EECS 545 HW3. Po-Jen Lai.

1. $\ell(w) = \sum_{i=1}^{n} -y_i \log h(x_i) - (1-y_i)\log(1-h(x_i))$

$= \sum_{i=1}^{n} -y_i w^T x_i + \log(e^{w^T x_i} + 1)$.

(a) $\frac{\partial \ell(w)}{\partial w_k} = \sum_{i=1}^{n} -x_k^{(i)}(y^{(i)} - \frac{1}{1+e^{-w^T x^{(i)}}}) \Rightarrow \nabla \ell(w) = \begin{bmatrix} \frac{\partial \ell(w)}{\partial w_1} \\ \vdots \\ \frac{\partial \ell(w)}{\partial w_m} \end{bmatrix}$ #

(b) $\frac{\partial^2 \ell(w)}{\partial w_k^2} = \sum_{i=1}^{n} (x_k^{(i)})^2 \cdot \frac{e^{-w^T x^{(i)}}}{(1+e^{-w^T x^{(i)}})^2}$, $\frac{\partial^2 \ell(w)}{\partial w_k w_j} = \sum_{i=1}^{n} x_k^{(i)} \cdot x_j^{(i)} \frac{e^{-w^T x^{(i)}}}{(1+e^{-w^T x^{(i)}})^2}$

$\Rightarrow H = \begin{bmatrix} \frac{\partial^2 \ell(w)}{\partial w_1^2} & \cdots & \frac{\partial^2 \ell(w)}{\partial w_1 \partial w_m} \\ \vdots & & \vdots \\ \frac{\partial^2 \ell(w)}{\partial w_m \partial w_1} & & \frac{\partial^2 \ell(w)}{\partial w_m^2} \end{bmatrix} = c \cdot u \cdot u^T$ where $c \in \mathbb{R}, c \geq 0, u \in \mathbb{R}^m$. #

$\Rightarrow$ for all $v \in \mathbb{R}^m$, $v^T H v = v^T (c u u^T) v = c \|v^T u\|^2 \Rightarrow H$ is PSD. #

2.

(a) $f(x; \alpha, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\alpha)^2}{2\sigma^2})$.

$\ell(\alpha, \sigma) = \frac{-n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\alpha)^2$

$\frac{\partial \ell}{\partial \alpha} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\alpha) = 0 \Rightarrow \hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} x_i$ #.

$\frac{\partial \ell}{\partial \sigma^2} = \frac{-n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\alpha)^2 \Rightarrow \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_j - \hat{\alpha})^2$ #.

(b) $\ell(\mu, \Sigma) = \frac{-nd}{2}\cdot\ln(2\pi) - \frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$.

$\frac{\partial \ell(\mu,\Sigma)}{\partial \mu} = \sum_{i=1}^{n}\Sigma^{-1}(x_i-\mu) \Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$. #

3.

(a) $H(X) + H(Y|X) = -\int p(x,y)\ln p(x)\,dx\,dy - \int p(x,y)\ln p(y|x)\,dx\,dy$.

$= -\int p(x,y)\cdot\ln p(x,y)\,dx\,dy = H(Y) + H(X|Y)$

$\Rightarrow H(X) - H(X|Y) = H(Y) - H(Y|X)$.

(b) $H(X|Y) = H(f(Y)|Y) = -\int p(f(y), y)\ln p(f(y)|y)\,dx\,dy$.

$\begin{cases} p(f(Y)=k \mid Y=y) = 1, & k = f(y) \\ p(f(Y)=k \mid Y=y) = 0, & k \neq f(y). \end{cases}$ $\Rightarrow H(X|Y) = 0$, same as $H(Y|X)$

$\Rightarrow I(X,Y) = H(X) - H(Y)$ #

3. (c) $\hat{p}(x) \triangleq \frac{1}{N}\sum_{i=1}^{N} I[x=x_i] = \frac{1}{N}\sum_{i=1}^{N}\delta(x-x_i)$

$D_{KL}(\hat{p}\|q) = -\int \hat{p}(x)\ln\frac{q(x|\theta)}{\hat{p}(x)}dx = -\int\hat{p}(x)\ln q(x|\theta)dx + \int\hat{p}(x)\ln\hat{p}(x)dx.$

