# Data mining: Homework 3

Vishvas Vasuki

October 26, 2009

## 1 1

$$
\begin{aligned}
E(w) &= -\sum_n [y_n \log(1+e^{-w^T x_n})^{-1} + (1-y_n)\log(1-(1+e^{-w^T x_n})^{-1})] \\
&= -\sum_n [y_n w^T x_n - w^T x_n - \log(1+e^{-w^T x_n})]
\end{aligned}
$$

$$
\begin{aligned}
\nabla_w(W(w)) &= -\sum_n [y_n x_n - x_n + (1+e^{-w^T x_n})^{-1} e^{-w^T x_n} x_n] \\
&= -\sum_n [y_n x_n - (1+e^{-w^T x_n})^{-1} x_n]
\end{aligned}
$$

Let X be the matrix whose ith column is $x_i$.

$$
\begin{aligned}
\frac{d^2 E(w)}{dw dw^T} &= \sum_n x_n x_n^T (1+e^{-w^T x_n})^{-2} e^{-w^T x_n} \\
&= XWX^T
\end{aligned}
$$

Above, W is diagonal, with $w_{n,n} = (1+e^{-w^T x_n})^{-2} e^{-w^T x_n} > 0$.

So, the Hessian matrix H is positive semidefinite, as $z^T H z = z^T XWX^T z = \left\|W^{1/2} X^T z\right\|_2^2 \geq 0$. So, E(w) is a convex function.

For E(w) to have a unique minimum, H should be positive definite. So, we want: $\forall z \neq 0 : \left\|W^{1/2} X^T z\right\|_2^2 > 0$. As $W^{1/2} > 0$, this happens when $X^T$, the $N \times d$ matrix of $\{x_i\}$ has full rank.

## 2 2

$$
\begin{aligned}
w_{t+1} &= w_t + y_t x_t \\
y_t(w^{*T} x_t) &\geq \gamma
\end{aligned}
$$

## 2.1 a

Base case:

$$\begin{aligned} w^{*T}w_1 &= w^{*T}w_0 + y_0 w^{*T}x_0 \\ &\geq \gamma \end{aligned}$$

Inductive hypothesis: Assume for t:

$$w^{*T}w_t \geq t\gamma$$

Induction: proof for t+1:

$$\begin{aligned} w^{*T}w_{t+1} &= w^{*T}w_t + y_t w^{*T}x_t \\ &\geq t\gamma + \gamma \\ &= (t+1)\gamma \end{aligned}$$

Hence proved by induction $\forall t > 0$.

## 2.2 b

Using the fact that $\|x_i\|^2 \leq R^2$, $w_0 = 0$ and triangle inequality:
Base case: t=1:

$$\begin{aligned} \|w_1\|_2^2 &= \|w_0 + y_0 x_0\|_2^2 \\ &= \|w_0\|^2 + \|y_0 x_0\|_2^2 + 2\langle w_0, y_0 x_0\rangle \\ &= \|x_0\|^2 \\ &\leq R^2 \end{aligned}$$

Inductive hypothesis:

$$\|w_t\|_2^2 \leq tR^2$$

Then:

$$\begin{aligned} \|w_{t+1}\|_2^2 &= \|w_t + y_t x_t\|_2^2 \\ &= \|w_t\|^2 + \|y_t x_t\|_2^2 + 2\langle w_t, x_t\rangle \\ &\leq tR^2 + \|x_t\|^2 \texttt{ as } \langle w_t, y_t x_t\rangle < 0 \\ &\leq (t+1)R^2 \end{aligned}$$

