EE6550 MACHINE LEARNING

HW#1, WORD PROBLEMS

102061210 王尊玄

Problem 2.7

(a) $\left(I, F, \tilde{P}\right)$: $a\ probability\ space$

c : $an\ unknown\ but\ fixed\ concept$

$S = (x_1, x_2, x_3, \ldots x_m)$: $a\ sample\ of\ m\ elements\ drawn\ i.i.d.\ from$ $I\ according\ to\ probability\ distribution\ \tilde{P}.$

$c(x_i), i = 1 \sim m$: $labels\ of\ x_i \in S$

$c'(x_i), i = 1 \sim m$: $labels\ of\ x_i \in S\ interfered\ with\ noise$ η

$d(h^*)$

$(probability\ that\ the\ label\ of\ a\ training\ point\ received\ c'(x_i)$
$disagree\ with\ that\ given\ by\ h^*)$

$= P_m(h^*(x_i) \neq c'(x_i))$

$(x_i\ is\ drawn\ i.i.d. and\ \eta\ is\ independent\ of\ \tilde{P}\ and\ is\ identically$
$added\ to\ every\ drawn)$

$= P(h^*(x_i) \neq c'(x_i))$

$= P(h^*(x_i) = c(x_i) \cap c(x_i) \neq c'(x_i))$

$(\eta\ is\ an\ independent\ additive\ noise)$

$= P(h^*(x_i) = c(x_i))P(c(x_i) \neq c'(x_i))$

$(h^*\ is\ the\ target\ concept\ which\ gives\ same\ label\ as\ c\ does$
$corresponding\ to\ sample\ S)$

$= 1 \times \eta$

$= \eta$

(b) $d(h)$

$= P(h(x_i) \neq c'(x_i))$

$= P(h(x_i) \neq h^*(x_i) \cap h^*(x_i) = c'(x_i)$
$\cup\ h(x_i) = h^*(x_i) \cap h(x_i) \neq c'(x_i))$

$(2\ events\ are\ mutually\ exclusive)$

$= P(h(x_i) \neq h^*(x_i) \cap h^*(x_i) = c'(x_i)) +$
$P(h(x_i) = h^*(x_i) \cap h(x_i) \neq c'(x_i))$

$(\eta\ is\ an\ independent\ additive\ noise)$

$= P(h(x_i) \neq h^*(x_i))P(h^*(x_i) = c'(x_i)) +$

$$P(h(x_i) = h^*(x_i))P(h(x_i) \neq c'(x_i))$$
$$= R(h) \times (1-\eta) + (1-R(h)) \times \eta$$
$$= \eta + (1-2\eta)R(h)$$

(c) $R(h) > \varepsilon\ implies\ that\ d(h) > \eta + (1-2\eta)\varepsilon$

$from\ (a), d(h^*) = \eta, we\ get$

$d(h) - d(h^*) > \eta + (1-2\eta)\varepsilon - \eta = (1-2\eta)\varepsilon$

$d(h) - d(h^*) > \varepsilon', \varepsilon' = (1-2\eta)\varepsilon$

(d) $According\ to\ definition\ of\ \hat{d}(h), we\ know\ that\ E(\hat{d}(h)) = d(h)$

$$P(\hat{d}(h^*) - d(h^*) > \frac{\varepsilon'}{2})$$
$$= P(\hat{d}(h^*) - E\left(\hat{d}(h^*)\right) > \frac{\varepsilon'}{2})$$
$(by\ Hoeffding's\ inequality)$
$$\leq \mathrm{e}^{-\frac{m\varepsilon'^2}{2}}, set\ this\ term\ to\ \frac{\delta}{2}, we\ get$$
$$P(\hat{d}(h^*) - d(h^*) > \frac{\varepsilon'}{2}) \leq \frac{\delta}{2}, which\ implies$$
$$P(\hat{d}(h^*) - d(h^*) \leq \frac{\varepsilon'}{2}) > 1 - \frac{\delta}{2}, for\ m \geq \frac{2}{\varepsilon'^2}\ln\frac{\delta}{2}$$

