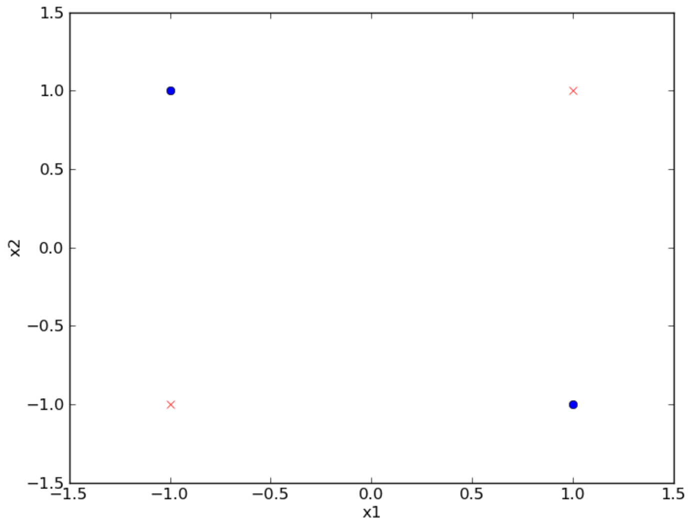
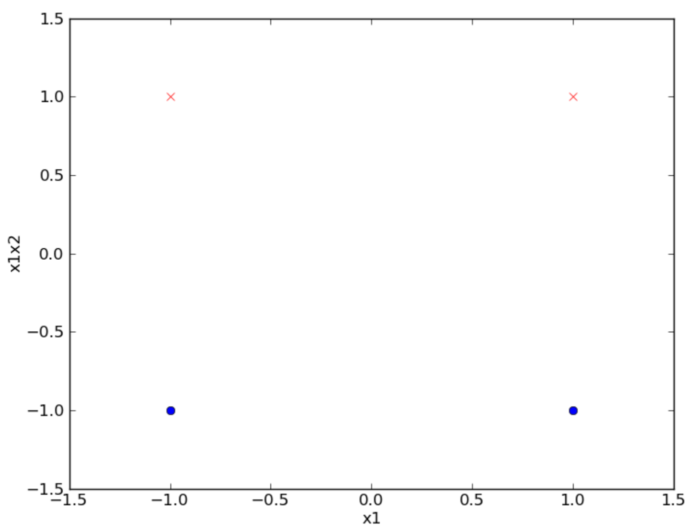
Franklin Hu, Sunil Pedapudi
SID: 20157715SID: 20247144
CS 194-10
2011-09-19
Assignment 2