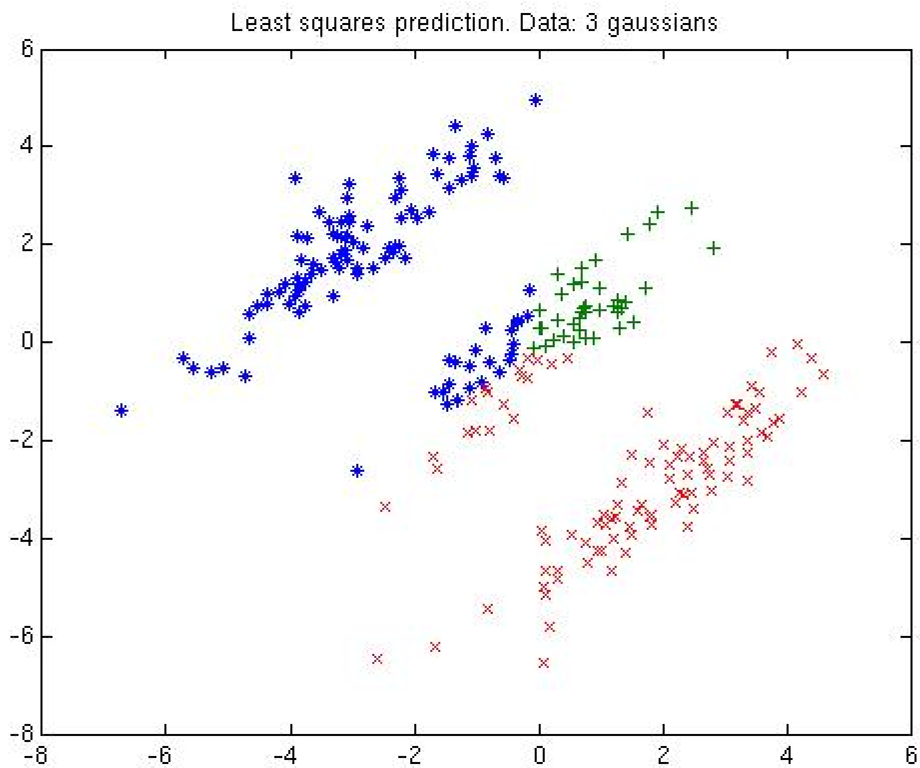
Hence, proved by induction.

### 2.3 c

$$\begin{aligned}
t\gamma &\leq w^{*T} w_t \\
&\leq \|w^*\| \|w_t\| \\
\frac{t\gamma}{\|w^*\|} &\leq \|w_t\| \\
(\frac{t\gamma}{\|w^*\|})^2 &\leq \|w_t\|^2 \leq tR^2 \\
t &\leq \frac{R^2 \|w^*\|^2}{\gamma^2}
\end{aligned}$$