(e) $we\ now\ want\ to\ find\ for\ all\ h \in H, (d)\ is\ true, i.e.$

$$P(\exists h \in H, \hat{d}(h) - d(h) > \frac{\varepsilon'}{2})$$
$(by\ union\ bound)$
$$\leq |H|P(\hat{d}(h_i) - d(h_i) > \frac{\varepsilon'}{2})$$
$(by\ Hoeffding's\ inequality)$
$$\leq |H|\mathrm{e}^{-\frac{m\varepsilon'^2}{2}}, set\ |H|\mathrm{e}^{-\frac{m\varepsilon'^2}{2}}\ to\ \frac{\delta}{2}, we\ get$$
$$P(\hat{d}(h) - d(h) > \frac{\varepsilon'}{2}) \leq \frac{\delta}{2}, which\ implies$$
$$P(\hat{d}(h) - d(h) \leq \frac{\varepsilon'}{2}) > 1 - \frac{\delta}{2}, for\ m \geq \frac{2}{\varepsilon'^2}\left(\ln|H| + \ln\frac{\delta}{2}\right)$$

(f) $(c)\ states\ that\ if\ R(h) > \varepsilon, d(h) - d(h^*) \geq \varepsilon'$

$$(d)\ states\ that\ \forall\delta > 0, P(\hat{d}(h^*) - d(h^*) \leq \frac{\varepsilon'}{2}) > 1 - \frac{\delta}{2},$$
$$for\ m \geq \frac{2}{\varepsilon'^2}\ln\frac{\delta}{2}$$

$$(e)\ states\ that\ \forall\delta > 0, P(\hat{d}(h) - d(h) \leq \frac{\varepsilon'}{2}) > 1 - \frac{\delta}{2}$$

$$for\ m \geq \frac{2}{\varepsilon'^2}\left(\ln|H| + \ln\frac{\delta}{2}\right)$$

$$P(\hat{d}(h) - \hat{d}(h^*) \geq 0)$$

$$= P(\left[\hat{d}(h) - d(h)\right] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)] \geq 0)$$

$$(the\ above\ can\ be\ intrpretted\ by\ (c), (d), (e))$$

$$= P((c) \cap (d) \cap (e) \geq -\frac{\varepsilon'}{2} + \varepsilon' - \frac{\varepsilon'}{2} = 0)$$

$$> 1 - 0 - \frac{\delta}{2} - \frac{\delta}{2}$$

$$= 1 - \delta$$

$$, where\ \left(m \geq \frac{2}{\varepsilon'^2}\ln\frac{\delta}{2}\right) \cap \left(m \geq \frac{2}{\varepsilon'^2}\left(\ln|H| + \ln\frac{\delta}{2}\right)\right)$$

$$= m \geq \frac{2}{\varepsilon'^2}\left(\ln|H| + \ln\frac{\delta}{2}\right)$$

$$So\ we\ get\ P\left(\hat{d}(h) - \hat{d}(h^*) \geq 0\right), for\ m \geq \frac{2}{\varepsilon'^2}\left(\ln|H| + \ln\frac{\delta}{2}\right)$$

Problem 2.9

$Given\ a\ PAC\ learning\ algorithm\ A.\ we\ can\ write,$

$P_{S\sim D^m}(R(h_s) \leq \varepsilon) \geq 1 - \delta$

$To\ approximate\ an\ unknown\ but\ fixed\ concept\ c,$

$assume\ a\ sample\ of\ m\ elements\ S = (x_1, x_2 \ldots x_m)\ is\ drawn\ i.i.d.$

$according\ to\ uniform\ distribution\ D, we\ can\ calculate\ generalization$

$error\ of\ hypothisis\ h_s\ over\ sample\ S, giving,$

$$R(h_s)$$

$$= E_{S\sim D^m}\left(\hat{R}_s(h_s)\right)$$

$$= E_{S\sim D^m}\left(\frac{1}{m}\sum_{i=1}^{m} 1_{h_s(x)\neq c(x_i)}\right)$$

$$(sample\ is\ drawn\ i.i.d.)$$

$$= E_{S\sim D}\left(\frac{1}{m}\sum_{i=1}^{m} 1_{h_s(x)\neq c(x_i)}\right), has\ its\ minimum\ step\ (one\ error)\ as\ \frac{1}{m},$$

$i.e.\ if\ only\ one\ error\ exists, cost\ R(h_s)\ will\ be \frac{1}{m}, giving$

$R(h_s) = \frac{k}{m}, k \in 0\ or\ N. This\ implies\ that$

$if\ R(h_s) < \frac{1}{m}, then\ R(h_s) = 0 \rightarrow consistent$

$We\ set\ \varepsilon = \frac{1}{m+1}, giving\ P_{S\sim D^m}\left(R(h_s) \leq \frac{1}{m+1}\right) \geq 1 - \delta$