1. Kernels

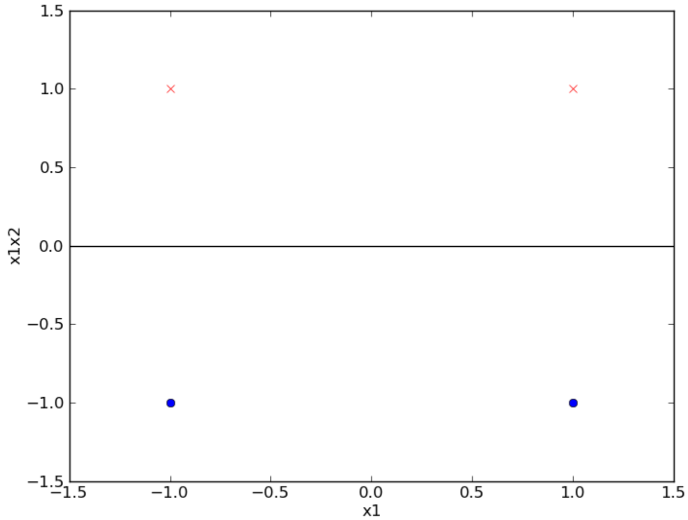
   (a) Original input



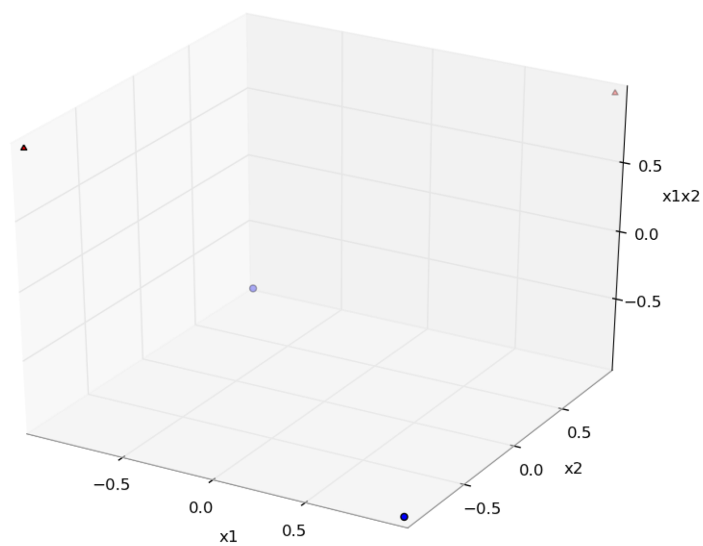
Input mapped onto space consisting of x_1 and x_1x_2:

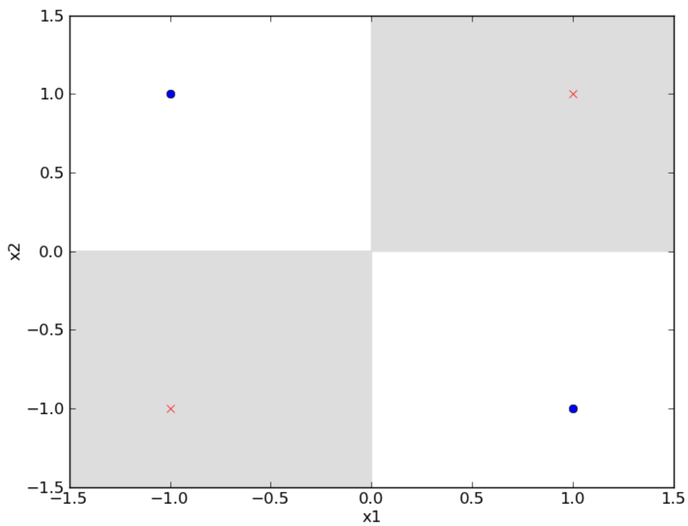


The maximum margin separator is the line $x_1x_2 = 0$.



The separating line on the original input space is a plane that rests at $x_1x_2 = 0$.





(b) Given

$$(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$$
$$x_1^2 - 2ax_1 + a^2 + x_2^2 - 2bx_2 + b^2 - r^2 = 0$$

let us pose the following:

$$\mathbf{w} = [-2a, -2b, 1, 1]$$
$$\mathbf{x} = [x_1, x_2, x_1^2, x_2^2]$$
$$\beta = a^2 + b^2 - r^2$$
$$\mathbf{w^T x} + \beta > 0, \text{ if } \mathbf{x} \text{ escapes the circle region}$$
$$\mathbf{w^T x} + \beta < 0, \text{ if } \mathbf{x} \text{ occupies the circle region}$$
$$\mathbf{w^T x} + \beta = 0, \text{ if } \mathbf{x} \text{ demarcates the circle region}$$

We then let $y_i = -1$ if $\mathbf{x}$ occupies the region inside the circle; $y_i = 1$ otherwise. Then, to satisfy the separability constraint, we note that

$$y_i(\mathbf{w^T x} + \beta) > 0, \forall i$$

Thus, we show that in feature space (x_1, x_2, x_1^2, x_2^2) , the region defined by $(x_1 - a)^2 + (x_2 - b)^2 - r^2 = 0$ is linearly separable.

