Subject: ۴۰۳۱۰۶۶۸۱
Date

محمدی مرادی

۱- الف) درست

در این الگوریتم برای ساده شدن، تمامی Feature ها را مستقل در نظر می‌گیرد

ب) درست

با انجام smoothing، احتمال وقایعی که در داده تست صفر بوده، از صفر بزرگتر می‌شود و دیگر overfitting نخواهیم داشت

ج) نادرست

یکی از قدرت های درخت تصمیم این است که می‌تواند با انعطاف بالا مسائل مختلف را مدل کند تا حدی که امکان overfitting دارد

د) درست

عمق بیشتر یعنی دانستن جزئیات بیشتر از داده تست که این احتمال overfit را بالا می‌برد

هـ) درست

$IG(X)=H(Y)-H(Y|X)$. هر چقدر مقادیر یکتا زیاد شود $H(Y|X)$ کم می‌شود و $IG(X)$ بالا می‌رود

Subject:
Date

ه) نادرست

درخت تصمیم می‌تواند ریزترین جزئیات یک مسئله را بیرون بکشد و ذاتا غیرخطی است

ز) درست

مدل وقتی هم روی train و validation بدعمل کند underfitting رخ داده

ح) نادرست

ابرپارامترها از داده Validation (Held-out) تنظیم می‌شود

CamScanner

Subject: Date:

۲- الف) بهترین ویژگی، بالاترین $IG(X)$ را دارد که یعنی پایین ترین $H(Y|X)$.

$$H(Y|X_1) = -\sum_{X} P(X) \sum_{Y} P(Y|X) \log_2 P(Y|X)$$

$$= -\left(\frac{4}{10}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{10}\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right)\right)$$

$$= (0/4 \times 0/811 + 0/6 + 0/918) = \boxed{0/874}$$

$$H(Y|X_2) = -\left(\frac{3}{10}\left(1\log_2 1 + 0\log_2 0\right) + \frac{7}{10}\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right)\right)$$

$$= (0/7 \times 0/985) = \boxed{0/689}$$

$$H(Y|X_3) = -\left(\frac{4}{10}\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{10}\left(\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right)\right)$$

$$= (0/4 \times 0/811 + 0/6 \times 0/918) = \boxed{0/874}$$

پس بهترین ویژگی جهت انشعاب X_2 است

ب) وقتی عمق درخت را تا ته می بریم و برگ ها pure می شوند، ما تمامی جزئیات ممکن درباره داده تست را به مدل آموزش می دهیم که منجر به overfitting می شود که اینگونه در مواجه با داده جدید، به دلیل عدم generalization، بدعمل می کند

DAT

Subject: Date:

ج)

$$\text{Gini Impurity} = 1 - \sum (p_i)^2$$

$$G(X_1) = \left(\frac{4}{10}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) + \frac{6}{10}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right)\right)$$

$$= 0/4 \times 0/375 + 0/6 \times 0/444 = \boxed{0/416}$$

$$G(X_2) = \left(\frac{3}{10} \times 0 + \frac{7}{10}\left(1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2\right)\right)$$

$$= 0/7 \times 0/489 = \boxed{0/342}$$

$$G(X_3) = \left(\frac{4}{10}\left(1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2\right) + \frac{6}{10}\left(1 - \left(\frac{2}{6}\right)^2 - \left(\frac{4}{6}\right)^2\right)\right)$$

$$= 0/4 \times 0/375 + 0/6 \times 0/444 = \boxed{0/416}$$

کمترین ناخالصی را X_2 داشت پس برای انشعاب انتخاب می‌شود

DAT

CamScanner

Subject: Date:

