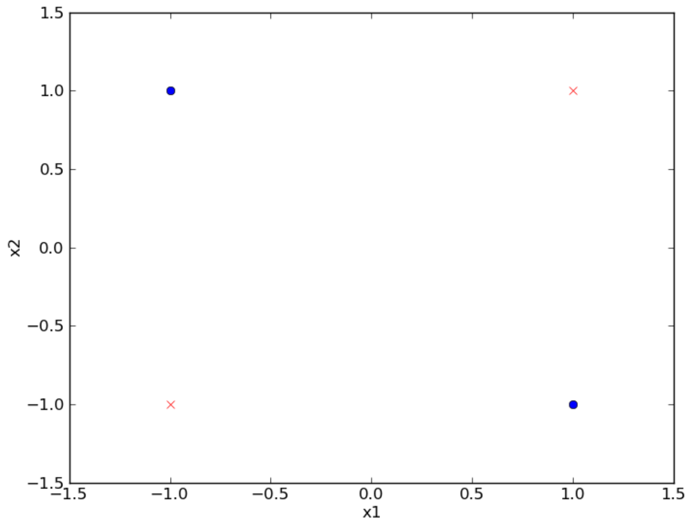
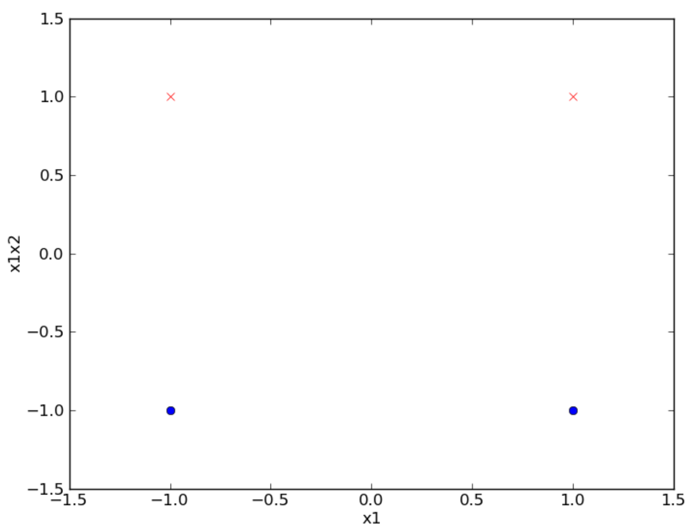
Franklin Hu, Sunil Pedapudi
SID: 20157715
CS 194-10
2011-09-19
Assignment 2