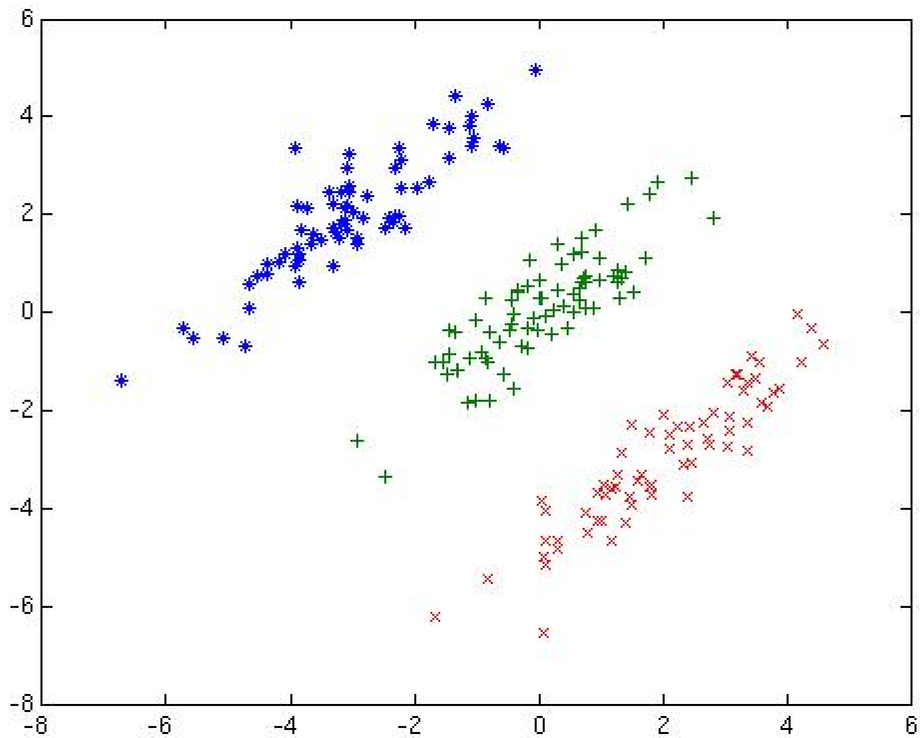
## 3 3

### 3.1 a




I observe that, in the middle cluster, many points are misclassified as belonging to other classes. This could probably be due to the sensitivity of least squares to outliers.

### 3.2 b




I observe that, using least squares regression, classification is almost perfect!

### 3.3 Code

```
%  load /u/vvasuki/vishvas/work/statistics/hw3/3gaussian/3gaussian

%  Common code
numLabels = numel(unique(labels));
%  returns 1,2,3

[numPoints,numFeatures] = size(X);


%  Add a feature.
X1 = X;
```

```
X1(:,3) = ones(numPoints,1);
numFeatures = numFeatures + 1;

%  Using 1 of 3 encoding for labels
Y = zeros(numPoints,numLabels);
label1spots = find(labels == 1);
label2spots = find(labels == 2);
label3spots = find(labels == 3);
Y(label1spots,1) = 1;
Y(label2spots,2) = 1;
Y(label3spots,3) = 1;

%  Do Least squares.
% Find W.
W = zeros(numFeatures,numLabels);
W = inv(X1'*X1)*(X1'*Y);
predictionLsq = X1*W;
%  Now,find the 1 of 3 vector with max correlation.
maxCorrelation = max(predictionLsq,[],2);
predictionLsq(:,1) = (predictionLsq(:,1) == maxCorrelation);
predictionLsq(:,2) = (predictionLsq(:,2) == maxCorrelation);
predictionLsq(:,3) = (predictionLsq(:,3) == maxCorrelation);

%  Plot actual data
%  plot(X(label1spots,1),X(label1spots,2),'*',X(label2spots,1),X(label2spots,2),'+',X(label3
label1spotsLsq = find(predictionLsq(:,1) == 1);
label2spotsLsq = find(predictionLsq(:,2) == 1);
label3spotsLsq = find(predictionLsq(:,3) == 1);

%  plot(X(label1spotsLsq,1),X(label1spotsLsq,2),'*',X(label2spotsLsq,1),X(label2spotsLsq,2),

%  Do logistic regression.
w = zeros(numFeatures*numLabels,1);
for k=1:100
W = reshape(w,numFeatures,numLabels);
expA_k = exp(X1*W);
InvSumExpA_k = inv(diag(sum(expA_k,2)));
Probability = InvSumExpA_k*expA_k;
%  Find the gradient
GradientMatrix = X1'*(Probability-Y);

%  Find the Hessian
H = zeros(numFeatures*numLabels);
I = eye(numLabels*numLabels);
for i = 1:numLabels
for j = 1:numLabels
```

```
wt = zeros(numPoints,1);
for n= 1:numPoints
wt(n) = Probability(n,i)*(I(i,j) -Probability(n,j));
H((i-1)*numFeatures + 1:i*numFeatures,(j-1)*numFeatures + 1:j*numFeatures)
= H((i-1)*numFeatures + 1:i*numFeatures,(j-1)*numFeatures + 1:j*numFeatures)
+ wt(n)*X1(n,:)'*X1(n,:);
end
end
end
w = w -inv(H)*GradientMatrix(:);
%      w
end

label1spotsPr = find(Probability(:,1) == 1);
label2spotsPr = find(Probability(:,2) == 1);
label3spotsPr = find(Probability(:,3) == 1);

plot(X(label1spotsPr,1),X(label1spotsPr,2),'*',X(label2spotsPr,1),X(label2spotsPr,2),'+',X(l
```

# 4 4

## 4.1 a

Let $\sum_{i=3}^{N} y_i\alpha_i = k$. Then, from condition 1, $y_1\alpha_1 + y_2\alpha_2 = -k$. The same condition should hold when $(\alpha_1, \alpha_2)$ are modified to $(\bar{\alpha}_1, \bar{\alpha}_2)$. So, we have $y_1\bar{\alpha}_1 + y_2\bar{\alpha}_2 = -k$.

