CS446: Machine Learning **Spring 2017**

Problem Set 4

Bangqi Wang *Handed In: April 11, 2017*

1. Answer to problem 1

   a. The algorithm needs to find the edges between negative data and positive data.

   - All Positive: find the smallest and the largest points.
   - All Negative: find two points that has largest distance between them.
   - Negative & Positive: find the smallest and the largest positive points.

Algorithm 1 Pseudocode:

procedure MyProcedure
  $queue \leftarrow PriorityQueue()$
  $hasPositive \leftarrow False$
  $left,\ right \leftarrow Null, Null$
  for `each point` P `in training set` **do**
    $queue$.push$((distance, P))$
    if P `is` $'+'$ **then**
      $hasPositive \leftarrow True$
    end if
  end for
  if $hasPositive$ `is` $True$ **then**
    $left \leftarrow$ smallest distance d_{left} among positive points
    $right \leftarrow$ largest distance d_{right} among positive points
  end if
  if $hasPositive$ `is` $False$ **then**
    $left \leftarrow$ smaller distance d_{left} of two negative points with largest distance between.
    $right \leftarrow$ larger distance d_{right} of two negative points with largest distance between.
  end if
  $r_1, r_2 \leftarrow left, right$
  return r_1, r_2
end procedure

   b. i. The hypothesis will only make mistakes on positive examples because it is tightest boundary for positive. Negative may never beyond the boundary but there may exist new positive that widen the boundary and make the boundary closer to the true boudnary.

   ii.

$$Pr_{each} = 1 - \epsilon$$

$$Pr_m = (1 - \epsilon)^m$$

c.

$$Pr_m = (1-\epsilon)^m < \delta$$
$$Pr_m = (e^{-\epsilon})^m < \delta$$
$$Pr_m = e^{-\epsilon m} < \delta$$
$$-\epsilon m > log(\delta)$$
$$\epsilon m > log(\frac{1}{\delta})$$
$$m > \frac{1}{\epsilon} * log(\frac{1}{\delta})$$

d. The VC Dimension is 1. There is an complexity bound with finite VC dimension. $m < \frac{1}{\epsilon}(8log\frac{13}{\epsilon} + 4log\frac{2}{\delta})$. The bound we found is tighter than the this bound using VC dimension.

2. Answer to problem 2

   a. The VC Dimension is 3. The function $ax^2 + bx + c$ is a parabola. The parabola can only split the axis into at most 3 parts. In this situation, there is no way to shatter data with more than 3 parts, such as $'+, -, +, -'$ or $'-, +, -, +'$.

3. Answer to problem 3

   a. From lecture note

      - weight vector $w = \sum_{1,m} r\alpha_i y_i x^i$
      - α_i is the mistake made on x^i
      - $f(x) = sgn(\sum_{z \in M} S(z)K(x,z))$
      - where $K(x,z) = \sum\limits_{i \in I} t_i(z)t_i(x)$, and $S(z) = \pm 1$

   b.

   $$K(\vec{\mathbf{x}}, \vec{\mathbf{z}}) = (\vec{\mathbf{x}}^T\vec{\mathbf{z}})^3 + 49(\vec{\mathbf{x}}^T\vec{\mathbf{z}} + 4)^2 + 64\vec{\mathbf{x}}^T\vec{\mathbf{z}}.$$
   $$K(\vec{\mathbf{x}}, \vec{\mathbf{z}}) = (\vec{\mathbf{x}}^T\vec{\mathbf{z}})^3 + 49(\vec{\mathbf{x}}^T\vec{\mathbf{z}})^2 + 456\vec{\mathbf{x}}^T\vec{\mathbf{z}} + 784$$

   Since $\vec{\mathbf{x}}^T\vec{\mathbf{z}}$ is a valid kernel. So, the polynomials of $\vec{\mathbf{x}}^T\vec{\mathbf{z}}$ is still a valid kernel.

   c. $K(\mathbf{x}, \mathbf{z}) = \sum_{c \in C} c(\mathbf{x})c(\mathbf{z})$, where C is combination C_m^k. It will take linear time $O(n)$ to find k features that are positive in both x and z. Traversing through x and z needs linear time $O(n)$.

