CS57800 Statistical Machine Learning

Homework 3 Solutions

1 Questions(50pts)

1. We see that C can shatter $2d+1$ points on a circle: proceeding clockwise around the circle, we pass a line from last of each run of consecutive included points to start of the next such run, for a total of at most d lines, and define our polygon by these lines (which extend beyond their defining points).
   Suppose we have $2d+2$ points. If any point is in the convex hull of the others, then C fails to shatter: any set in C is convex, and thus can not include the extreme points while excluding an interior point. Otherwise, number the points in their clockwise order as vertices of the convex hull, and consider an alternating sign pattern. In order to shatter, there must be a face of the polygon from C passing between each excluded point and its two included neighbors; these $d+1$ faces must all be distinct, or else we could show that the given points violated our convexity assumption; but no convex polygon with d vertices has more than d faces, a contradiction. So C can not shatter $2d+2$ points.

2. Let's take the first order and second order derivation, then we will have:
$$\frac{d}{d\hat{y}} = \frac{1}{log2}\frac{1}{1+e^{-y\hat{y}}}(-ye^{-y\hat{y}})$$
$$\frac{d^2}{d\hat{y}^2} = \frac{1}{log2}\frac{y^2e^{-y\hat{y}}}{(1+e^{-y\hat{y}})^2}$$
   We can see that the second derivation is always positive for any value of $y \in \{+1, -1\}$.

3. The training error of the final classifier, H, is bounded by:
   $\frac{1}{n}\sum_{i=1}^{n}\delta(H(x_i) \neq y_i) \leq exp(-2\sum_{t=1}^{T}(0.5-\epsilon_t)^2)$
   The training error should be bounded by $\frac{1}{n}$, and thus we choose T such that
   $\frac{1}{n} \geq exp(-2\sum_{t=1}^{T}(0.5-\epsilon_t)^2)$
   $\frac{1}{n} \geq exp(-2T(0.5-\gamma_t)^2)$
   $\Rightarrow ln(n) \leq 2T(0.5-\gamma_t)^2$
   $\Rightarrow \frac{ln(n)}{2(0.5-\gamma_t)^2} \leq T$

4. We can start with the probabilities $p_1(x_i) = 0.1$ for all the examples(here i is an index of examples). For the first round, the best threshold A is 5. Then,

$$h_1(x_i) = I(x_i > 5)$$
$$\epsilon_1 = 0.2$$
$$\alpha_1 = \frac{1}{2} ln(\frac{1-\epsilon_1}{\epsilon_1}) = 0.6931$$

Now we need to update the probabilities for each example as follows:

$$p_{t+1}(i) = \frac{p_t(i)}{Z_t} e^{-\alpha_t y_i h_t(x_i)}$$

where Z_t is the constant normalizing factor.

index	1	2	3	4	5	6	7	8	9	10
correct	Y	Y	Y	Y	Y	Y	Y	Y	N	N
$p_1(x_i)$	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
$p_2(x_i)$	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.2500	0.2500

For the second round, we again choose the best threshold that makes the error minimized with the new updated probabilities $p_2(x_i)$. The best threshold is 3. Then,

$$h_2(x_i) = I(x_i > 3)$$
$$\epsilon_2 = 0.375$$
$$\alpha_2 = \frac{1}{2} ln(\frac{1-\epsilon_2}{\epsilon_2}) = 0.2554$$

The updated probabilities are

index	1	2	3	4	5	6	7	8	9	10
correct	Y	N	Y	N	Y	Y	Y	Y	Y	N
$p_2(x_i)$	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625	0.2500	0.2500
$p_3(x_i)$	0.0500	0.0833	0.0500	0.0833	0.0500	0.0500	0.0500	0.0500	0.2000	0.3333

The final hypothesis after two rounds is

$$\begin{aligned} h_{final}(x_i) &= 0.6931 h_1(x_i) + 0.2554 h_2(x_i) \\ &= sign(0.6931 I(x_i > 5) + 0.2554 I(x_i > 3)) \end{aligned}$$

5. If $K_1(\vec{x}, \vec{y}) = \phi_1(\vec{x})^T \phi_1(\vec{y})$ and $K_2(\vec{x}, \vec{y}) = \phi_2(\vec{x})^T \phi_2(\vec{y})$, we could define a projection $\phi_3(\vec{x})^T = [\sqrt{\alpha}\phi_1(\vec{x}) \ \sqrt{\beta}\phi_2(\vec{x})]$. It would then follow that $K(\vec{x}, \vec{y}) = \alpha K_1(\vec{x}, \vec{y}) + \beta K_2(\vec{x}, \vec{y})$ is also a kernel function.

6. $E_i = \frac{1}{2}|y_i - sign(x_i w + b)| \neq 0$ iff $\xi_i > 1$

$$\Rightarrow E = \sum_{i=1}^{M} E_i = \sum_{i=1, \xi_i > 1}^{M} E_i \leq \sum_{i=1, \xi_i > 1}^{M} \xi_i$$

$$\xi_i \geq 0, \forall i = 1, \ldots, M \Rightarrow E \leq \sum_{i=1, \xi_i > 1}^{M} \xi_i \leq \sum_{i=1}^{M} \xi_i$$