CIS 520, Machine Learning, Fall 2011: Assignment 4

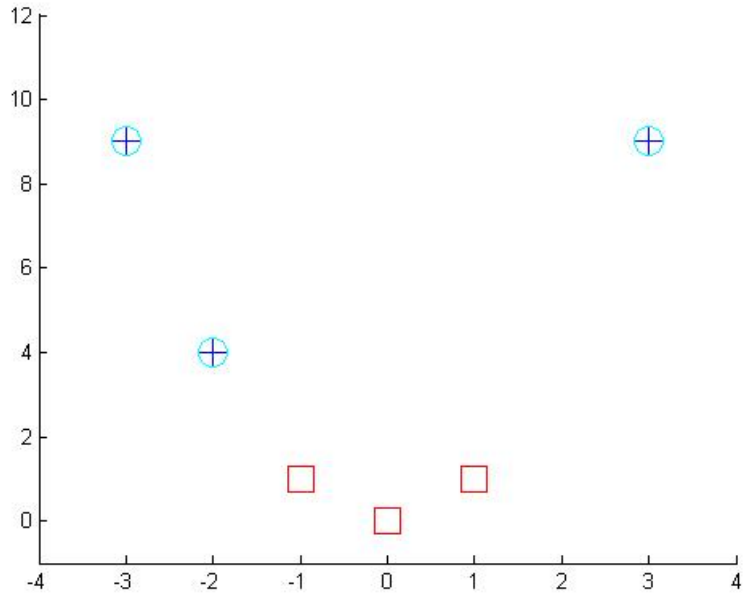
Xiaojun Feng

November 5, 2011

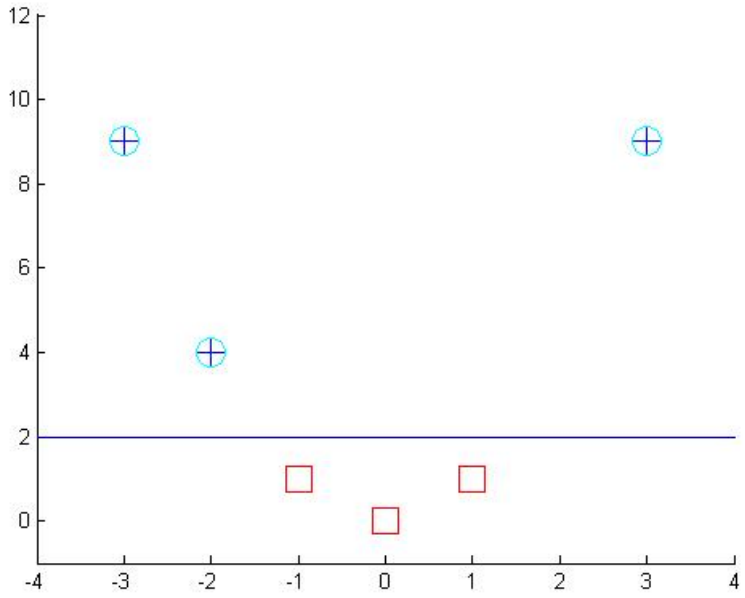
1 SVMs

1. Finite Features

   (a) $\phi(u) = (u, u^2)$



   (b) Line seperator



(c) define $k(X_1, X_1^{'})$

$$\begin{aligned} k(X_1, X_1^{'}) = & \quad \phi(X_1)^T \phi(X_1^{'}) \\ = & \quad (X_1, X_1^2).*(X_1^{'}, X_1^{'2}) \\ = & \quad X_1 X_1^{'} + (X_1 X_1^{'})^2 \end{aligned}$$

(d) By geometric intuition of the graph above, we choose the positive point $Y^{(1)}$ (-2,4) and the negative point $Y^{(2)}$(-1,1) to constitude the suportive vector. Hence we can get the following equations

$$\begin{aligned} w^T Y^{(1)} + c &= 1 \\ w^T Y^{(2)} + c &= -1 \end{aligned}$$

Since the distance to the hyerplane is

$$\begin{aligned} distance &= \frac{w}{2||w||_2}(x_1 - x_2) \\ &= \frac{|w| * |x_1 - x_2|}{2||w||_2} cos(w, (x_1 - x_2)) \end{aligned}$$

we can see that when $cos(w, (x_1 - x_2)) = 1$ the distance is achieve maximum, a.k.a the direction of w is the same as the vector $(x_1 - x_2)$, hence we get the third function:

$$\frac{w_1}{w_2} = \frac{1-2}{-1+4} = -\frac{1}{3}$$

extend Y and w, we get

$$\begin{aligned} -2w_1 + 4w_2 + c &= 1 \\ -w_1 + w_2 + c &= -1 \\ 3w_1 + w_2 &= 0 \end{aligned}$$

