DATE / / SUBJECT:

Exercise chapter 18:

18_2:

۱- الگوریتم ابتدا feature ای را انتخاب می‌کند که بیشترین information gain را داشته باشد:

$IG(X) = H(Y) - H(Y|X)$

$P(X_1=1) = \frac{3}{4} \quad P(X_2=1) = \frac{1}{2} \quad P(X_3=1) = \frac{1}{2}$

$P(X_1=0) = \frac{1}{4} \quad P(X_2=0) = \frac{1}{2} \quad P(X_3=0) = \frac{1}{2}$

$$H(Y) = \overbrace{-P(Y=1)\log_2 P(Y=1)}^{1/2} \overbrace{- P(Y=0)\log_2 P(Y=0)}^{1/2} = \frac{1}{2} + \frac{1}{2} = 1$$

$$H(Y|X_1) = -\frac{3}{4}\left(P(Y=1|X_1=1)\log_2 P(Y=1|X_1=1) + P(Y=0|X_1=1)\log_2 P(Y=0|X_1=1)\right)$$

$$-\frac{1}{4}\left(P(Y=1|X_1=0)\log_2 P(Y=1|X_1=0) + P(Y=0|X_1=0)\log_2 P(Y=0|X_1=0)\right) =$$

$$-\frac{3}{4}\left(\frac{2}{3}\log_2 \frac{2}{3} + \frac{1}{3}\log_2 \frac{1}{3}\right) - \frac{1}{4}\left(0\log_2 0 + 1\log_2 1\right) = 0{,}69$$

$$H(Y|X_2) = -\frac{1}{2}\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{2}\log_2 \frac{1}{2}\right) - \frac{1}{2}\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{2}\log_2 \frac{1}{2}\right) = 1$$

$H(Y|X_3) = 1$

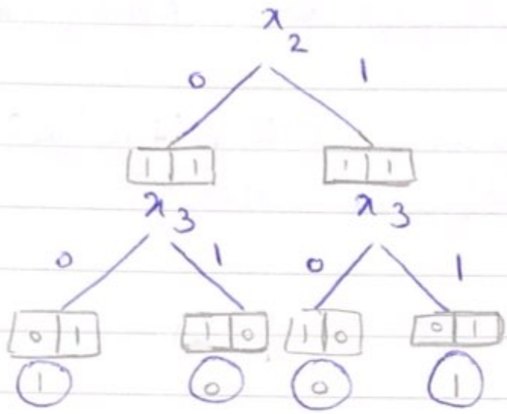
$IG(X_1) = 0{,}31 \quad IG(X_2) = 0 \quad IG(X_3) = 0$

بنابراین X_1 انتخاب می‌شود.

اکنون در هر برگ داریم ((1,1,0),0) - ((1,0,0),1) - ((1,1,1),1)

در یک زیرشاخه قرار می‌گیرند بنابراین هر lable ای به این زیرشاخه داده شود، حداقل یک نقطه از training set را اشتباه پیش‌بینی می‌کند. و چون ۴ نقطه در training set داریم بنابراین حداقل میزان خطا برابر $\frac{1}{4}$ است.

_2



Exercises chapter 10 :

10_1 :

k قسمت به اندازه $m_H(\epsilon/2)$ را انتخاب می‌کنیم و الگوریتم A را روی هر k قسمت اجرا می‌کنیم:

داریم: $\hat{h}_1, \dots, \hat{h}_k$

می‌دانیم $P(\min_{i\in[k]} L_D(\hat{h}_i) \le \min L_D(h) + \epsilon/2)$ حداقل $1-\delta_0^k \ge 1-\delta/2$ است.

اکنون ... ERM را روی کلاس $H = \{\hat{h}_1, \dots, \hat{h}_k\}$ اجرا می‌کنیم و حاصل را $\hat{h}$ نمایش می‌دهیم.

اندازه‌ی آخرین قسمت (training set) را $\left\lceil \frac{2\log(4k/\delta)}{\epsilon^2} \right\rceil$ در نظر می‌گیریم.

با استفاده از corollary 4.6 نتیجه می‌گیریم، با احتمال حداقل $1-\delta/2$

$$L_D(\hat{h}) \le \min_{i\in[k]} L_D(\hat{h}_i) + \epsilon/2$$

union bound: $L_D(\hat{h}) \le \min L_D(\hat{h}_i) + \epsilon/2 \le \min_{h\in H} L_D(h) + \epsilon$ ☑

DATE / / SUBJECT:

۱۰-۴

۱) اگر X یک مجموعه‌ی متناهی با اندازه‌ی n باشد، و B مجموعه همه توابع موجود از X به $\{0,1\}$،
در این صورت $L(B,T) = B$، و هر دو مجموعه متناهی هستند بنابراین طبق
قضیه اساسی یادگیری ماشین برای هر $T \geq 1$

$$VCdim(B) = VCdim(L(B,T)) = \log 2^n = n$$

-2

$$B = \{ Sign(\theta - x_j).b : j \in [d], b \in \{-1,1\}, \theta \in \mathbb{R}\}$$

برای هر $j \in [d]$ قرار می‌دهیم: $B_j = \{Sign(\theta - x_j).b : b \in \{-1,1\}, \theta \in \mathbb{R}\}$

در نتیجه $VCdim(B_j) = 2$،

$$B = \bigcup_{j=1}^{d} B_j$$

طبق راهنمایی در صورت سؤال با استفاده از تمرین 11:

$$VCdim(B) \leq 16 + 2\log d$$

د- با توجه به راهنمایی برای هر $i \in [\frac{Tk}{2}]$ قرار می‌دهیم $x_i = [\frac{i}{k}] A_{i,\rightarrow}$.

