_0} \sum_{i=1}^{n} x_{i1} 1(y=1)$. n_0 means n when $y=0$

therefore, combining the two results we can derive that

$\hat{\theta}_y^{(1)} = \frac{1}{n(y=y_i)} \sum_{i=1}^{n} x_{i1} \cdot 1(y=y_i) \quad y_i \in \{0,1\}$

(c) $\nabla_{\hat{\theta}_y^{(2)}} \left(\sum_{i=1}^{n} \ln P(y_i|x) + \sum_{i=1}^{n} \ln P(x_{i1}|\theta_{y_i}^{(1)}) + \sum_{i=1}^{n} \ln P(x_{i2}|\theta_{y_i}^{(2)}) \right) = 0$

Similar as problem (b). Since θ_{y0} and θ_{y1} is independent with each other we can derive θ_{y0} and θ_{y1} independently and then combine the result to θ_y

when $y=0$ $\nabla_{\theta_{y0}^{(2)}} \sum_{i=1}^{n} \ln(x_{i1}|\theta_{y0}^{(2)}) \cdot \mathbb{I}(y=0)$ $\nabla_{\theta_{y0}^{(2)}} \sum_{i=1}^{n} (\ln \theta_{y0}^{(2)} - (\theta_{y0}^{(2)}+1) \ln x_{i2}) \cdot \mathbb{I}(y=0) = 0$

$\sum_{i=1}^{n} \left(\frac{1}{\theta_{y0}^{(2)}} - \ln x_{i2} \right) \cdot \mathbb{I}(y=0) = 0$ $\frac{n(y=0)}{\theta_{y0}^{(2)}} = \sum_{i=1}^{n} \ln x_{i2} \cdot \mathbb{I}(y=0)$ $\hat{\theta}_{y0}^{(2)} = \frac{n(y=0)}{\sum_{i=1}^{n} \ln x_{i2} \cdot \mathbb{I}(y=0)}$

Similarly, we can derive that $\hat{\theta}_{y1}^{(2)} = \frac{n(y=1)}{\sum_{i=1}^{n} \ln x_{i2} \cdot \mathbb{I}(y=1)}$

therefore combining these two results, we can derive that

$$\hat{\theta}_y^{(2)} = \frac{n(y=y_i)}{\sum_{i=1}^{n} \ln x_{i2} \cdot \mathbb{I}(y=y_i)}, \quad y_i \in \{0,1\}$$

Programming part

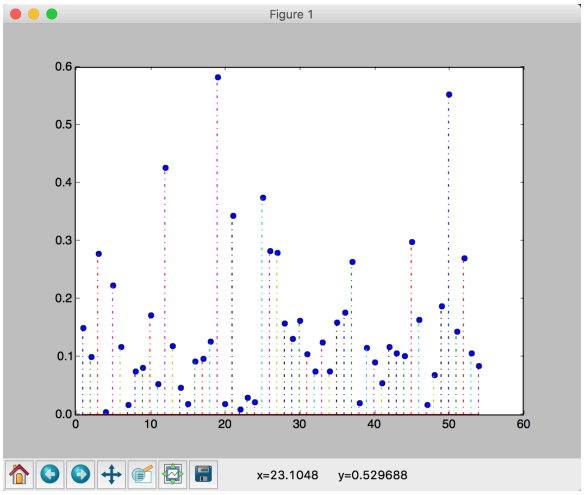
(a)

	y =1	y=0
y’ = 1	32	2
y’ = 0	5	54

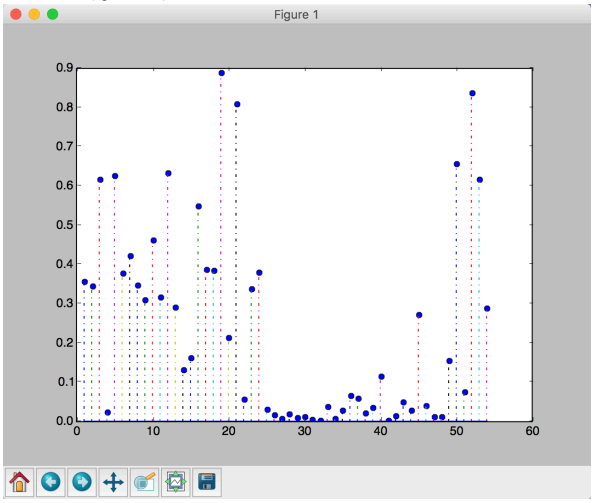
Prediction accuracy: (32+54)/93 = 92.47%