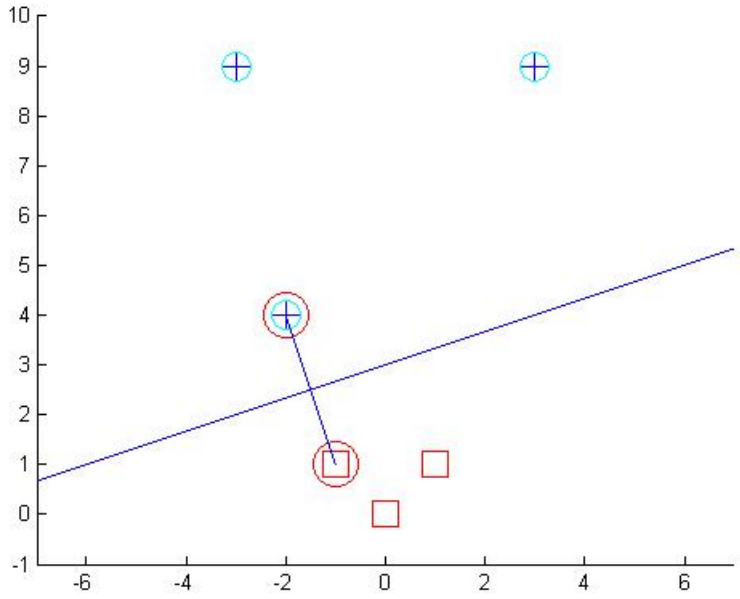
solve the three equations we can get

$$\begin{aligned} w_1 &= -0.2 \\ w_2 &= 0.6 \\ c &= -1.8 \end{aligned}$$

hence

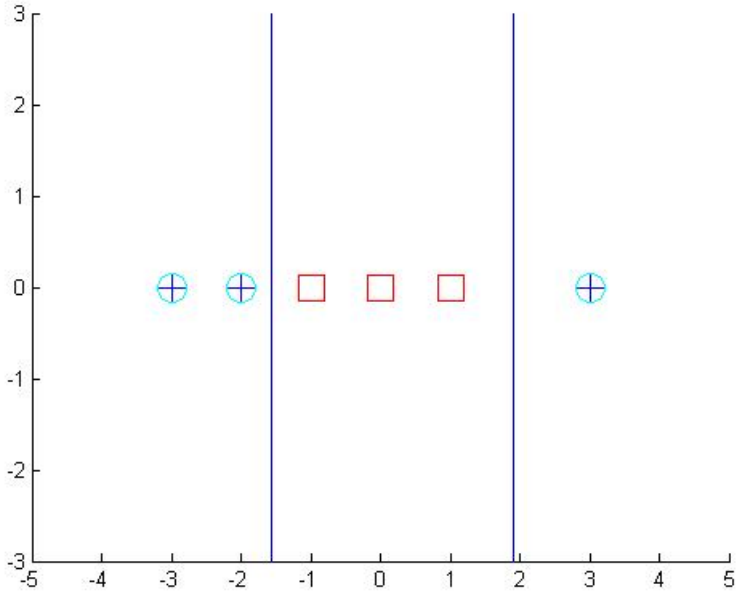
$$\begin{aligned} distance &= \frac{w}{2||w||_2}(x_2 - x_1) \\ &= \frac{2}{2\sqrt{0.4}} \\ &= \frac{\sqrt{10}}{2} \end{aligned}$$

(e) graph



(f) Transfer the hyperlane into R^1feature space

$$\begin{aligned} w^T Y + c &= 0 \\ w_1 x + w_2 x^2 + c &= 0 \\ -0.2x + 0.6x^2 - 1.8 &= 0 \\ 3x^2 - x - 9 &= 0 \\ x &= \frac{1 \pm \sqrt{1 + 4 * 3 * 9}}{6} \\ x &= \frac{1 \pm \sqrt{109}}{6} \end{aligned}$$