$from\ the\ above\ we\ can\ further\ get\ P_{S\sim D^m}(R(h_s) = 0) \geq 1 - \delta$

$Thus, we\ can\ use\ algorithm\ A\ and\ sample\ S\ to\ find\ in\ polynomial$

$time\ P\left(m+1, \frac{1}{\delta}\right) a\ hypothisis\ h_s\ consistent\ with\ \left(x_i, c(x_i)\right), with$

$high\ probability\ 1 - \delta.$

Problem 3.9

$the\ set\ of\ all\ closed\ ball\ in\ \boldsymbol{R}^n\ \{x \in \boldsymbol{R}^n: \|x - a\|^2 \leq r\}\ can$

$be\ rewritten\ as\ \sum_{i=1}^n (x_i - a_i)^2 - \sum_{i=1}^n r_i \leq 0.$

$$\sum_{i=1}^n (x_i - a_i)^2 - \sum_{i=1}^n r_i$$

$$= \sum_{i=1}^n x_i^2 + \sum_{i=1}^n 2a_i x_i + \sum_{i=1}^n (a_i^2 - r_i) \leq 0$$

$the\ above\ inequality\ refers\ to\ a\ halfspace\ in\ \boldsymbol{R}^{n+1}$

$\boldsymbol{W^T} \cdot \boldsymbol{x} + \boldsymbol{b} \leq 0, where$

$$\boldsymbol{W} = \begin{bmatrix} 1 \\ 2a_1 \\ 2a_2 \\ \vdots \\ 2a_n \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} \sum_{i=1}^n x_i^2 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, b = \sum_{i=1}^n (a_i^2 - r_i)$$

$Thus, VC\ dimension\ of\ all\ closed\ balls\ in\ \boldsymbol{R}^n, which\ is\ equivalent\ to\ that$
$of\ halfspaces\ or\ hyperplanes\ in\ \boldsymbol{R}^{n+1}, is\ at\ most\ n+2, as\ known$
$that\ VC\ dimension\ of\ halfspaces\ in\ \boldsymbol{R}^n\ is\ n+1. Besides, if\ we\ try$
$to\ reduce\ the\ dimension\ of\ hypersphere, i.e. assign\ several\ x_i = 0$
$VC\ dimension\ becomes\ less\ than\ n+2.$

Problem 3.12

(a) $Consider\ a\ set\ of\ point\ \{x, 2x, 3x, 4x\}, assume\ that\ dichotomy$
$\{-,-,+,-\}\ can\ be\ realized. Then\ from\ triangometric\ function,$
$sin(4wx) = 2\sin(2wx)\cos(2wx) < 0$
$as\ sin(2wx) < 0, we\ can\ know\ \cos(2wx) > 0$

$$\cos(2wx) = 1 - 2\sin^2 wx > 0 \rightarrow sin^2(wx) < \frac{1}{2}$$

$$as\ sin(wx) < 0, we\ can\ know\ sin(wx) > -\frac{1}{\sqrt{2}}$$

$Besides, sin(3wx) = 3sin(wx) - 4\, sin^3(wx) > 0$

$$as\ sin(wx) < 0, then\ we\ get\ 3 - 4\, sin^2 wx < 0 \rightarrow sin^2 wx > \frac{3}{4}$$

$$as\ sin(wx) < 0, we\ can\ know\ sin(wx) < -\frac{\sqrt{3}}{2}, which$$

$$contradicts\ to\ above\ statement\ sin(wx) > -\frac{1}{\sqrt{2}}$$

$Thus, dichotomy\ \{-,-,+,-\}\ cannot\ be\ realized\ and\ the$
$set\ of\ point\ \{x, 2x, 3x, 4x\}\ cannot\ be\ shattered\ by\ the$
$family\ of\ sine\ function.$

(b) $Given\ a\ set\ of\ points\ with\ m\ elements\ in\ \boldsymbol{R}, \{2^{-i}: i \le m\},$

$family\ of\ sine\ functions\ can\ shatter\ the\ set. Also, if\ for\ all$
$m > 0, the\ aforementioned\ statement\ is\ true,\ then\ VC$
$dimension\ of\ sine\ function\ is\ infinite.$

