CSC411 A1

Q1.1.
Proof:

$$p(y=1|x) = \frac{p(x|y=1)*p(y=1)}{p(x)} = \frac{p(x|y=1)*p(y=1)}{p(x|y=1)*p(y=1)+p(x|y=0)*p(y=0)}$$

$$= \frac{1}{1+\frac{p(x|y=0)*p(y=0)}{p(x|y=1)*p(y=1)}} = \frac{1}{1+\exp\left(\ln\left(\frac{p(x|y=0)*p(y=0)}{p(x|y=1)*p(y=1)}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\ln\left(\frac{p(x|y=0)}{p(x|y=1)}\right)+\ln\left(\frac{p(y=0)}{p(y=1)}\right)\right)} = \frac{1}{1+\exp\left(\ln\left(\frac{p(x|y=0)}{p(x|y=1)}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\ln\left(\frac{p(x_i|y=0)}{p(x_i|y=1)}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)} \ (by\ conditionally\ independent)$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\ln\left(\frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_{i0})^2\right)}{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_{i1})^2\right)}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\ln\left(\frac{\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_{i0})^2\right)}{\exp\left(-\frac{1}{2\sigma^2}(x_i-\mu_{i1})^2\right)}\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\ln\left(\exp\left(\frac{1}{2\sigma^2}(x_i-\mu_{i1})^2-\frac{1}{2\sigma^2}(x_i-\mu_{i0})^2\right)\right)+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\frac{(x_i-\mu_{i1})^2-(x_i-\mu_{i0})^2}{2\sigma^2}+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1+\exp\left(\sum_{i=1}^{D}\frac{({\mu_{i1}}^2-{\mu_{i0}}^2)+2x_i(\mu_{i0}-\mu_{i1})}{2\sigma^2}+\ln\left(\frac{1-\alpha}{\alpha}\right)\right)}$$

$$= \frac{1}{1 + \exp\left(\sum_{i=1}^{D} \frac{(\mu_{i0} - \mu_{i1})}{\sigma^2} x_i + \ln\left(\frac{1-\alpha}{\alpha}\right) + \sum_{i=1}^{D} \frac{({\mu_{i1}}^2 - {\mu_{i0}}^2)}{2\sigma^2}\right)}$$

From the equation above, it shows that p(y = 1|x) takes the form of a logistic function:

p(y = 1|x) = $\sigma(w^T x + b) = \frac{1}{1+\exp(-\sum_{i=1}^{D} w_i x_i - b)}$

where $w_i = -\frac{(\mu_{i0} - \mu_{i1})}{\sigma^2}$ b = $-\left(\ln\left(\frac{1-\alpha}{\alpha}\right) + \sum_{i=1}^{D} \frac{({\mu_{i1}}^2 - {\mu_{i0}}^2)}{2\sigma^2}\right)$

Q1.2.

$$E(w, b) = -ln\left(\prod_{i=1}^{N} p\left(y^i = 0 \middle| x^i\right)^{1-y^i} * p\left(y^i = 1 \middle| x^i\right)^{y^i}\right)$$

$$= -\ln\left(\prod_{i=1}^{N} p\left(y^i = 0 \middle| x^i\right)^{1-y^i}\right) - \ln\left(\prod_{i=1}^{N} p\left(y^i = 1 \middle| x^i\right)^{y^i}\right)$$

$$= -(1 - y^i)\ln\left(\prod_{i=1}^{N} p\left(y^i = 0 \middle| x^i\right)\right) - y^i * \ln\left(\prod_{i=1}^{N} p\left(y^i = 1 \middle| x^i\right)\right)$$

$$= \sum_{i=0}^{N} -(1 - y^i)\ln\left(p\left(y^i = 0 \middle| x^i\right)\right) - \sum_{i=0}^{N} y^i \ln\left(p\left(y^i = 1 \middle| x^i\right)\right)$$

$$let\ z^i = w^T x_i + b, then$$

$$= \sum_{i=0}^{N} (1 - y^i)\ln\left(1 + \exp\left(-z^i\right)\right) + \sum_{i=0}^{N} (1 - y^i)\ln\left(\exp\left(z^i\right)\right) + \sum_{i=0}^{N} y^i \ln\left(1 + \exp\left(-z^i\right)\right)$$

$$= \sum_{i=0}^{N} \ln\left(1 + \exp\left(-z^i\right)\right) + \sum_{i=0}^{N} (1 - y^i) z^i$$