(b)

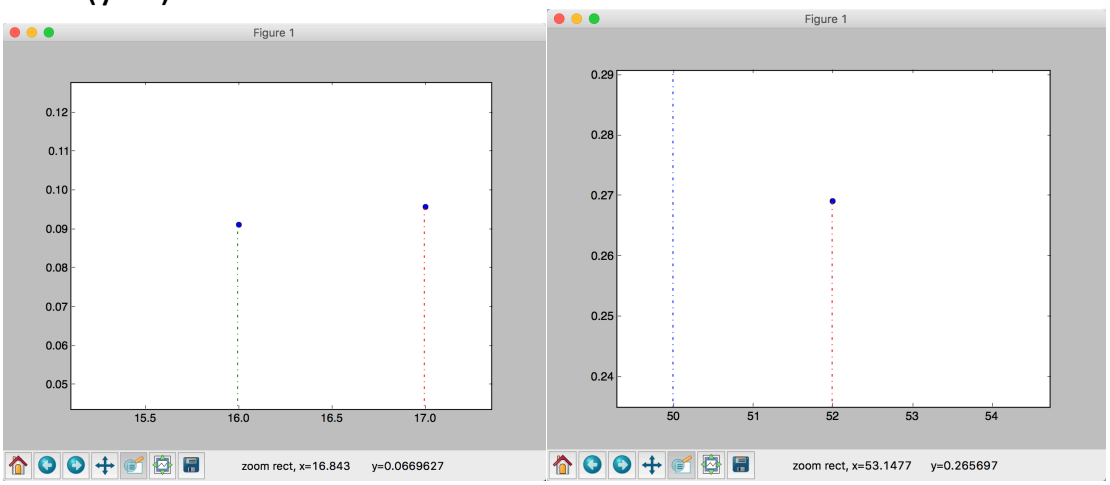
class(y=0)



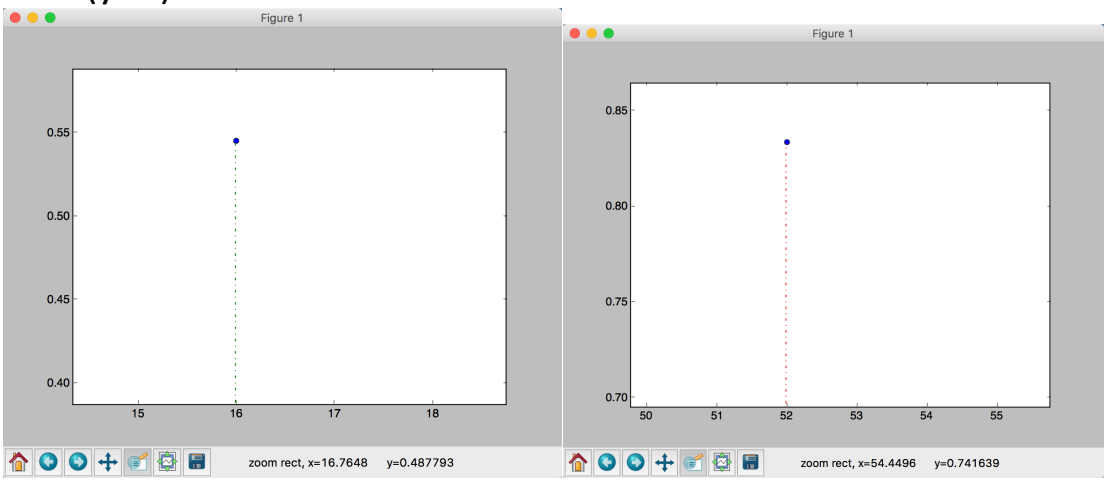
class(y=1)



class(y=0)



class(y=1)