4. Answer to problem 4

   a. 1. $w = < -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} >$, $\theta = 0$
      2. $w = < -\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2} >$, $\theta = 0$
      3. SVM optimization will make the margin between positive and negative as large as possible. Usually it will draw two parallel lines that will pass the closest positive point and the closest negative point. Then draw the middle line of those two lines. This line is perpendicular bisector of those two points.

b. 1. $< -2, 0 >$, $< 0, 2 >$
   2. $\alpha = \{\frac{\sqrt{2}}{4}, -\frac{\sqrt{2}}{4}\}$
   3. $Objective function value = \frac{1}{2}||\mathbf{w}||^2 = \frac{1}{2} \times \sqrt{(\frac{\sqrt{2}}{2})^2 + (\frac{\sqrt{2}}{2})^2} = \frac{1}{2}$

c. Slack variable ξ allows some instances to fall off the margin, but penalize them. C trades-off margin width and misclassification.
   - $C \to 0$: the optimization problem gives the hyperplan found in $(a) - 2$.
   - $C \to 1$: get the hyperplan that balance the margin and hinge loss.
   - $C \to \infty$: get the hyperplan closer to hard-margin solution.

5. Answer to problem 5

i	Label	Hypothesis 1				Hypothesis 2			
		D_0	$f_1 \equiv [x > 2]$	$f_2 \equiv [y > 4]$	$h_1 \equiv [x > 2]$	D_1	$f_1 \equiv [x > 9]$	$f_2 \equiv [y > 11]$	$h_2 \equiv [y > 11]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	$-$	$\frac{1}{10}$	$-$	$+$	$-$	$\frac{1}{16}$	$-$	$-$	$-$
2	$-$	$\frac{1}{10}$	$-$	$-$	$-$	$\frac{1}{16}$	$-$	$-$	$-$
3	$+$	$\frac{1}{10}$	$+$	$+$	$+$	$\frac{1}{16}$	$-$	$-$	$-$
4	$-$	$\frac{1}{10}$	$-$	$-$	$-$	$\frac{1}{16}$	$-$	$-$	$-$
5	$-$	$\frac{1}{10}$	$-$	$+$	$-$	$\frac{1}{16}$	$-$	$+$	$+$
6	$-$	$\frac{1}{10}$	$+$	$+$	$+$	$\frac{1}{4}$	$-$	$-$	$-$
7	$+$	$\frac{1}{10}$	$+$	$+$	$+$	$\frac{1}{16}$	$+$	$-$	$-$
8	$-$	$\frac{1}{10}$	$-$	$-$	$-$	$\frac{1}{16}$	$-$	$-$	$-$
9	$+$	$\frac{1}{10}$	$-$	$+$	$-$	$\frac{1}{4}$	$-$	$+$	$+$
10	$+$	$\frac{1}{10}$	$+$	$+$	$+$	$\frac{1}{16}$	$-$	$-$	$-$

a. All elements have same weights $\frac{1}{10}$.

b. The best hypothesis is $x > 2$. $\epsilon_0 = 0.2$, $\alpha_0 = \frac{1}{2} ln(\frac{1-\epsilon_0}{\epsilon_0}) = 0.69$

c. For weights of D_1:
   - $D_{t+1} = D_t / (z_t \times e^{-\alpha_t y_i h_t(x_i)})$, where $z_0 = \sum_i D_0(i) e^{-\alpha_0 y_i h_0(x_i)} = 0.8$
   - $D_{1_{correct}} = D_0 e^{\alpha_0} / z_0 = \frac{1}{16}$
   - $D_{1_{error}} = D_0 e^{-\alpha_0} / z_0 = \frac{1}{4}$

d. Final hypothesis:
   - $\epsilon_1 = 0.25$, $\alpha_1 = \frac{1}{2} ln(\frac{1-\epsilon_1}{\epsilon_1}) = 0.55$
   - $H_{final}(x) = sgn(\sum_t \alpha_t h_t(x)) = sgn(0.69 \cdot [x > 2] + 0.55 \cdot [y > 11])$