So, $y_1\alpha_1 + y_2\alpha_2 = y_1\bar{\alpha}_1 + y_2\bar{\alpha}_2 = -k$.

When $y_1 = y_2$, multiplying both sides by $y_1$, we get: $\alpha_1 + \alpha_2 = \bar{\alpha}_1 + \bar{\alpha}_2$. From condition 2, all these are non-negative. So, $\bar{\alpha}_2 = \alpha_1 + \alpha_2 - \bar{\alpha}_1 \geq \alpha_1 + \alpha_2$.

When $y_1 \neq y_2$, as $y_1y_2 = -1$, multiplying both sides by $y_2$, we get: $\alpha_2 - \alpha_1 = -\bar{\alpha}_1 + \bar{\alpha}_2$. From condition 2, all these are non-negative. So, $\hat{\alpha}_2 \geq \alpha_2 - \alpha_1$.

## 4.2 b

As noted earlier, $y_1\alpha_1 + y_2\alpha_2 = y_1\bar{\alpha}_1 + y_2\bar{\alpha}_2 = -k$. Using $s = y_1y_2$, by multiplication by $y_1$: $\alpha_1 + y_1y_2\alpha_2 = -y_1k$, which yields $\alpha_1 + s\alpha_2 = -y_1k = \gamma$.

Let $v_i = \sum_{j=3}^{N} y_j\alpha_jK_{i,j}$ for i = 1 or 2.

Using the above identities, we get:

$$\begin{aligned}
\sum_{i=1}^{N} \alpha_i &= \gamma - s\alpha_2 + \alpha_2 + k'' \\
\sum_i \sum_j y_iy_j\alpha_i\alpha_jK_{i,j} &= K_{1,1}(\gamma - s\alpha)^2 + K_{2,2}\alpha_2^2 + \\
& \quad 2sK_{1,2}(\gamma - s\alpha_2)\alpha_2 + 2v_1y_1(\gamma - s\alpha_2) + 2y_2\alpha_2v_2 + k'
\end{aligned}$$

Above, $k' = \sum_{i=3}^{N} \sum_{i=3}^{N} y_i y_j \alpha_i \alpha_j K_{i,j}$ and $k''$ are constants wrt $\alpha_2$, but can depend on other $\alpha_i$.

$$\begin{aligned}
\therefore w(\alpha_2) &= \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j y_i y_j \alpha_i \alpha_j K_{i,j} \\
&= \gamma - s\alpha_2 + \alpha_2 + k'' - 2^{-1}[K_{1,1}(\gamma - s\alpha)^2 + K_{2,2}\alpha_2^2 + \\
&\quad 2sK_{1,2}(\gamma - s\alpha_2)\alpha_2 + 2v_1 y_1(\gamma - s\alpha_2) + 2y_2\alpha_2 v_2 + k']
\end{aligned}$$

### 4.3 c

$\frac{w(\alpha_2)}{d\alpha_2} = -s+1+sK_{1,1}(\gamma - s\alpha_2) - K_{2,2}\alpha_2 - sK_{1,2}(\gamma - 2s\alpha_2) + y_1 s v_1 2^{-1} - y_2 v_2 2^{-1}$.

Setting this to 0, and using $s^2 = 1, y_1 s = y_2$ and $d_{1,2} = K_{1,1} + K_{2,2} - 2K_{1,2}$, we get: $d_{1,2}\bar{\alpha}_2 = -s + 1 + sK_{1,1}\gamma - sK_{1,2}\gamma + 2^{-1}y_2 v_1 - 2^{-1}y_2 v_2$.