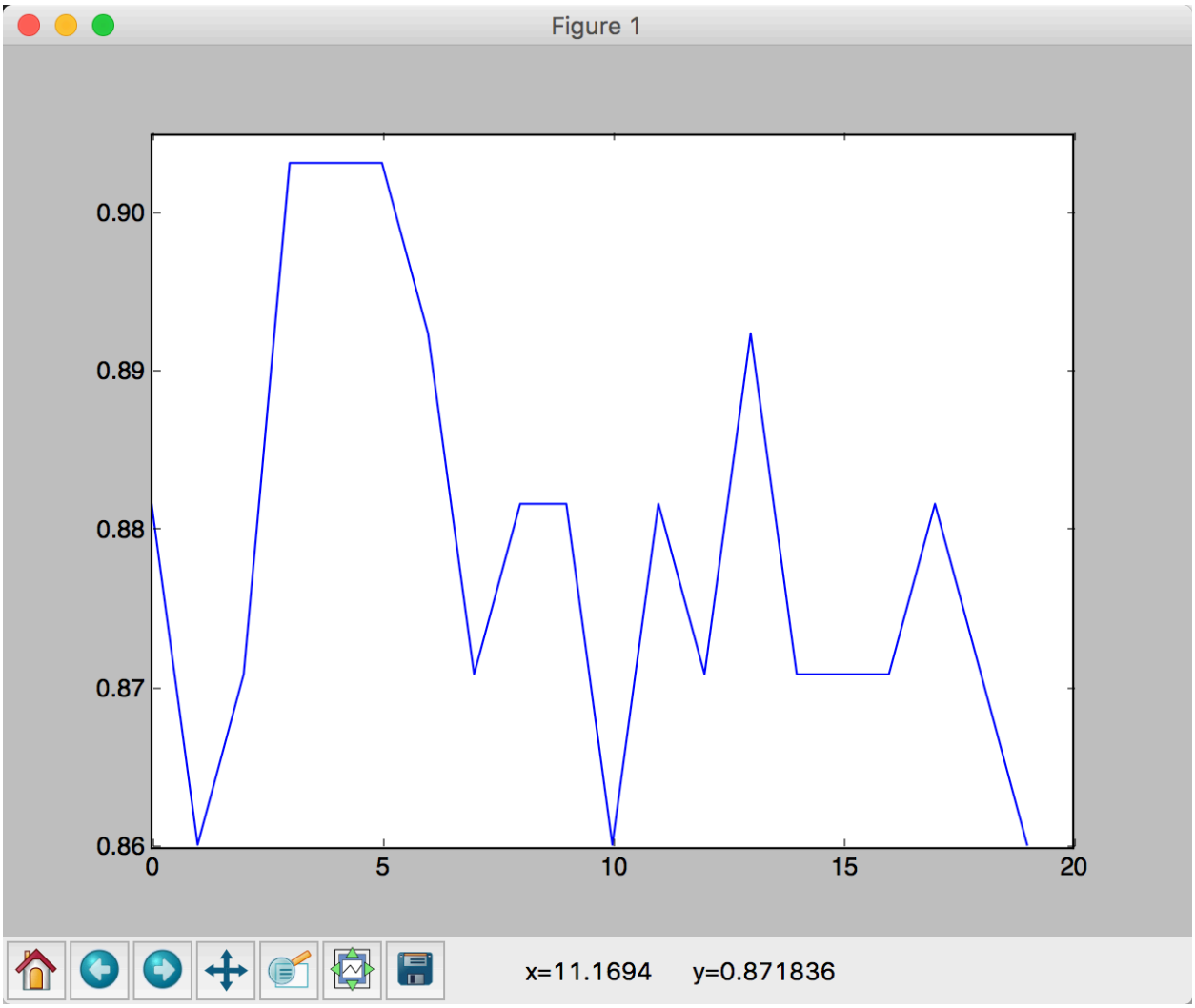
In the file spambase.names, we can know that the 16th dimension is word_freq_free, the 52th dimension is char_freq_!.