$Our\ goal: sign(\sin w2^{-j})\ can\ be\ \pm 1, \forall j > 0$

$Let\ w = \pi\left(1 + \sum_{i=1}^{m} 2^i \frac{1-y_i}{2}\right), y_i\ is\ the\ label\ of\ i^{th}\ element.$

$$sin(w2^{-j})$$

$$= sin(\pi\left(1 + \sum_{i=1}^{m} 2^i \frac{1-y_i}{2}\right) 2^{-j})$$

$$= sin\left(\pi\left(2^{-j} + \sum_{i=1}^{m} 2^{i-j} \frac{1-y_i}{2}\right)\right)$$

$(Since\ \frac{1-y_i}{2}\ can\ only\ be\ 0\ or\ 1,\ \sum_{i=j+1}^{m} 2^{i-j} \frac{1-y_i}{2}\ only$

$contribute\ 2\pi\ to\ sine\ function, which\ doesn't\ matter.)$

$$= sin(\pi(2^{-j} + \sum_{i=1}^{j-1} 2^{i-j} \frac{1-y_i}{2} + \frac{1-y_j}{2}))$$

$(upper\ bound\ occurs\ as\ \frac{1-y_i}{2}\ is\ always\ 1)$

$$\le sin(\pi(2^{-j} + \sum_{i=1}^{j-1} 2^{i-j} + \tfrac{1-y_j}{2}))$$

$$= sin(\pi(\sum_{i=0}^{j-1} 2^{i-j} + \tfrac{1-y_j}{2}))$$

$$(\sum_{i=0}^{j-1} 2^{i-j} = \tfrac{2^{-j}(1-2^j)}{1-2} = 1 - 2^{-j} < 1)$$

$$< sin(\pi(1 + \frac{1-y_j}{2}))$$

Besides, we can know the lower bound,

$$sin(\pi(2^{-j} + \sum_{i=1}^{j-1} 2^{i-j} \tfrac{1-y_i}{2} + \tfrac{1-y_j}{2})) > \ sin(\pi \tfrac{1-y_j}{2})$$

Thus we get, $\forall j > 0$

$$sin(\pi \tfrac{1-y_j}{2}) < sin(w2^{-j}) < sin(\pi(1 + \tfrac{1-y_j}{2}))$$

From the above inequality, if $y_j = +1$, *then* $sgn(\sin(wx_j)) = +1$. *If* $y_j = -1$, *then* $sgn(\sin(wx_j)) = -1$. *Hence our goal is a true statement, saying VC dimension of hypothesis family of sine functions is infinite.*

Problem 3.21

(a) $P(\sup_{h\in H_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\varepsilon}{2})$

$(h \in H_S$ *is consistent over sample* S, *which makes* $\hat{R}_S(h) = 0$.
However, $\hat{R}_{S'}$ *is the empirical error over sample* S', *which may not necessary make* $\hat{R}_{S'}(h) = 0.)$

$$= P(\sup_{h\in H_S} \hat{R}_{S'}(h) > \frac{\varepsilon}{2})$$

$(for\ all\ h \in H_S\ makes\ a\ stricter\ set\ than\ \sup_{h\in H_S})$

$$\ge P(\forall h \in H_S\colon \hat{R}_{S'}(h) > \frac{\varepsilon}{2})$$

$$= P(\hat{R}_{S'}(h_0) > \frac{\varepsilon}{2})$$

$$\ge P(\hat{R}_{S'}(h_0) > \frac{\varepsilon}{2} \cap R(h_0) > \varepsilon)$$

$$\ge P\left(\hat{R}_{S'}(h_0) > \tfrac{\varepsilon}{2}\right) P(R(h_0) > \varepsilon)$$

$(since\ \hat{R}_{S'}(h_0) = \frac{1}{m}\sum_{i=1}^{m} 1_{h_0(x_i)\neq c(x_i)})$

$= P(\sum_{i=1}^{m} 1_{h_0(x_i)\neq c(x_i)} > \frac{m\varepsilon}{2})P(R(h_0) > \varepsilon)$

$(B(m,\varepsilon)\ means\ w.r.t.\ a\ sample\ of\ m\ elements\ and\ probability\ of\ success\ (in\ our\ case\ incorrect\ prediction) = \varepsilon,$ the $number\ of\ success\ (incorrect\ prediction).\ Besides,\ R(h_0) > \varepsilon\ implies\ the\ probability\ of\ incorrect\ prediction\ using\ h_0\ over\ a\ new\ testing\ input\ is\ at\ least\ \varepsilon.$