1. Kernels

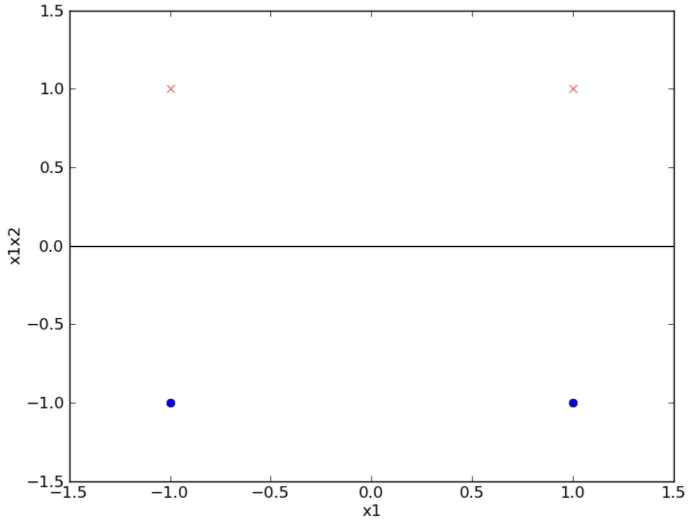
   (a) Original input



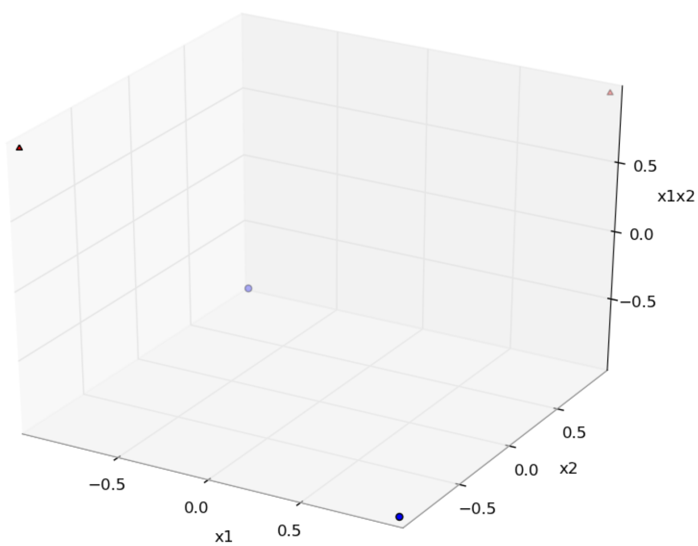
Input mapped onto space consisting of x_1 and x_1x_2:



The maximum margin separator is the line $x_1x_2 = 0$.



The separating line on the original input space is a plane that rests at $x_1 x_2 = 0$.



(b)

$$\begin{aligned} (x_1 - a)^2 + (x_2 - b)^2 - r^2 =& 0 \\ x_1^2 - 2ax_1 + a^2 + x_2^2 - 2bx_2 + b^2 - r^2 =& 0 \end{aligned}$$

If we let $x = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix}$:

$$\begin{pmatrix} -2a & -2b & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} + (a^2 + b^2 - r^2) = 0$$

If we would like to classify points inside the circle with -1 and points outside the circle with $+1$, we can adjust this to be:

$$\hat{y} = sign(\begin{pmatrix} -2a & -2b & 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} + (a^2 + b^2 - r^2))$$

(c)

$$\begin{aligned}
c(x_1 - a)^2 + d(x_2 - b)^2 - 1 =& 0 \\
cx_1^2 - 2acx_1 + ca^2 + dx_2^2 - 2bdx_2 + db^2 - 1 =& 0 \\
\begin{pmatrix} ca^2 + db^2 - 1 & -2ac & -2bd & c & d & 0 \end{pmatrix} \begin{pmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \\ x_1 x_2 \end{pmatrix} =& 0
\end{aligned}$$

2. Logistic Regression
   Given:

$$L(w) = -\sum_{i=1}^{N} log(\frac{1}{1 + e^{y_i(w^T x_i + b)}}) + \lambda \|w\|_2^2$$

(a)

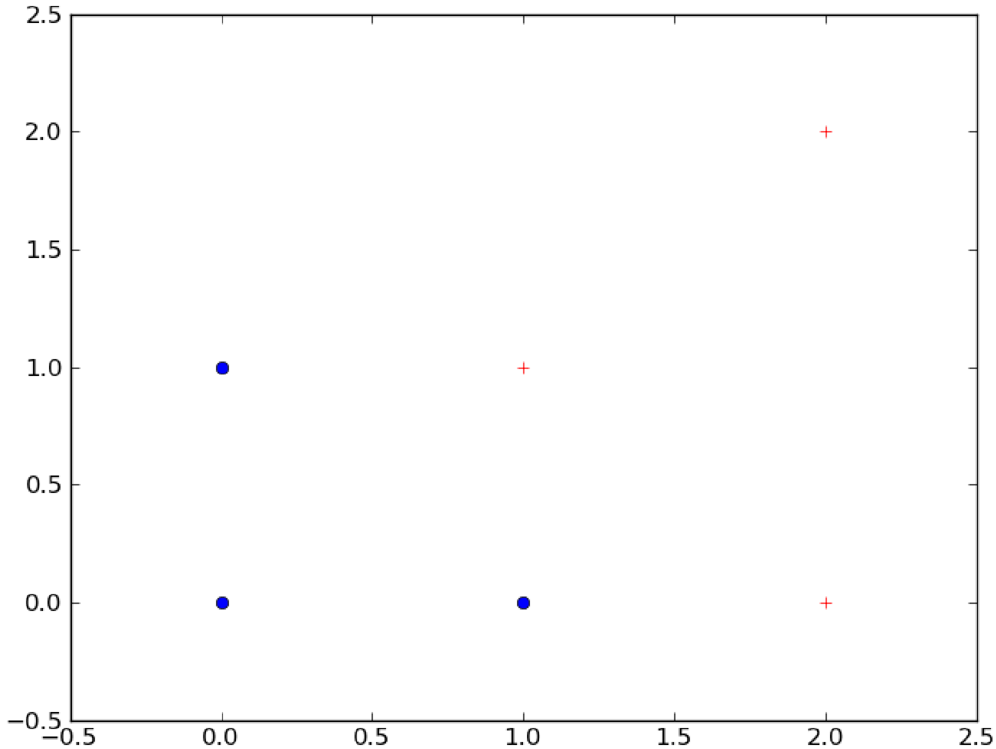
$$\begin{aligned}
\frac{\partial L}{\partial w_j} =& -\sum_{i=1}^{N}(1 + e^{y_i(w^T x_i + b)}) \cdot -1 \cdot (1 + e^{y_i(w^T x_i + b)})^{-2}(e^{y_i(w^T x_i + b)}) \cdot x_j y_j + \frac{\partial}{\partial w_j}(\lambda \|w\|_2^2) \\
=& -\sum_{i=1}^{N} \frac{-e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})} \cdot x_j y_j + 2\lambda w_j \\
=& x_j y_j \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})} + 2\lambda w_j
\end{aligned}$$