Since the word “free” is open in the commercial advertisement email, so it is an important feature to be a spam email but a very less important feature to be a non-spam email.

The character “!” is a common character in email. But in spam email like advertisement they often brag their product by this character. Therefore, it is very likely to be in a spam email and became an important feature. In non-spam it still often has “!” in the email, but not as important as in spam email.

(c)

knn prediction accuracy with x axis as k



(d)

objective function with x axis as iteration times

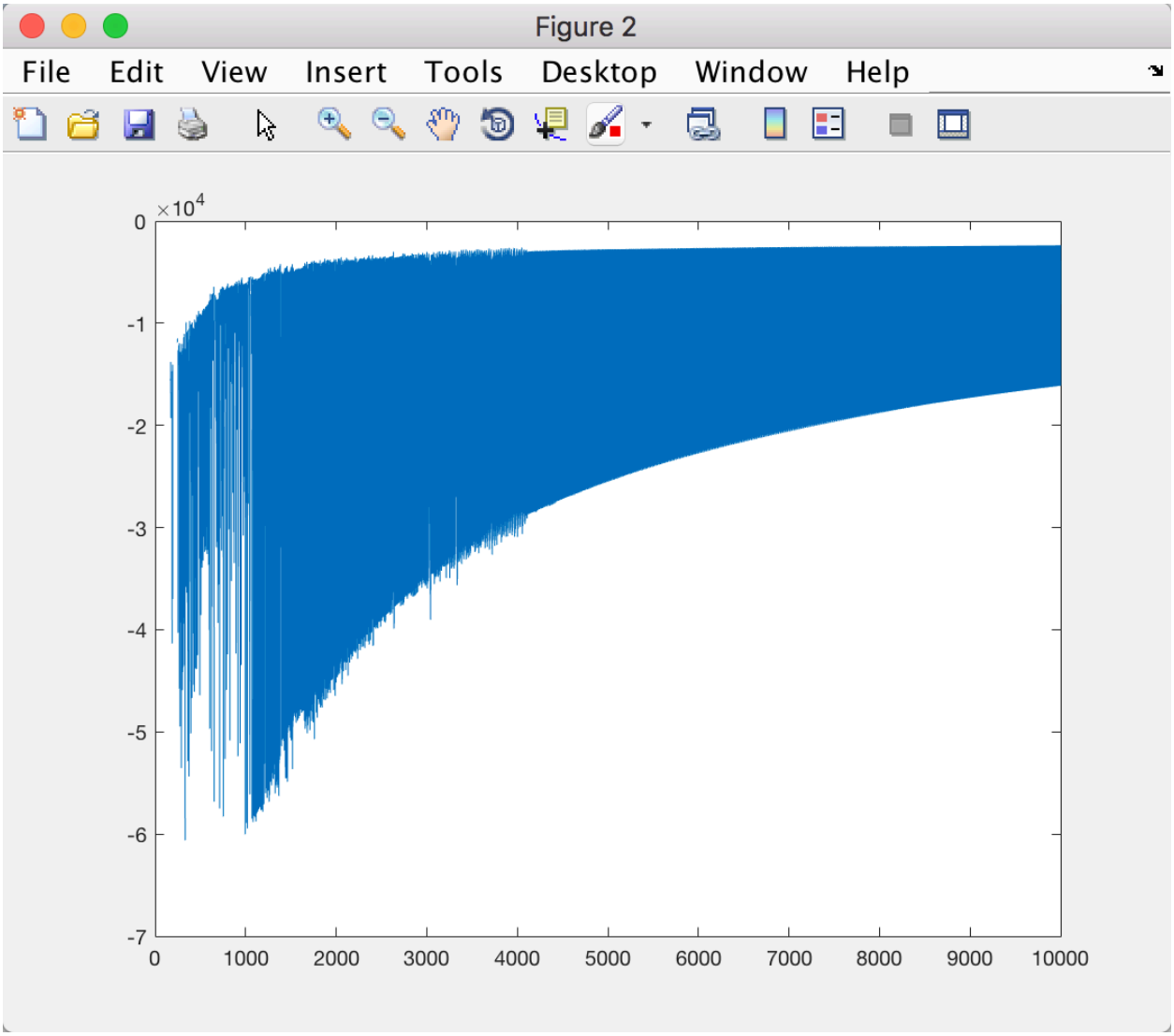
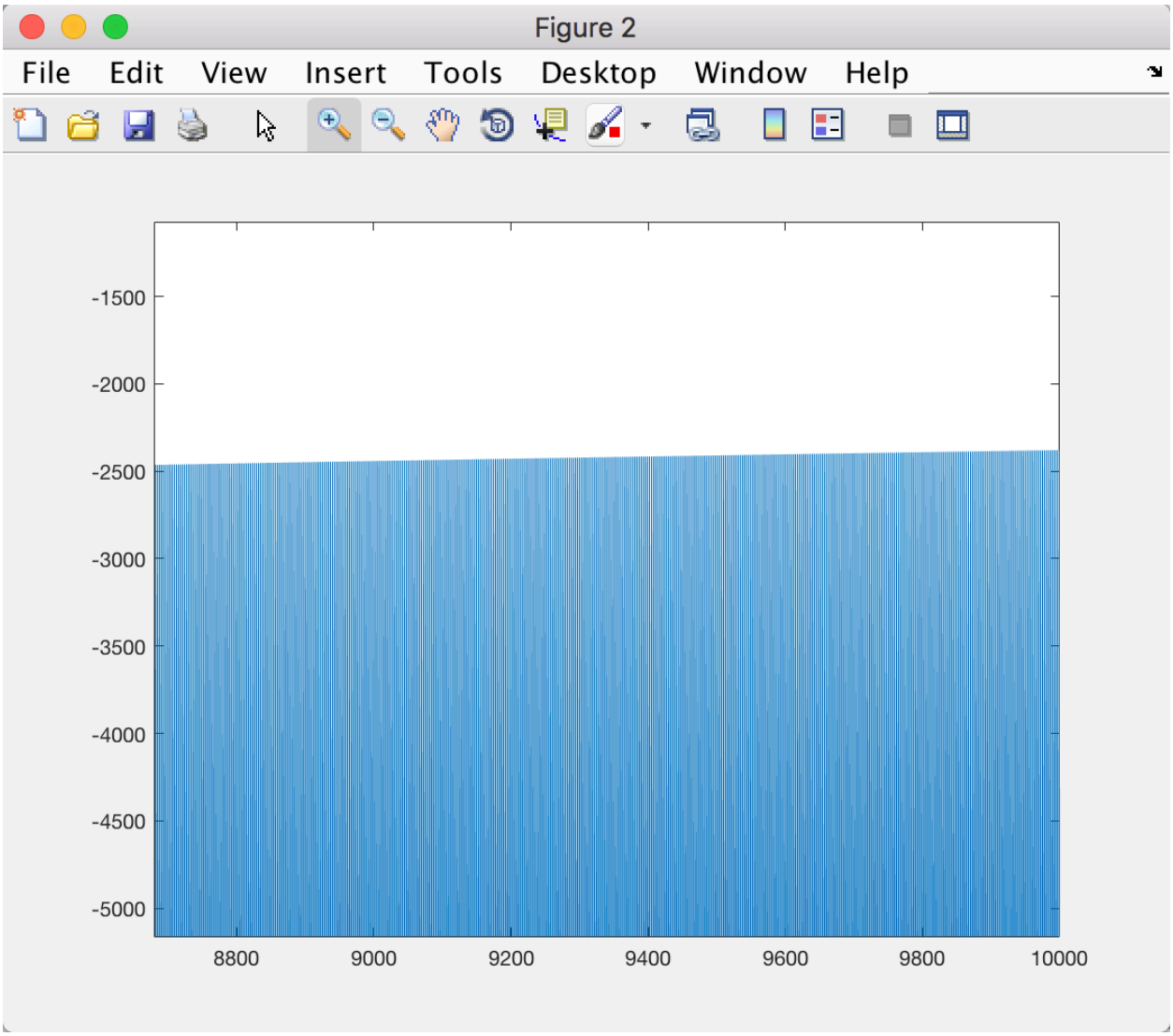