(c) Given

$$\begin{aligned} K(\mathbf{u}, \mathbf{v}) &= (1 + \mathbf{u^T v})^2 \\ &= 1 + 2\mathbf{u^T v} + (\mathbf{u^T v})^2 \\ &= 1 + 2u_1v_1 + 2u_2v_2 + (u_1^2v_1^2 + 2u_1v_1u_2 + v_2 + u_2^2v_2^2) \end{aligned}$$

Let us realize that this kernel suggests a feature space $[1, \sqrt{2}u_1, \sqrt{2}u_2, u_1^2, u_2^2, \sqrt{2}u_1u_2]$. For simplicity, we adapt this feature space more generally as $[1, x_1, x_2, x_1^2, x_2^2, x_1x_2]$ and drop the constant multipliers as suggested. Then, given an ellipse is defined by

$$c(x_1 - a)^2 + d(x_2 - b)^2 = 1$$
$$cx_1 - 2acx_1 + ca_2^2 + dx_2^2 - 2dbx_2 + db^2 - 1 = 0$$

we wish to recycle the proof from 1b. To do this, we form the following vector

$$\mathbf{w} = [ca^2 + db^2, -2ac, -2db, c, d, 0]$$

Then, we define $y_i = -1$ if a point lies within the ellipse, $y_i = 1$ otherwise. We simply adopt the inequalities from 1b and claim that

$$\mathbf{w^T x} + \beta > 0, \text{ if } \mathbf{x} \text{ escapes the ellipse region}$$
$$\mathbf{w^T x} + \beta < 0, \text{ if } \mathbf{x} \text{ occupies the ellipse region}$$
$$\mathbf{w^T x} + \beta = 0, \text{ if } \mathbf{x} \text{ demarcates the ellipse region}$$

which satisfies the separability constraint $y_i(\mathbf{w^T x} + \beta) > 0, \forall i$

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)

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

(b)

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

Since $y_i^2 = 1$, we simply rewrite this as

$$\sum_{i=1}^{N} x_{ij}x_{ik} \cdot \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \tag{1}$$

(c) Then, we wish to show

$$\mathbf{a^T H a} \equiv \sum_{j,k} a_j a_k H_{j,k} \geq 0$$

Note that summation (1) indicates the j, k^{th} element of the Hessian which allows us to rewrite the the above inequality as

$$\begin{aligned}
\sum_{j,k} a_j a_k H_{j,k} &= \sum_{j,k} a_j a_k \sum_{i=1}^{N} x_{ij}x_{ik} \cdot \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \\
&= \sum_{j,k} a_j a_k \sum_{i=1}^{N} x_{ij}x_{ik} \cdot \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \sum_{j,k} a_j a_k x_{ij} x_{ik} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \sum_{j} a_j x_{ij} \sum_{k} a_k x_{ik} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \sum_{j,k} \mathbf{a^T x} \sum_{k} a_k x_{ik} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \sum_{j,k} \mathbf{a^T x a^T x} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \sum_{j,k} \mathbf{a^T x a^T x} \\
&= \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \cdot (\mathbf{a^T x})^2 \geq 0
\end{aligned}$$

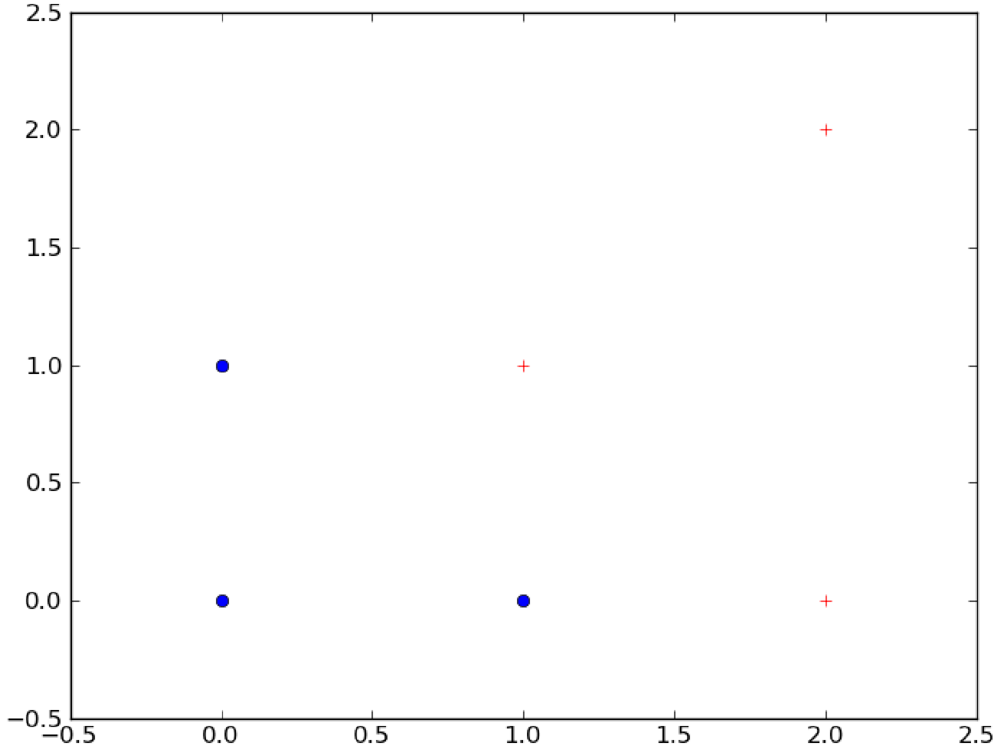
We show this summation is non-negative by showing each component of the summation is non-negative. Consider