$\arg\min_\theta D_{KL}(\hat{p}\|q) = \arg\max_\theta \int\hat{p}(x)\ln q(x|\theta)dx$

$\int\hat{p}(x)\ln q(x|\theta)dx = \frac{1}{N}\int\sum_{i=1}^{N} N\delta(x-x_i)\ln q(x|\theta)dx = \frac{1}{N}\sum_{i=1}^{N}\ln p(x_i|\theta).$

$\Rightarrow$ The minimum of $D_{KL}(\hat{p}\|q)$ is obtained by the maximum likelihood estimation #

(d) To maxmize $-\int p(x)\ln p(x)dx$. We consider $F(p,\lambda_1,\lambda_2,\lambda_3)$

$= -\int p(x)\ln(x)dx + \lambda_1(\int p(x)dx - 1) + \lambda_2(\int xp(x)dx - \mu) + \lambda_3(\int(x-\mu)^2p(x)dx - \sigma^2)$

$\frac{\partial F}{\partial p} = -1 - \ln p + \lambda_1 + \lambda_2 x + \lambda_3(x-\mu)^2 = 0 \Rightarrow p = \exp(\lambda_1 - 1 + \lambda_2 x + \lambda_3(x-\mu)^2)$

For $\int p(x)dx$ to be finite requires $\lambda_2 = 0$ and $\lambda_3 < 0$.

$\Rightarrow P(x) = e^a e^{-b(x-\mu)^2}$. where $a = \lambda_1 - 1$. $b = -\lambda_3 > 0$.

$\int p(x)dx = e^a\sqrt{\frac{\pi}{b}} = 1 \Rightarrow P(x) = \sqrt{\frac{b}{\pi}}e^{-b(x-\mu)^2}$. $\int(x-\mu)^2p(x)dx = \frac{1}{2b} = \sigma^2$

$\Rightarrow b = \frac{1}{2\sigma^2} \Rightarrow \underline{P(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp(\frac{-1}{2\sigma^2}(x-\mu)^2)}$. #

4. (a) $L(\beta,b) = L(w) = \sum_{i=1}^{n} c_i(y_i - \beta^T x_i - b)^2 = (Y-Xw)^T C(Y-Xw)$

where $Y\in\mathbb{R}^n$. $X\in\mathbb{R}^{n\times m}$. $w\in\mathbb{R}^m$. $C\in\mathbb{R}^{n\times n}$. $C_{ij} = c_i$. $i=j$; $C_{ij}=0$. $i\neq j$.

$\frac{\partial L}{\partial w} = 2(X^TCXw - X^TCY) = 0 \Rightarrow \underline{w = (X^TCX)^{-1}X^TCY}$ #

$\epsilon_i = y_i - x_i\cdot w. \Rightarrow l(w) = \frac{-n}{2}\ln(\sigma^2) = \frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}(y_i - x_iw)^2$

$\Rightarrow \arg\min_w \sum_{i=1}^{n}(y_i - x_iw) = \arg\max_w l(w).$

(b) $l(w) = \sum_{i=1}^{n}\frac{-1}{2}\ln(\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_i - x_iw)^2.$

$\Rightarrow \arg\min_w \sum_{i=1}^{n} c_i(y_i - x_iw) = \arg\max_w l(w)$. where $c_i = \frac{1}{\sigma_i^2}$

5. (a) $t^{(i)}(w^Tx^{(i)}+b) \geq 1-\xi_i \Rightarrow \xi_i \geq 1 - t^{(i)}(w^Tx^{(i)}+b)$ and $\xi_i \geq 0$.

$\therefore \min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \iff \min_{w,b}\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\max(0, 1-t^{(i)}(w^Tx^{(i)}+b))$

5. (b) $\xi_i^* = 1 - t^{(i)}((w^*)^T x^{(i)} + b^*)$. margin hyperplane : $(w^*)^T x + b = \pm 1$.

distance $= \frac{|(w^*)^T x^{(i)} + b - 1|}{\|w^*\|}$ for $t^{(i)} = 1$. $\frac{|(w^*)^T x^{(i)} + b + 1|}{\|w^*\|}$ for $t^{(i)} = -1$.