نشان می‌دهیم که مجموعه $C = \{x_i : i \in [\frac{Tk}{2}]\}$ توسط $L(B_d, T)$، shatter می‌شود.

قرار می‌دهیم: $I \subseteq [\frac{Tk}{2}]$ بنابراین $I = I_1 \cup \dots \cup I_{T/2}$،
در آن I_t یک زیرمجموعه از $\{(t-1)k+1, \dots, tk\}$ است
برای هر $t \in [T/2]$، j_t را ستون A در نظر می‌گیریم یعنی اگر $(t-1)k + i \in I_t$
$A_{ij} = 1$.

$$h(x) = Sign((h_{j1,-1,1/2} + h_{j1,1,3/2} + h_{j2,-1,3/2} + h_{j2,1,5/2} + \dots$$

$h_{j,b,\theta}$

$+ h_{j_{T/2-1}, -1, T/2-3/2} + h_{j_{T/2-1}, 1, T/2-1/2} + h_{j_{T/2}, 1, T/2-1/2})(x))$

بنابراین $h(x_i) = 1$ اگر $i \in I$ ✓.

Exercises chapter 11:

11-1:

اگر S یک نمونه iid و h فرضیه الگوریتم توضیح داده شده در صورت سؤال باشد. چون h یک تابع ثابت است، $L_D(h) = 1/2$

برای محاسبه $L_V(h)$ فرض می‌کنیم parity S، 1 باشد. اگر با توجه به الگوریتم Leave-one-out، یک fold را به صورت زیر در نظر بگیریم:

$\{(x, y)\} \subseteq S$

در این صورت دو حالت داریم:

1- Parity مجموعه $S - \{x\}$ برابر با 1 باشد، یعنی x صفر بوده است. بنابراین با توجه به الگوریتم برای training set: $S - \{x\}$، فرضیه $h(x) = 1$ است. پس خطای تخمین leave-one-out برای این fold برابر با 1 است.

2- Parity مجموعه $S - \{x\}$ برابر صفر باشد، یعنی x یک است. پس فرضیه الگوریتم $h(x) = 0$ می‌شود (برای $S - \{x\}$) و باز هم مقدار leave-one-out برابر با 1 می‌شود (نقطه اشتباه predict می‌شود)

بنابراین خطای تخمین نهایی L.O.O برای تمام fold ها برابر با 1 می‌شود. در نتیجه تفاوت خطای حقیقی و تخمین برابر با: $1 - 1/2 = \frac{1}{2}$ است.

2-11:

سلسله‌ای را در نظر می‌گیریم که $H_1 \subseteq H_2 \subseteq \dots \subseteq H_k$ و برای هر $i \in k$: $|H_i| = 2^i$

در H_k ~~...~~ Agnostic PAC model داریم زیرا برای یک فرضیه ERM h ادعا می‌کند:

$$L_D(h) \leq \min_{h \in H_k} L_D(h) + \sqrt{\frac{2(k+1+\log(1/\epsilon))}{m}}$$

DATE / / SUBJECT:

فرض کنید j کوچکترین index ای است که شامل $h^* \in \underset{h\in H}{\operatorname{argmin}} L_D(h)$ است. r را یک عدد ثابت عضو $[k]$ در نظر می‌گیریم. طبق نامساوی Hoeffding با احتمال حداقل $1-\frac{\delta}{2k}$:

$$|L_D(\hat{h}_r)-L_V(\hat{h}_r)| \leq \sqrt{\frac{1}{2\alpha m}\log\frac{4}{\delta}}$$

با استفاده از union bound، با احتمال حداقل $1-\delta/2$:

$$L(\hat{h}) \leq L_V(\hat{h})+\sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}} \leq L_V(\hat{h}_r)+\sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}}$$

$$\leq L_D(\hat{h}_r)+2\sqrt{\frac{1}{2\alpha m}\log\frac{4k}{\delta}} = L_D(\hat{h}_r)+\sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}$$

پس با احتمال حداقل $1-\delta/2$ داریم:

$$L_D(\hat{h}) \leq L_D(\hat{h}_j)+\sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}$$

$$L_D(\hat{h}_j) \leq L_D(h^*)+\sqrt{\frac{2}{(1-\alpha)m}\log\frac{4|H_j|}{\delta}}$$

با استفاده از union bound:

$$L_D(\hat{h}) \leq L_D(h^*)+\sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}+\sqrt{\frac{2}{(1+\alpha)m}\log\frac{4|H_j|}{\delta}}$$

$$L_D(\hat{h}) \leq L_D(h^*)+\sqrt{\frac{2}{\alpha m}\log\frac{4k}{\delta}}+\sqrt{\frac{2}{(1-\alpha)m}(j+\log\frac{4}{\delta}}$$

با مقایسه این دو کران می‌بینیم وقتی optimal index j به طور قابل توجهی کوچکتر از k باشد

کران ما با استفاده از model selection به دست می‌آید بسیار کمتر است. اگر j برحسب k
لگاریتمی باشد، به logarithmic improvement دست می‌یابیم.