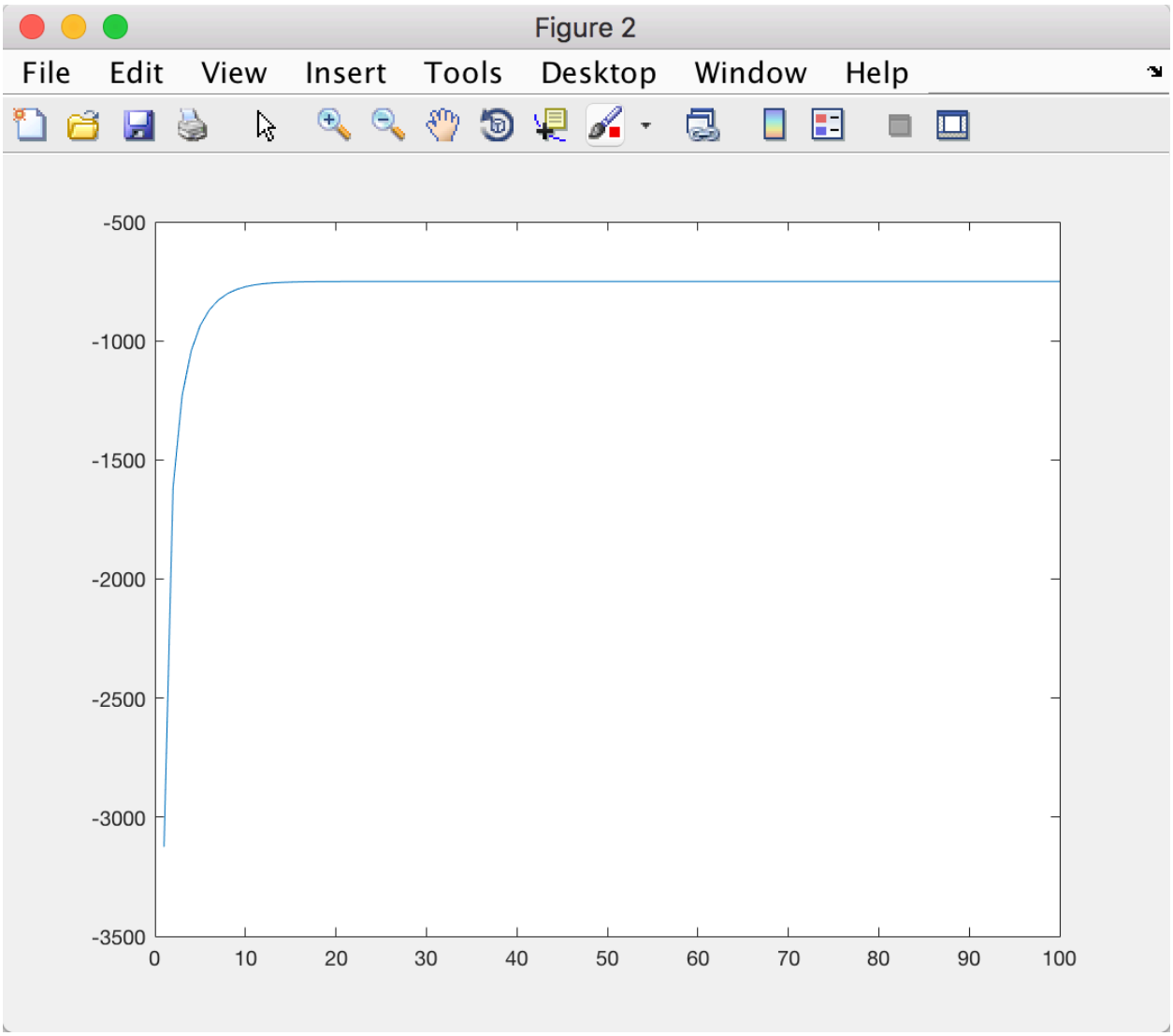
Figure 2
File Edit View Insert Tools Desktop Window Help
×10^4
0
-1
-2
-3
-4
-5
-6
-7
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000