$\Rightarrow$ distance $= \frac{\xi_i^*}{\|w^*\|} \propto \xi_i^*$. #

(c)

$C \to \infty$. ξ_i has to be zero.

$\Rightarrow \min \frac{1}{2}\|w\|^2$ s.t. $t^{(i)}(w^T x^{(i)} + b) \geq 0$. $\Rightarrow$ hard-margin SVM. #.

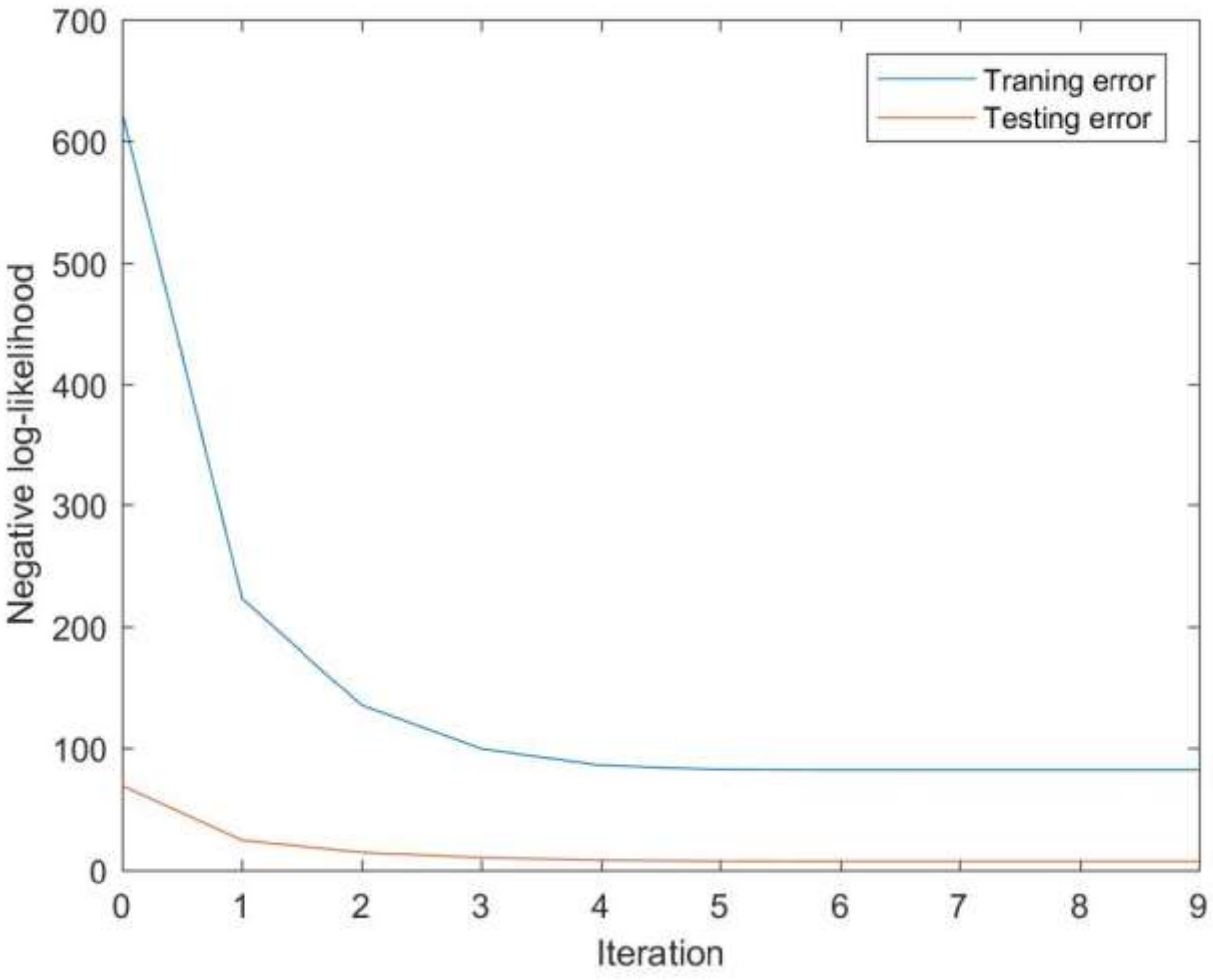
1.(c)

```
clear;clc;
filename = 'train_features.dat';
train_X = importdata(filename);
train_X = [ones(length(train_X),1) train_X];
filename = 'train_labels.dat';
train_Y = importdata(filename);
filename = 'test_features.dat';
test_X = importdata(filename);
test_X = [ones(length(test_X),1) test_X];
filename = 'test_labels.dat';
test_Y = importdata(filename);

eps = 1e-8;
w_update = zeros(3,1);
w = zeros(3,1);
diff = 1;
iter = 0;
training_error = [log_likelihood(train_Y,train_X,w)];
testing_error = [log_likelihood(test_Y,test_X,w)];
while diff >= eps
    iter = iter + 1;
    g = Gradient(train_Y,train_X,w);
    h_inv = inv(Hessian(train_X,w));
    w_update = w - h_inv*g;
    error1 = log_likelihood(train_Y,train_X,w_update);
    error2 = log_likelihood(test_Y,test_X,w_update);
    training_error = [training_error error1];
    testing_error = [testing_error error2];
    diff = abs(error1 - training_error(end-1));
    w = w_update;
end
x = 0:1:iter;
plot(x,training_error,x,testing_error);
xlabel('Iteration');
ylabel('Negative log-likelihood');
legend('Traning error','Testing error');

function l_w = log_likelihood(y,x,w)
    % y: n by 1, x: n by m+1, w: m+1 by 1
    l_w = -transpose(y)*x*w + sum(log(exp(x*w)+1));
end
```

```
function g = Gradient(y,x,w)
    linear_comb = -x*w;
    a = y - 1./(1+exp(linear_comb));
    g = -transpose(x)*a;
end
function H = Hessian(x,w)
    linear_comb = -x*w;
    a = exp(linear_comb);
    b = 1./(1 + exp(linear_comb)).^2;
    c = a.*b;
    xx = x.*repmat(c,1,3);
    H = transpose(x)*xx;
end
```