$$\begin{aligned}
\sum_{i=1}^{N} \frac{\alpha}{\beta} \cdot \epsilon = \sum_{i=1}^{N} \frac{e^{y_i(w^T x_i + b)}}{(1 + e^{y_i(w^T x_i + b)})^2} \cdot (\mathbf{a^T x})^2 \geq 0 \text{ Then, we realize that} \\
\alpha = e^{y_i(w^T x_i + b)} > 0, \text{since } e^z \text{ is always positive} \\
\beta = (1 + e^{y_i(w^T x_i + b)})^2 > 0 \\
\epsilon = (\mathbf{a^T x})^2 \geq 0
\end{aligned}$$

Therefore, L is convex.

3. Training data

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



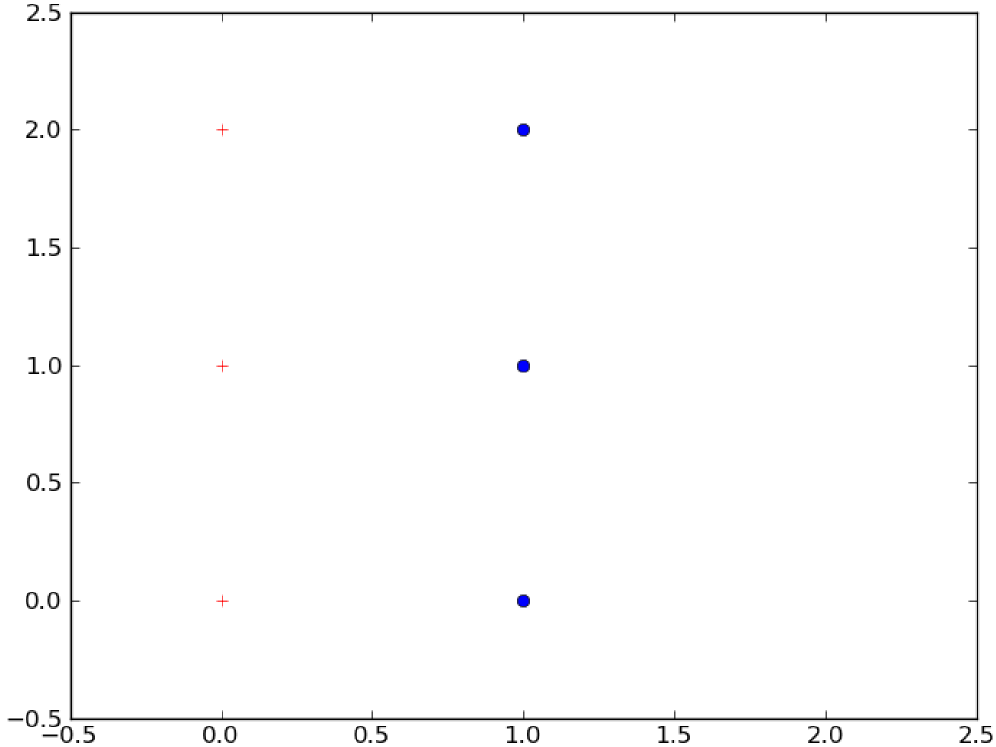
(b) The best hyperplane by inspection is:

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

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

   (a) No

   (b)

$$
\begin{aligned}
\phi(x_1) &= [1, 0, 0]^T \\
\phi(x_2) &= [1, -\sqrt{2}, 1]^T \\
\phi(x_3) &= [1, \sqrt{2}, 1]^T
\end{aligned}
$$

   Yes, this is linearly separable with the hyperplane $x^2 = \frac{1}{2}$

   (c) Let

$$
\begin{aligned}
x_1 &= 0 \\
x_2 &= -1 \\
x_3 &= 1 \\
y_1 &= 1 \\
y_2 &= -1 \\
y_3 &= -1 \\
\Lambda(w_1, w_2, w_3, b, \lambda, \mu, \varepsilon) &= \frac{1}{2}\|w\|_2^2 \\
&\quad + \lambda(y_1(w_1 + b) - 1) \\
&\quad + \mu(y_2(w_1 - \sqrt{2}w_2 + w_3 + b) - 1) \\
&\quad + \varepsilon(y_3(w_1 + \sqrt{2}w_2 + w_3 + b) - 1)
\end{aligned}
$$

Then, using the method of Lagrange multipliers,

$$\frac{\partial \Lambda}{\partial w_1} = \frac{1}{2}w_1^2 + \lambda - \mu - \varepsilon = 0 \quad (2)$$

$$\frac{\partial \Lambda}{\partial w_2} = \frac{1}{2}w_2^2 + \sqrt{2}\mu - \sqrt{2}\varepsilon = 0 \quad (3)$$

$$\frac{\partial \Lambda}{\partial w_3} = \frac{1}{2}w_3^2 - \mu - \varepsilon = 0 \quad (4)$$

$$\frac{\partial \Lambda}{\partial b} = \lambda - \mu - \varepsilon = 0 \quad (5)$$

$$\frac{\partial \Lambda}{\partial \lambda} = w_1 + b - 1 = 0 \quad (6)$$

$$\frac{\partial \Lambda}{\partial \mu} = -(w_1 - \sqrt{2}w_2 + w_3 + b) - 1 = 0 \quad (7)$$

$$\frac{\partial \Lambda}{\partial \varepsilon} = -(w_1 + \sqrt{2}w_2 + w_3 + b) - 1 = 0 \quad (8)$$

We inspect these equations to arrive at the following conclusions:
From (4), we know $\lambda - \mu - \varepsilon = 0$ so in (1), we realize that $\frac{1}{2}w_1^2 + \lambda - \mu - \varepsilon = \frac{1}{2}w_1^2 = 0$, therefore $w_1 = 0$. Then, in (5), $w_1 + b - 1 = 0 + b - 1 = 0$, therefore $b = 1$. Then, (6) and (7) render a system of simple equations.

$$-(0 - \sqrt{2}w_2 + w_3 + 1) - 1 = 0$$
$$-(0 + \sqrt{2}w_2 + w_3 + 1) - 1 = 0$$

Solving this system of equations renders $w_3 = -1$ and $w_2 = 0$.

(d) Generalizing the solution to 4c. renders that $b = \rho$ from (5). Given ρ_1 and ρ_2 , let us say that 4c. expresses $b, \mathbf{w}$ for some ρ_1 . Then, for some ρ_2 , we find $b = \rho_2, \mathbf{w} = [0, 0, -2\rho_2]$. We realize that our function classifies according to the sign of $\rho(\mathbf{w^T x} + b)$ instead of simply $\mathbf{w^T x} + b$. Knowing that $\rho \geq 1$, we realize that $sign(\mathbf{w^T x} + b) = sign(\rho(\mathbf{w^T x} + b))$ so the classification remains the same for all such ρ .

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

station	iphase accuracy (%)	classifier accuracy (%)
7	97.75	88.08
24	87.23	92.15
3	83.86	92.02
80	95.64	88.10
19	67.56	88.87
38	94.74	90.69
63	91.09	88.18
12	82.33	88.77
74	81.44	89.56
65	81.56	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