(g) The seperable SVM dual is

$$\begin{aligned} max_{\alpha \geq 0} \quad & \sum \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j Y^{(i)T} Y^{(j)} \\ s.t. \quad & \sum \alpha_i y_i = 0 \end{aligned}$$

plugging in with the support vector$(u_1, u_2) = ((-2, +), (-1, -))$

$$\begin{aligned} \sum \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j Y^{(i)T} Y^{(j)} &= \alpha_1 + \alpha_2 - \frac{1}{2}(\alpha_1^2(4+16) - 2\alpha_1\alpha_2(2+4) + \alpha_2^2(1+1)) \\ &= \alpha_1 + \alpha_2 - 10\alpha_1^2 + 6\alpha_1\alpha_2 - \alpha_2^2 \\ \sum \alpha_i y_i = 0 \implies & \alpha_1 - \alpha_2 = 0 \end{aligned}$$

now the SVM dual (plugin with $\alpha_1 = \alpha_2$)is

$$max_{\alpha \geq 0} \qquad 2\alpha_1 - 5\alpha_1^2$$

by derivative, we get $\alpha_1 = \alpha_2 = 0.2$

$$\begin{aligned} w &= \sum_i \alpha_i y_i Y^{(i)} \\ &= 0.2 * (-2, 4) - 0.2 * (-1, 1) \\ &= (-0.2, 0.6) \\ b &= y_1 - w^T Y^{(i)} \\ &= 1 - (0.4 + 2.4) \\ &= -1.8 \end{aligned}$$

(h) Both the hyperlane and margin will not change, since the distance between the point and the hyperplane is too far to make the point selected as the support vector

2. Infinite Features Spaces and Kernel Magic

   (a) If the dataset is consistenct, then there is no finite set of points that cannot be linearly seperated. This is because if we map the hyperlane to the original feature space, it is a bunch of vertical lines the nubmer of which can be infinite, hence, no matter the dataset is we can simply draw the decision boundary between any two adjacent points, so that any point can be sperated.

   (b) Assume the two vetors as as follows

$$\begin{aligned} a &= \phi_\infty(x_1) \\ b &= \phi_\infty(x_2) \\ k(a,b) &= \sum_{i=1}^{\infty} a_i b_i \\ &= e^{-\frac{x_1^2}{2}} e^{-\frac{x_2^2}{2}} + x_1 x_2 e^{-\frac{x_1^2}{2}} e^{-\frac{x_2^2}{2}} + ... + \frac{x_1^i}{\sqrt{i!}} * \frac{x_2^i}{\sqrt{i!}} e^{-\frac{x_1^2}{2}} e^{-\frac{x_2^2}{2}} + ... \\ &= e^{-\frac{x_1^2}{2} - \frac{x_2^2}{2}} (1 + x_1 x_2 + ... + \frac{(x_1 x_2)^i}{i!} + ...) \\ &= e^{-\frac{x_1^2}{2} - \frac{x_2^2}{2}} e^{x_1 x_2} \\ &= e^{-\frac{1}{2}(x_1 - x_2)^2} \end{aligned}$$

   (c) Prove or disprove each claime

      i. Will not change, since

$$k(x_i^{'}, x_j^{'}) = e^{-\frac{1}{2}(x_i + x_0 - x_j - x_0)^2} = e^{-\frac{1}{2}(x_i - x_j)^2} = k(x_i, x_j)$$

      ii. Will not change, since

$$k(x_i^{'}, x_j^{'}) = e^{-\frac{1}{2}(-x_i + x_j)^2} = e^{-\frac{1}{2}(x_i - x_j)^2} = k(x_i, x_j)$$

      iii. It will change, since

$$k(x_i^{'}, x_j^{'}) = e^{-\frac{1}{2}(ax_i - ax_j)^2} = e^{-\frac{1}{2}a^2(x_i - x_j)^2} \neq k(x_i, x_j)$$

2 VC Dimension

1. VC=2k
   We mark the 2k points in an increasing way $x_i < x_i < ...x_{2k}$, and $\delta_i = \frac{x_{i+1} - x_i}{2}$, *and* $0 < \delta < min\delta_i$ and also k incrasing intervals $b_1 < a_2, b_2 < a_{3,...} b_{k-1} < a_k$
   First prove $VC(H) \geq 2k$, we can achieve all 1s by setting a large interval including all points $a_1 = x_i - \delta$, $b_1 = x_{2k} + \delta$, and achieve all 0s by setting all intervals larger than $a_1 = x_{2k} + \delta$, and mark $n \leq k$ points to be positive, by setting n intervals so that $a_i = x_i - \delta, b_i = x_i + \delta$, if n>k, then there are 2k-n negative points, which can seperate the positive points into at most $2k - n + 1 < k + 1 \leq k$ clusters, cluster here means consecutive positive points, for the ith cluster, we set an interval a_i= first point in cluster-δ, b_i=laster point in cluster+δ.
   Second prove $VC(H) \leq 2k$, proove by contradiction, suppose there are 2k+1 points, including k negative points, and k+1 positive points, and the k+1 positivle points are seperated by the k negative points, since the k intervals can include all k+1 positive points, then at least two positive points should be in one interval, meaning that the two points must be adjacent, which contradict our assumption.