$\geq P(B(m,\varepsilon) > \frac{m\varepsilon}{2})P(R(h_0) > \varepsilon)$

(b) $P(B(m,\varepsilon) \leq \frac{m\varepsilon}{2})$

$= P(B(m,\varepsilon) > (1 - \frac{1}{2})m\varepsilon)$

$(by\ Chernoff\ bound)$

$\leq e^{-\frac{\left(\frac{1}{2}\right)^2}{2}m\varepsilon} = e^{-\frac{1}{8}m\varepsilon}$

$(since\ m\varepsilon \geq 8\ given\ by\ this\ problem)$

$\leq \mathrm{e}^{-1} \leq \frac{1}{2}$

$Thus, we\ get\ P\left(B(m,\varepsilon) \leq \frac{m\varepsilon}{2}\right) \leq \frac{1}{2}.\ This\ implies$

$P\left(B(m,\varepsilon) > \frac{m\varepsilon}{2}\right) \geq \frac{1}{2}, and\ from\ (a) we\ then\ get$

$P(\sup_{h\in H_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\varepsilon}{2}) \geq \frac{1}{2}P(R(h_0) > \varepsilon)$

$P(R(h_0) > \varepsilon) \leq 2P(\sup_{h\in H_S}\left|\hat{R}_S(h) - \hat{R}_{S'}(h)\right| > \frac{\varepsilon}{2})$

(c) $Now\ we\ uniformly\ at\ random\ split\ 2m\ points\ of\ T\ into\ S\ and\ S', implying\ every\ point\ in\ T\ have \frac{1}{2}\ chance\ to\ be\ in\ S\ or\ S'.\ l\ is\ the\ total\ number\ of\ errors\ h_0\ makes\ on\ sample\ T, and\ therefore\ probability\ of\ every\ error\ falling\ into\ S'\ is \frac{1}{2}.\ Let\ E\ be\ the\ set\ of\ all\ errors\ on\ T,$

$$\underset{T\sim D^{2m},T\to[S,S']}{P}(\hat{R}_S(h_0)=0\cap\hat{R}_{S'}(h_0)>\frac{\varepsilon}{2}|\hat{R}_T(h_0)>\frac{\varepsilon}{2})$$

$$=\underset{T\sim D^{2m},T\to[S,S']}{P}(\forall e\in E: e\in S'|E\subset T)$$

$$\leq\left(\frac{1}{2}\right)^l=2^{-l}$$

(d) $From\ (c)\ we\ know,$

$$\underset{T\sim D^{2m},T\to[S,S']}{P}(\hat{R}_S(h)=0\cap\hat{R}_{S'}(h)>\tfrac{\varepsilon}{2}|\hat{R}_T(h_0)>\tfrac{\varepsilon}{2})\leq 2^{-l}$$

$and\ given\ from\ (c), l>-\frac{m\varepsilon}{2}, we\ can\ have\ a\ looser\ bound$

$$\underset{T\sim D^{2m},T\to[S,S']}{P}(\hat{R}_S(h)=0\cap\hat{R}_{S'}(h)>\tfrac{\varepsilon}{2})\leq 2^{-\frac{m\varepsilon}{2}}$$

(e) $P(R(h)>\varepsilon)$

$(h_0\in H_S,\ h\in H, H_S\subset H)$

$=P(R(h_0)>\varepsilon\cap R(h)>\varepsilon)$

$\leq P(R(h_0)>\varepsilon)$

$(from\ (b))$

$$\leq 2P(\sup_{h\in H_S}\left|\hat{R}_S(h)-\hat{R}_{S'}(h)\right|>\tfrac{\varepsilon}{2})$$

$(rewritten\ version\ from\ (c))$

$$=2\underset{T\sim D^{2m},T\to[S,S']}{P}(\exists h\in H:\hat{R}_S(h)=0\cap\hat{R}_{S'}(h)>\tfrac{\varepsilon}{2})$$

$(by\ union\ bound)$

$$\leq 2|H|\underset{T\sim D^{2m},T\to[S,S']}{P}(\hat{R}_S(h)=0\cap\hat{R}_{S'}(h)>\tfrac{\varepsilon}{2})$$

$$=2\prod_H(2m)P(\hat{R}_S(h)=0\cap\hat{R}_{S'}(h)>\tfrac{\varepsilon}{2})$$

$(by\ (d)\ and\ corollary\ 3.3\ in\ textbook)$

$$\leq 2\left(\frac{e2m}{d}\right)^d 2^{-\frac{\varepsilon m}{2}}$$