۳_الف)

$$p(Y=spam \mid W_1=best, W_2=price)$$

$$= \underbrace{p(Y=spam)}_{p} \times \frac{1}{8} \times \frac{1}{4} = \frac{p}{32}$$

$$p(Y=ham \mid W_1=best, W_2=price)$$

$$= (1-p) \times \frac{1}{12} \times \frac{1}{6} = \frac{1-p}{72}$$

$$\Longrightarrow \frac{p}{\cancel{32}_{4}} > \frac{1-p}{\cancel{72}_{9}} \Longrightarrow 9p > 4-4p \Longrightarrow \boxed{p > \frac{4}{13}}$$

ب)

$$p(W=just \mid Y=spam) = \frac{2}{10}$$

$2 \rightarrow$ تعداد تکرار just

$10 \rightarrow$ تعداد کلمات دو لیبل spam

$$p(W=call \mid Y=ham) = \frac{1}{8}$$

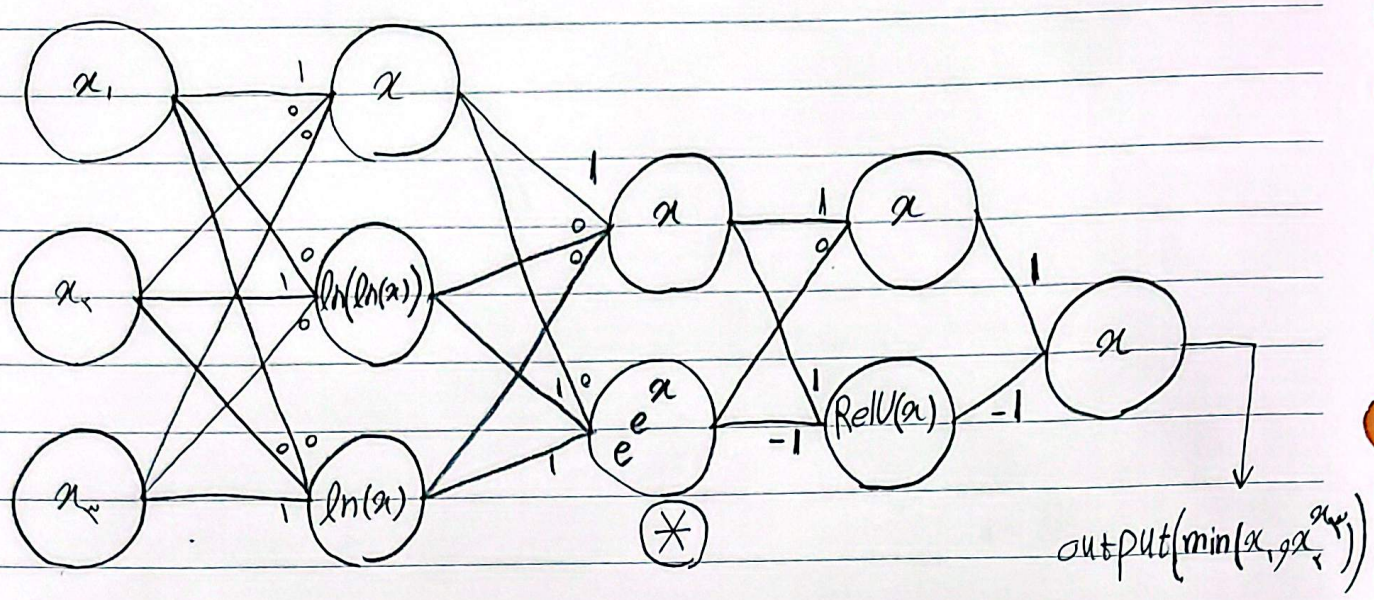
$$p(W=offer \mid Y=ham) = \frac{0}{8} = 0$$

$$p(Y=spam) = \frac{2}{5}$$

DAT

Subject: Date:

۴ - الف)



در نورون (*) ما داریم:

$$e^{e^{\ln(\ln(x_2))+\ln(x_3)}} = e^{x_3 \ln(x_2)} = e^{\ln(x_2^{x_3})} = x_2^{x_3}$$