(b)

$$\begin{aligned}
\frac{\partial^2 L}{\partial w_j \partial w_k} =& \frac{\partial L}{\partial w_k}(x_j y_j \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})} + 2\lambda w_j) \\
=& x_j y_j \sum_{i=1}^{N} \frac{(1 + e^{y_i(w^T x_i + b)}) \cdot \frac{\partial L}{\partial w_k}(e^{y_i(w^T x_i + b)}) - e^{y_i(w^T x_i + b)} \cdot \frac{\partial L}{\partial w_k}(1 + e^{y_i(w^T x_i + b)})}{(1 + e^{y_i(w^T x_i + b)})^2} \\
=& x_j y_j \sum_{i=1}^{N} \frac{(1 + e^{y_i(w^T x_i + b)}) \cdot \frac{\partial L}{\partial w_k}(e^{y_i(w^T x_i + b)}) - e^{y_i(w^T x_i + b)} \cdot \frac{\partial L}{\partial w_k}(e^{y_i(w^T x_i + b)})}{(1 + e^{y_i(w^T x_i + b)})^2} \\
=& x_j y_j \sum_{i=1}^{N} \frac{(1 + e^{y_i(w^T x_i + b)} - e^{y_i(w^T x_i + b)}) \cdot \frac{\partial L}{\partial w_k}(e^{y_i(w^T x_i + b)})}{(1 + e^{y_i(w^T x_i + b)})^2} \\
=& x_j y_j \sum_{i=1}^{N} \frac{\frac{\partial L}{\partial w_k}(e^{y_i(w^T x_i + b)})}{(1 + e^{y_i(w^T x_i + b)})^2} \\
=& x_j y_j \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)} x_k y_k}{(1 + e^{y_i(w^T x_i + b)})^2} \\
=& x_j x_k y_j y_k \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2}
\end{aligned}$$

3. Training data

(a) Yes the classes {+,-} are linearly separable. The - class is represented by circles in the graph below.



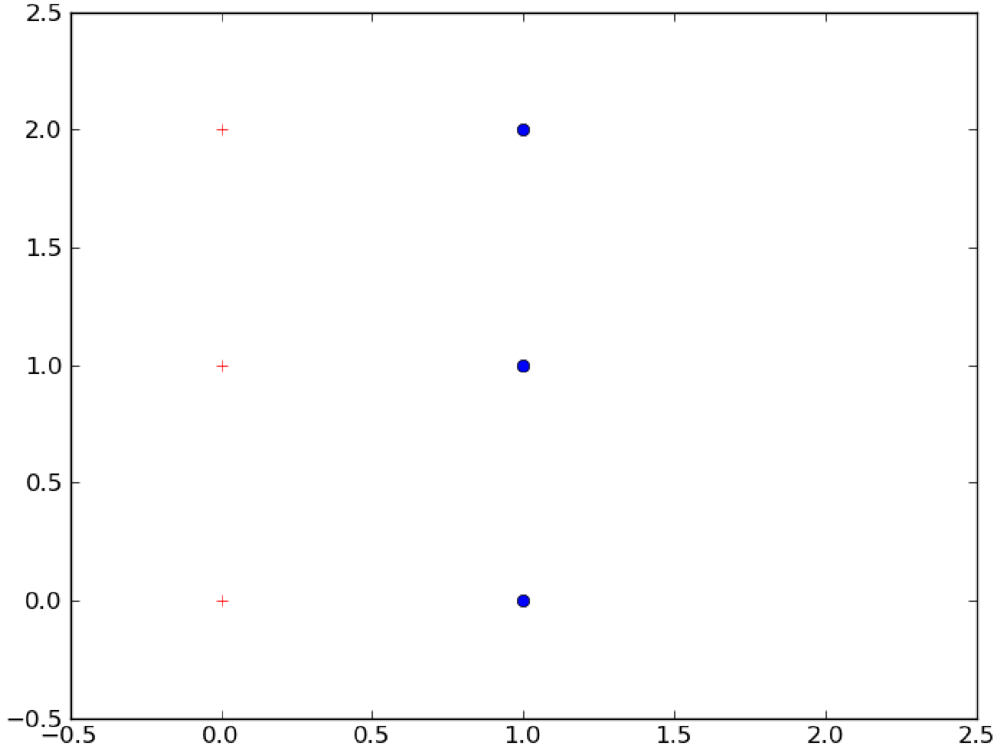
(b) The best hyperplane by inspection is:

$$
\begin{aligned}
x_2 =& - x_1 + 1.5 \\
x_1 + x_2 - 1.5 =& 0 \\
\begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} - 1.5 =& 0
\end{aligned}
$$

So therefore $w = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $b = -1.5$. The support vectors are $(1,0), (0,1), (2,0), (1,1)$.

(c) If we remove a support vector, then the optimal margin will increase since there are fewer constraints.

(d) The answer for (c) is not always true. Consider if we have a class + with points $(0,0), (0,1), (0,2)$ and a class - with points $(1,0), (1,1), (1,2)$. If we remove either $(0,1)$ or $(1,1)$, the best hyperplane does not change and thus the optimal margin remains the same.



4. 3 point dataset
5. Seismic waves

   (a) phase, iphase frequencies

   - phase

     –

phase	absolute frequency	relative frequency
Lg	1594	0.017811
P	61779	0.690322
PKP	5974	0.066754
Pg	403	0.004503
Pn	10762	0.120255
Rg	11	0.000123
S	4685	0.052350
Sn	4285	0.047881

   - iphase

–

iphase	absolute frequency	relative frequency
Lg	2171	0.024259
N	10683	0.119372
P	50815	0.567810
Pg	5291	0.059122
Pn	12610	0.140905
Px	365	0.004079
Rg	444	0.004961
Sn	318	0.003553
Sx	4179	0.046696
tx	2617	0.029243

(b) Confusion matrix (empty cells are zero)

•

		phase										
		Lg	PKP	P	S	Rg	Sn	Pn	Pg		Total	Accuracy (%)
iphase	Lg	293	2	114	860	5	859	34	4		2171	13.496
	Sx	297	61	971	1257	3	1191	393	6		4179	58.579
	tx	17	383	2039	26		18	111	23		2617	0
	Px	30	13	101	46		68	61	46		365	60.548
	N	431	564	6097	1278	1	1133	1149	30		10683	0
	P	105	4586	42600	336		153	2993	42		50815	83.834
	Rg	83		8	182	2	169				444	0.450
	Pg	218	120	2716	318		303	1509	107		5291	2.022
	Pn	95	244	7123	243	256		4504	145		12610	35.717
	Sn	25	1	10	139		135	8			318	42.453

(c) Top stations

i. 7: 8751 detections
ii. 24: 5794 detections
iii. 3: 2677 detections
iv. 80: 2528 detections
v. 19: 2478 detections
vi. 38: 2429 detections
vii. 63: 2411 detections
viii. 12: 2343 detections
ix. 74: 2265 detections
x. 65: 2227 detections

(d) Data munging

i.

station	classifier accuracy (%)	
7	88.08	
24	92.15	
3	92.02	
80	88.10	
19	88.87	
38	90.69	
63	88.18	
12	88.77	
74	89.56	
65	92.51	

(e) Optimal c

station	c	accuracy
7	0.42	93.23
24	0.5	86.64
	0.1	87.43
	0.2	87.03
	0.05	87.98
	0.01	89.79
	0.001	92.04
	0.0001	92.20
	0	92.15