Then derive expressions for the derivatives of E with respect to each of the model parameters:

$$\frac{E(w, b)}{\partial b} = -\sum_{i=0}^{N} \frac{\exp\left(-\left(wx^i + b\right)\right)}{1 + \exp\left(-(wx^i + b)\right)} + \sum_{i=0}^{N} (1 - y^i) = -\sum_{i=0}^{N} \frac{\exp\left(-z^i\right)}{1 + \exp(-z^i)} + \sum_{i=0}^{N} (1 - y^i)$$

$$\frac{E(w,b)}{\partial w_j} = -\sum_{i=0}^{N} {x_j}^i \frac{\exp\left(-\left(wx^i + b\right)\right)}{1 + \exp\left(-\left(wx^i + b\right)\right)} + \sum_{i=0}^{N} \left(1 - y^i\right) {x_j}^i$$
$$= {x_j}^i \left(-\sum_{i=0}^{N} \frac{\exp\left(-z^i\right)}{1 + \exp(-z^i)} + \sum_{i=0}^{N} \left(1 - y^i\right) \right)$$

Q1.3.

$$p(w,b|D) \propto p(D|w,b)p(w,b)$$

$$p(w,b) = \prod_{i=1}^{D} N\left(w_i \middle| 0, {}^{1}/_{\lambda}\right) N\left(b \middle| 0, {}^{1}/_{\lambda}\right)$$

$$let\ z = wx + b$$

$$p(D|w,b) = \prod_{i=1}^{N} \frac{1}{(1 + \exp(-z))}^{y^i} \frac{\exp(-z)}{(1 + \exp(-z))}^{1-y^i}$$

$$p(w,b|D) \propto \left(\prod_{i=1}^{N} \frac{1}{(1 + \exp(-z))}^{y^i} \frac{\exp(-z)}{(1 + \exp(-z))}^{1-y^i} \right) \prod_{i=1}^{D} N\left(w_i \middle| 0, {}^{1}/_{\lambda}\right) N\left(b \middle| 0, {}^{1}/_{\lambda}\right)$$

$$L(w,b)$$
$$= -\ln\left(\left(\prod_{i=1}^{N} \frac{1}{(1 + exp(-z))}^{y^i} \frac{exp(-z)}{(1 + exp(-z))}^{1-y^i} \right) \left(\prod_{i=1}^{D} N\left(w_i \middle| 0, {}^{1}/_{\lambda}\right) N\left(b \middle| 0, {}^{1}/_{\lambda}\right) \right) \right)$$

$$= -\ln\left(\prod_{i=1}^{N} \frac{1}{(1 + exp(-z))}^{y^i} \frac{exp(-z)}{(1 + exp(-z))}^{1-y^i} \right) - \ln\left(\prod_{i=1}^{D} N\left(w_i \middle| 0, {}^{1}/_{\lambda}\right) N\left(b \middle| 0, {}^{1}/_{\lambda}\right) \right)$$

$$= E(w,b) - \sum_{i=0}^{D} \ln\left(N\left(w_i \middle| 0, {}^{1}/_{\lambda}\right) \right) - \ln\left(N\left(b \middle| 0, {}^{1}/_{\lambda}\right) \right)$$

$$= E(w,b) - \sum_{i=0}^{D} -\frac{{w_i}^2 * \lambda}{2} - D * \ln\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right) - \left(-\frac{b^2 * \lambda}{2}\right) - \ln\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)$$
$$= E(w,b) + \frac{\lambda}{2} \sum_{i=0}^{D} {w_i}^2 + \left(\frac{\lambda}{2}\right) b^2 - (D+1) \ln\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)$$

$$= E(w,b) + \frac{\lambda}{2}\sum_{i=0}^{D} {w_i}^2 + \frac{\lambda}{2}b^2 + c(\lambda)$$