We confirm that this is the maximum by the following: $\frac{d^2 w(\alpha_2)}{d\alpha_2^2} = -K_{1,1} - K_{2,2} + 2K_{1,2} = -\|x_1 - x_2\|^2 \leq 0$.

### 4.4 d

Take $E_i = \sum_j \alpha_j y_j K_{i,j} + w_0 - y_i$ for i = 1 or 2.

$$\begin{aligned}
E_1 - E_2 &= y_2 - y_1 + \sum_j \alpha_j y_j K_{1,j} - \sum_j \alpha_j y_j K_{2,j} && (1) \\
&= y_2 - y_1 + \alpha_1 y_1 K_{1,1} + \alpha_2 y_2 K_{1,2} + v_1 && (2) \\
&\quad -\alpha_1 y_1 K_{2,1} - \alpha_2 y_2 K_{2,2} - v_2 && (3) \\
&= y_2 - y_1 + \gamma y_1 K_{1,1} + v_1 - v_2 - \gamma y_1 K_{2,1} - y_2 \alpha_2 K_{1,1} && (4) \\
&\quad +\alpha_2 y_2 K_{1,2} + y_2 \alpha_2 K_{2,1} - \alpha_2 y_2 K_{2,2} && (5) \\
& && (6)
\end{aligned}$$

Above, we have used the definitions of $\gamma$ and $v_1, v_2$ seen in part c, and the fact that $sy_1 = y_2$.

So, $y_2(E_1 - E_2) = 1 - s + sK_{1,1}\gamma + y_2 v_1 - y_2 v_2 - s\gamma K_{1,2} - \alpha_2(K_{1,1} - 2K_{1,2} + K_{2,2})$.

Thus, comparing this with the equation for $\bar{\alpha}$ derived in part c, we get: $\bar{\alpha}_2 = \alpha_2 + \frac{y_2(E_1 - E_2)}{d_{1,2}}$.

#### 4.4.1 Taking care of the constraints from part a

The expression for $\bar{\alpha}_2$ above gives the best step length for maximizing $W(\alpha_2)$ if we don't care about the constraints mentioned in part a. Also note that $W(\alpha_2)$ is a concave function, as we saw in part c. So, if the constraints do not allow us to select the maximal $\bar{\alpha_2}$, we should select $\bar{\alpha}_2$ to be a value within the feasible region, which is closest to the maximal $\bar{\alpha}_2$. Below, we use this fact when incorporating the constraints from part a.

When $y_1 = y_2$: $\bar{\alpha}_2 \geq \alpha_1 + \alpha_2$. So, to satisfy this constraint, we pick: $\bar{\alpha}_2 = \min(\bar{\alpha}_2, \alpha_1 + \alpha_2)$.

When $y_1 \neq y_2$: $\hat{\alpha}_2 \geq \alpha_2 - \alpha_1$. So, $\bar{\alpha}_2 = \max(\bar{\alpha}_2, \alpha_2 - \alpha_1)$.

#### 4.4.2 Also satisfying the nonnegativity constraint

We also want to impose the constraint: $\bar{\alpha} \geq 0$ to the values of $\bar{\alpha}_2$ specified above. Using the same reasoning, we want to pick a value within the feasible region (where all constraints are satisfied), which is closest to the maximal $\bar{\alpha}_2$ we derived earlier.

So, when $y_1 = y_2$: $\bar{\alpha}_2 = \max(0, \min(\bar{\alpha}_2, \alpha_1 + \alpha_2))$

So, when $y_1 \neq y_2$: $\bar{\alpha}_2 = \max(0, \bar{\alpha}_2, \alpha_2 - \alpha_1)$

#### 4.4.3 Finding $\bar{\alpha}_1$

We saw earlier that $y_1\alpha_1 + y_2\alpha_2 = y_1\bar{\alpha}_1 + y_2\bar{\alpha}_2$. Multiplying both sides by $y_1$ and solving for $\bar{\alpha}_1$, we get: $\bar{\alpha}_1 = \alpha_1 + y_1 y_2(\alpha_2 - \bar{\alpha}_2)$.