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

1. Kernels

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)

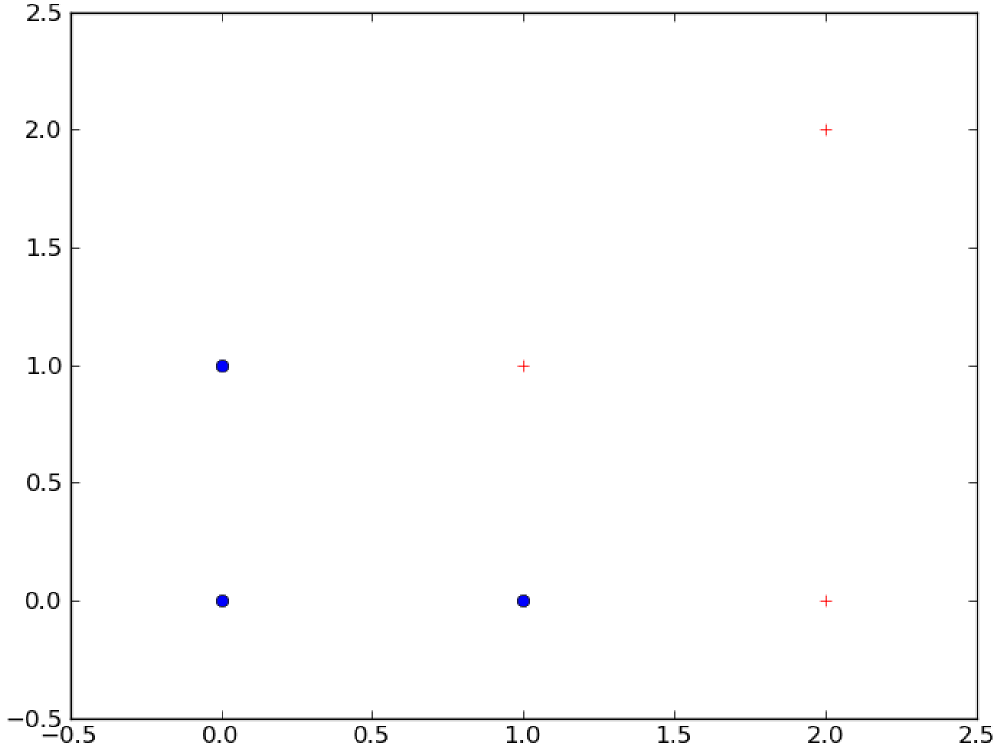
$$\begin{aligned}
\frac{\partial L}{\partial w_j} &= -\sum_{i=1}^{N}(1+e^{y_i(w^T x_i+b)}) \cdot -1 \cdot (1+e^{y_i(w^T x_i+b)})^{-2}(e^{y_i(w^T x_i+b)}) \cdot x_j y_j + \frac{\partial}{\partial w_j}(\lambda \|w\|_2^2) \\
&= -\sum_{i=1}^{N} \frac{-e^{y_i(w^T x_i+b)}}{(1+e^{y_i(w^T x_i+b)})} \cdot x_j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

(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.



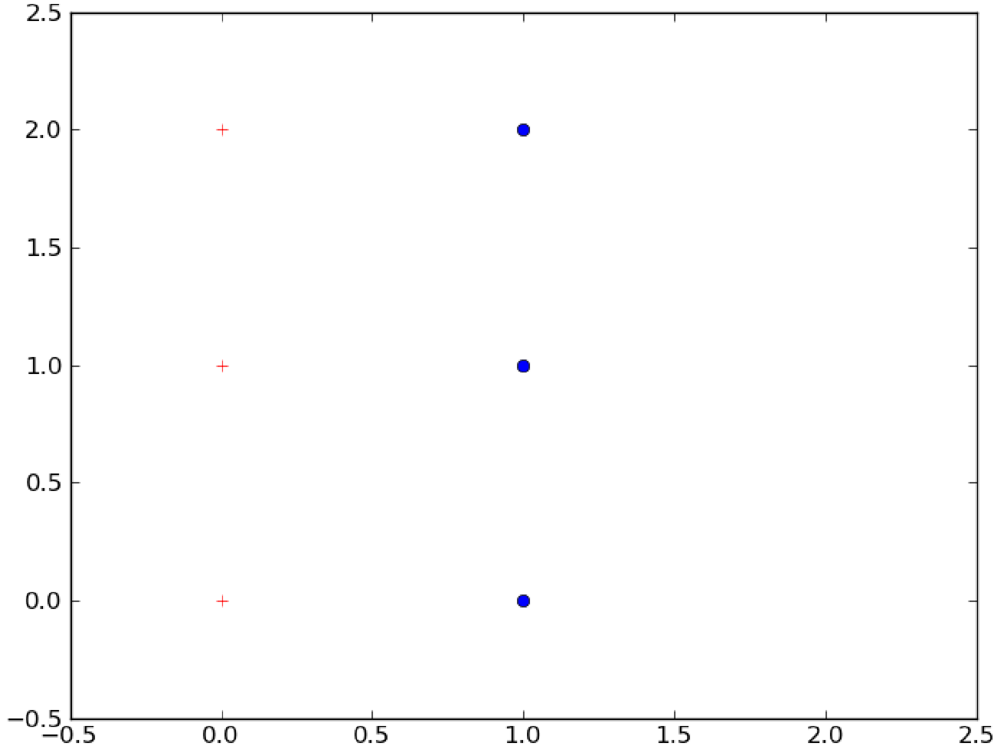
(b) The best hyperplane by inspection is:

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

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
   - Lg,1594,0.0178114489401
   - P,61779,0.690322148101
   - PKP,5974,0.0667538243215
   - Pg,403,0.00450314549741
   - Pn,10762,0.120255215492
   - Rg,11,0.000122914641369
   - S,4685,0.0523504631647
   - Sn,4285,0.0478808398422

   (b) iphase
   - Lg,2171,0.0242588805828
   - N,10683,0.119372464886
   - P,50815,0.567809772831
   - Pg,5291,0.0591219424983
   - Pn,12610,0.140904875242
   - Px,365,0.00407853128178
   - Rg,444,0.00496128188797
   - Sn,318,0.00355335054138
   - Sx,4179,0.0466963896618

- tx,2617,0.0292425105874

(c) TODO

(d) 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