Then derive expressions for the derivatives of L with respect to each of the model parameters:

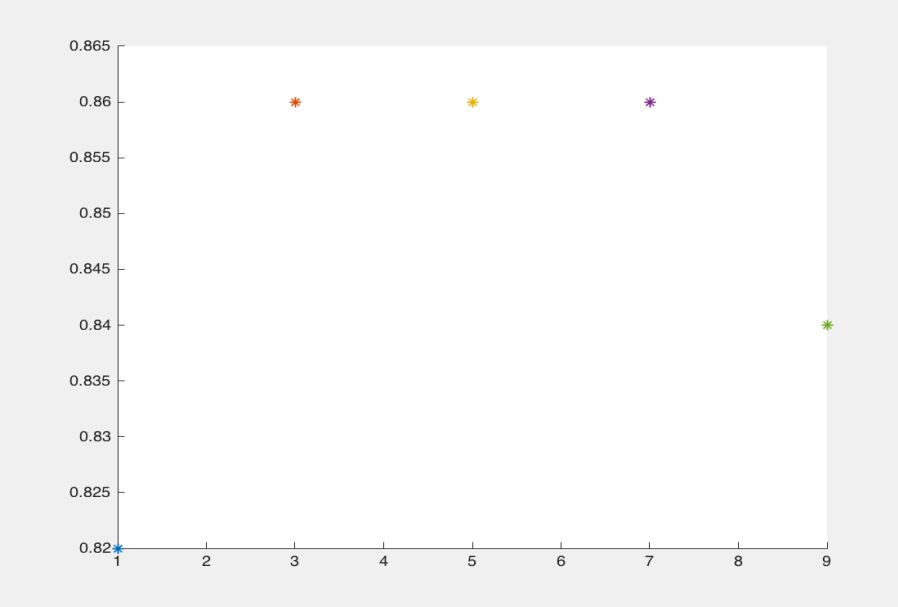
$$\frac{L(w,b)}{\partial b} = -\sum_{i=0}^{N} \frac{\exp(-z^i)}{1+\exp(-z^i)} + \sum_{i=0}^{N}(1-y^i) + \lambda b$$

$$\frac{L(w,b)}{\partial w_j} = -\sum_{i=0}^{N} {x_j}^i \frac{\exp\left(-(wx^i+b)\right)}{1+\exp(-(wx^i+b))} + \sum_{i=0}^{N}(1-y^i){x_j}^i + \lambda \sum_{i=0}^{D} w_i$$

$$= {x_j}^i\left(-\sum_{i=0}^{N} \frac{\exp(-z^i)}{1+\exp(-z^i)} + \sum_{i=0}^{N}(1-y^i)\right) + \lambda \sum_{i=0}^{D} w_i$$

Q2.1.

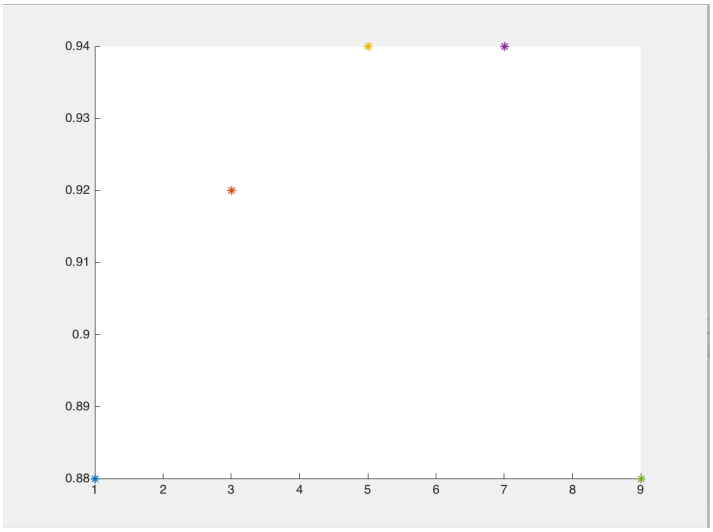
After running the script, we can get the figure:



form this picture, we can get that the classification rate of validation data reaches the highest value when k = 3,5,7

Since with larger k, error estimation tends to be more accurate, so we choose 7 here.