2. VC(H)=2
   First prove $VC(H) \geq 2$, assume the two points are $[x_{11}, x_{12}, ..., x_{1n}]$, $[x_{21}, x_{22}, ..., x_{2n}]$,$\delta_i = \frac{|x_{1i} - x_{2i}|}{2}, 0 <$

$\delta < min\delta_i$we can achieve all 1s setting the intervals in every dimention as $a_i = min(x_{1i}, x_{2i}) - \delta$, $b_i = max(x_{1i}, x_{2i}) + \delta$, so that the two points can be covered in all dimension. We can achieve all 0s by setting the intervals simply not including any point. If we achieve single 1, either x_1 *or* x_2,without lose general, suppose $x_1$is labeled 1we can set $a_i = x_{1i} - \delta, b_i = x_{1i} + \delta$, since $|x_{2i} - x_{1i}| > 2\delta > \delta$,so $x_{2i} \notin [a_i, b_i]$.
Second prove $VC(H) \leq 2$, suppose there are 3 points, $h(x_1) = 1, h(x_2) = 0, h(x_3) = 1$,and we arrange the 3 points so that $\exists i, x_{1i} < x_{2i} < x_{3i}$, in this case since $x_{1i} \in [a_i, b_i]$ $x_{2i} \notin [a_i, b_i]$ and $x_{3i} \in [a_i, b_i]$, $[a_i, b_i]$ cannot be closed interval, which generate the contradiction.

3. Labeling

   (a) N points in 1 dimention:
   In one dimention, meaning adjacent points must have the same label except for the two points nearest the seperator. There are the following cases.

      i. All points are labeled the same positive or negative, 2 ways
      ii. all left k points are label the same, the rest are labeled the other, $0 < k < N$, #(k)=N-1, since the left k points can be labeled in two ways, positive or negative, 2*#(k)=2N-2.

   So, the total number is 2N-2+2=2N.

   (b) 4 points in 2 dimention. We arrange the four points in a way so that there are no 3 points in the same line, then 3 of the points construnct a triangle, and the fourth point can be either inside or outside the triangle.

      i. If inside, then if the inside point is labeled different from all other 3 points, it is not seperable, since the inside point can be labeled either positive of negative, so the number is $2^4 - 2 = 14$.
      ii. If outside, then if we label the points at the same diagonal line with same color, and different with the others, the points are not seperable, also it can be labeled in 2 ways, so , the number is also 14

   (c) Like the 4 points in 2 dimention, we arrange the 5 points so that no 4 points are at the same plane, then with any 4 points we can construct tetrahedron, and the 5th point is either inside or outside the tetrahedron, similar to the problem above, for each situation there are 2 ways making the points unseperable, so the number is $2^5 - 2 = 30$.

3 Online Learning

1. Since the algorithm is deterministic, so the adversary knows what the the prediction result of any point. If VC(H)=m, it means there is some set of m points that can be shattered by H. So, the adversary can arrange these points as the first m points in the sequence, since the adversary knows the prediction result, he can always label these points different from the majority, so that m mistakes would be forced to make. And because VC(H)=m, so no matter how the m points are labeled, there must be some $f \in H$, for every point x_i of the m points, $f(x_i) = y_i$.

2. In the case that there are k categories, every time we make mistake, we can delete all the hypothesis that predict wrong, since majority is wrong, means that at most $\frac{C_t}{2} - 1$predict correct and can be remained, and in this worst case, the wrong hypothesises generate at most 2 wrong labels, one is produced by$\frac{C_t}{2}$ hypothesises which is the prediction label, and the other is produced by a single one, otherwise the right label would be the majority, or the number of the remaining hypothesis would be less than $\frac{C_t}{2} - 1$. So $C_{t+1} < \frac{C_t}{2}$, so the upper bound is still $log_2|H|$.