W = [-4.73878262951508, 4.40214932791691, -1.51521664732469]

Iteration = 9

5.(d)

```
clear;clc;
filename = 'diabetes_scale.csv';
data = csvread(filename);

train_X = data(1:500,2:end);
train_Y = data(1:500,1);
test_X = data(501:end,2:end);
test_Y = data(501:end,1);
C = linspace(0.1, 2, 20);
idx = crossvalind('Kfold', 500, 5);
rng(42);
%Soft-Margin
ce = zeros(20,1);
for i = 1:20
    SM_mdl = fitcsvm(train_X,train_Y,'KFold',5,'BoxConstraint',C(i));
    ce(i) = kfoldLoss(SM_mdl);
end
[~,I] = min(ce);
C_best = C(I);
SM_mdl = fitcsvm(train_X,train_Y,'BoxConstraint',C_best);
SM_label = predict(SM_mdl,test_X);
SM_accuracy = sum(SM_label==test_Y)/length(test_Y);

%Hard-Margin
HM_mdl = fitcsvm(train_X,train_Y,'BoxConstraint',1e6);
HM_label = predict(HM_mdl,test_X);
HM_accuracy = sum(HM_label==test_Y)/length(test_Y);
```

(i) Best C = 1.4, accuracy = 0.787313432835821

(ii) Accuracy = 0.317164179104478.

The C parameter means how much we want to avoid misclassifying each training example. For large values of C, the optimization will choose a smaller-margin hyperplane if that hyperplane does a better job of getting all the training points classified correctly. It makes the cost of misclassification high, thus forcing the algorithm to explain the input data stricter and potentially overfit. Therefore, it would cause a lower accuracy.

On the contrary, a small value of C will cause the optimizer to look for a larger-margin separating hyperplane, even if that hyperplane misclassifies more points. It makes the cost of misclassification low, allowing more of them due to a wider margin.