Test data:



form this picture, it shows that
k = 5, classification rate = 0.94
k = 7, classification rate = 0.94
k = 9, classification rate = 0.88
we can get that test performance correspond to the validation performance, because K-NN is a non-parametric method, the value of k is the only factor that influence K-NN.

Q2.2
Training data:
- Learning rate: 0.1
- Num_iteration: 80
- Initialize the weights: All zeros

Training data small:
- Learning rate: 0.1
- Num_iteration: 239
- Initialize the weights: All zeros

After running the code on two kinds of data:
Training data:
- Training set:
  - Cross entropy: 24.147825
  - Classification Error: 0
- Validation set:
  - Cross entropy: 13.438001
  - Classification Error: 0.1
- Test set:
  - Cross entropy: 12.2186
  - Classification Error: 0.08

Training data small:

Training set:

Cross entropy: 1.488820

Classification Error: 0

Validation set:

Cross entropy: 3.535339

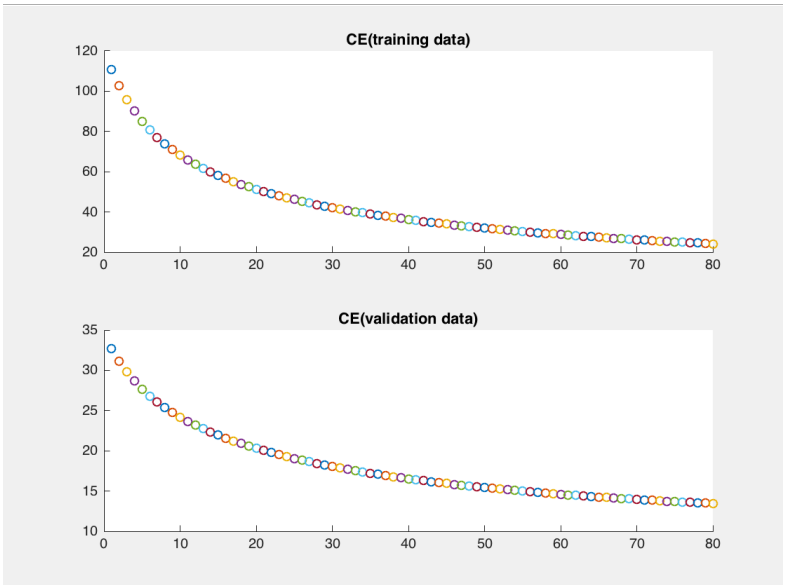
Classification Error: 0.38

Test set:

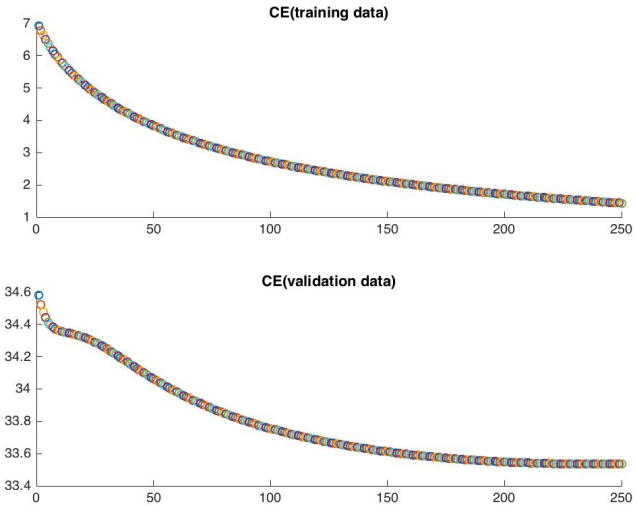
Cross entropy: 27.05

Classification Error: 0.24

Train data:



Train data small:

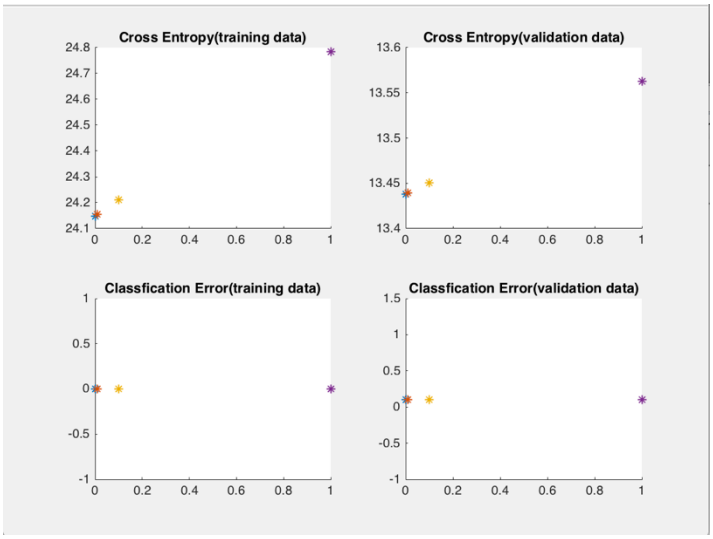


The results change when I change the settings, after I try many times, I find for large train data, when the number of iteration reaches 80, the cross entropy stops decreasing and rising up; for

small train data, 239 is the number that cross entropy reaches the lowest point. So, they are the best settings.

Q2.3

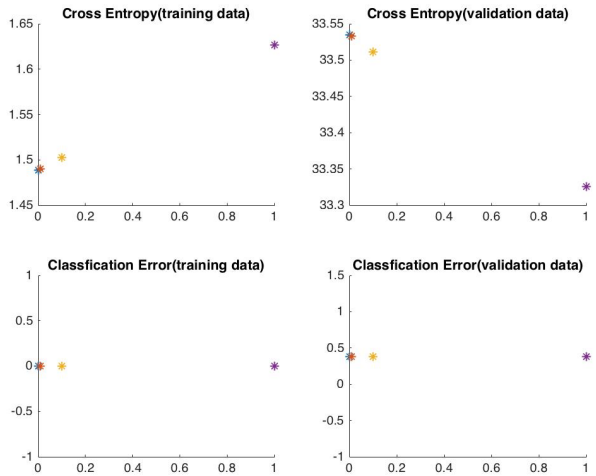
Train data:



When we use the training data, we can find that the tendency of cross entropy of training data and validation data are almost same, both rise up when lambda become larger. And training data is large enough to reduce the appearance of over fit. So, the lambda should be very small, in this case, 0.001 is better.
Test set:

Cross entropy: 12.21874
Classification Error: 0.08

Train data small:



When we use the training small data, we can find that the tendency of cross entropy of training data and validation data are totally different, one rise up and the other going down while lambda become larger, it should caused by the over fit problem, the training data here is small So, the lambda should be very large, in this case, 1is better.

Test set:

- Cross entropy: 27.0855
- Classification Error: 0.24

Comparison:

- Train data:
  - Penalty:
    - Cross entropy: 12.21874
    - Classification Error: 0.08
  - No Penalty:
    - Cross entropy: 12.2186
    - Classification Error: 0.08
- Train data small:
  - Penalty:
    - Cross entropy: 27.0855
    - Classification Error: 0.24
  - No Penalty:
    - Cross entropy: 27.05
    - Classification Error: 0.24

After comparing logistic regression with penalty and without penalty, we can find that the data of training data are almost the same, and for training data small, there exits some difference between penalized and not penalized, the reason should be the data is small, there exists some outlier, penalized can solve this problem.