zoomed in

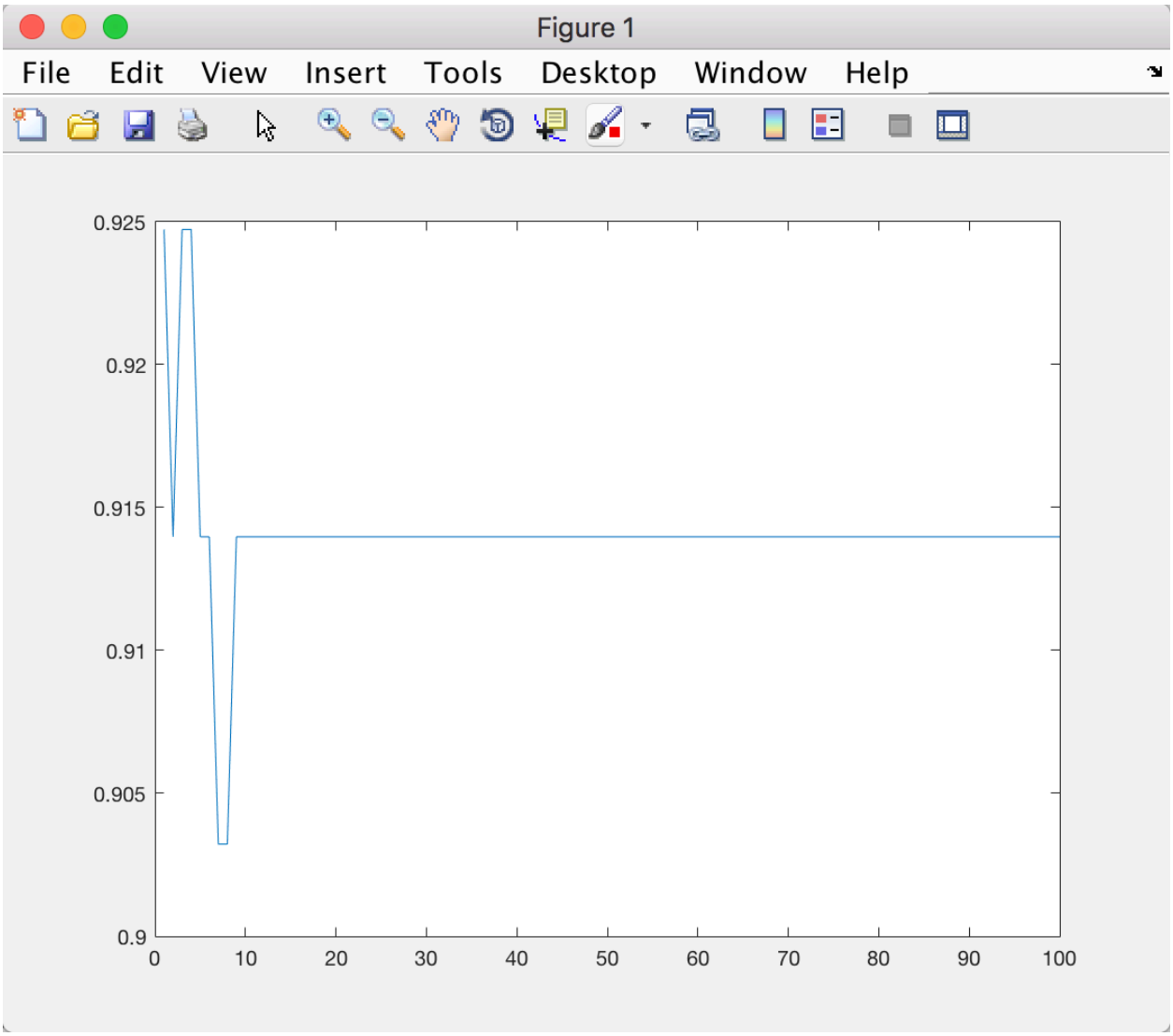


(e)

objective function with x axis as iteration times



test accuracy with x axis as iteration times



the final accuracy after 100 iteration is 91.4%