۱۰،۱

۲،۱۱، ۱،۱۱ و ۳،۱، ۱،۱، ۱،۰

اگر $\epsilon, \delta \in (0,1)$ باشند. k بخش از سایز m انتخاب کرده و الگوریتم A را روی هر کدام از این بخش ها اجرا کنید برای بدست آوردن $\hat{h}_1, \dots, \hat{h}_k$.

در هر بخش دیدیم که احتمال $\min_{i \in K} L_D(\hat{h}_i) \leq \min L_D(h) + \epsilon/2$ حداقل $1 - \delta_0/2$ است. حال الگوریتم ERM را روی کلاس $\hat{H} = \{\hat{h}_1, \dots, \hat{h}_k\}$ با استفاده از داده های آموزش باقیمانده حداقل $\left[m = \frac{2\log(4k/\delta)}{\epsilon^2} \right]$ در نظر می گیریم.

می دانیم که $L_D(\hat{h}) \leq \min L_D(\hat{h}_i) + \epsilon/2$ که با اجرای union bound می‌توان نتیجه گرفت که

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

۱۰,۲

① اگر X یک مجموعه متناهی به سایز n باشد، اگر B مجموعه توابع شامل
تمام توابعی از $X \to \{0,1\}$ باشد آنگاه $B = L(B,T)$ است که هر دو
متناهی هستند. بنابراین برای $T \geq 1$:

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

② با توجه به راهنمایی، هر $i \in [TK/2]$ باشد و $X_i = [i/k]A_i$ یک مجموعه ای
$\mathcal{C} = \{X_i : i \in [TK/2]\}$ که توسط $L(B_d,T)$ شکسته می‌شود، اگر $I \subseteq [TK/2]$
باشد پس $I = I_1 \cup \dots \cup I_{T/2}$

$h(x_i) = 1$ if $i \in I$

, $h \in L(B_d, T)$

$\boxed{11.1}$

اگر s مستقل و هم توزیع باشد. و h یک مدل خوب از الگوریتم باشد.

پس تا زمانی که h یک تابع ثابت باشد $L_D(h) = 1/2$

حال می خواهیم $L_V(h)$ را محاسبه کنیم.

/ $S/x = 1$ می شود: اگر $y = 0$ باشد وقتی با گذرش را انجام می دهیم از S/x که استفاده
می کنیم خروجی الگوریتم $h(x) = 1$ را پیش بینی می کند بنابراین با برآورد leave one out
از fold $= 1$ استفاده کنید.

✓ اگر $S/x = 0$ شود: اگر $y = 1$ باشد با توجه به توضیح های قبل $h(x) = 0$
و با استدلال از آن بالا نیز fold 1 می گیرد.

/ تفاوت بین خطای واقعی و خطای برآورد 6 $\frac{1}{2}$ است

۱۸.۲

الف)

از گره اول شروع می‌کنیم و دنبال ویژگی هایی می‌گردیم که آنتروپی ما را ماکسیمم کنند. برای feature 1، آنتروپی برابر:

$$H(1/2) - \left(3/4\, H(2/3) + 1/4\, H(0)\right) \approx 0.32$$

برای feature 2 و 3:

$$H(1/2) - \left(1/2\, H(1/2) + 1/2\, H(1/2)\right) = 0$$

اگر $X_1 = 0$ بگذاریم هر عمقی از درخت نمی‌توان تمام داده ها را طبقه‌بندی کرد

بنابراین درم حداقل یکی از مثال‌ها اشتباه لیبل می‌دهد مثلاً در ۴ مثال احتمال

خطای آموزشی حداقل $1/4$ است.

ب)

$x_3 = 0$

ب)