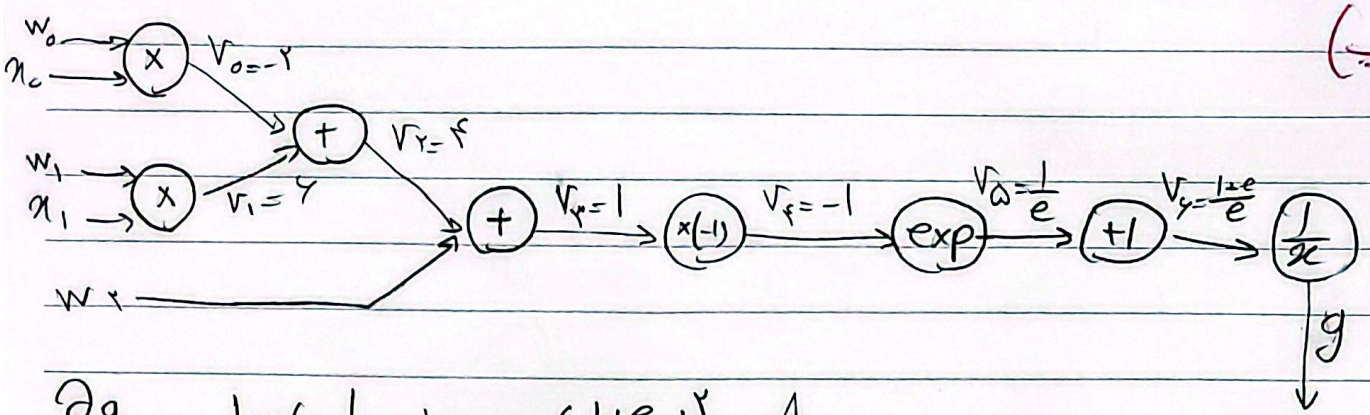
و $a - \mathrm{ReLU}(a-b) = \min(a, b)$

$\Longrightarrow x_1 - \mathrm{ReLU}(x_1 - x_2^{x_3}) = \min(x_1, x_2^{x_3})$ در دقایق آخر داریم:

DAT

Subject: Date:

ب)



$$\frac{\partial g}{\partial v_6} = 1\times\left(-\frac{1}{(v_6)^2}\right) = -\left(\frac{1+e}{e}\right)^2 = A$$

$$\frac{\partial g}{\partial v_5} = \frac{\partial g}{\partial v_6}\cdot\frac{\partial v_6}{\partial v_5} = A\times 1 = A$$

$$\frac{\partial g}{\partial v_4} = \frac{\partial g}{\partial v_5}\cdot\frac{\partial v_5}{\partial v_4} = A\times e^{v_4} = \frac{A}{e}$$

$$\frac{\partial g}{\partial v_3} = \frac{\partial g}{\partial v_4}\cdot\frac{\partial v_4}{\partial v_3} = \frac{A}{e}\times(-1) = -\frac{A}{e}$$

$$\frac{\partial g}{\partial v_2} = \frac{\partial g}{\partial v_3}\cdot\frac{\partial v_3}{\partial v_2} = \left(-\frac{A}{e}\right)\times 1 = -\frac{A}{e}$$

$$\frac{\partial g}{\partial v_1} = \frac{\partial g}{\partial v_2}\cdot\frac{\partial v_2}{\partial v_1} = \left(-\frac{A}{e}\right)\times 1 = -\frac{A}{e}$$

$$\frac{\partial g}{\partial v_0} = \frac{\partial g}{\partial v_2}\cdot\frac{\partial v_2}{\partial v_0} = \left(-\frac{A}{e}\right)\times 1 = -\frac{A}{e}$$

$$\frac{\partial g}{\partial w_2} = \frac{\partial g}{\partial v_3}\cdot\frac{\partial v_3}{\partial w_2} = \left(-\frac{A}{e}\right)\times 1 = -\frac{A}{e}$$

$$\frac{\partial g}{\partial w_1} = \frac{\partial g}{\partial v_1}\cdot\frac{\partial v_1}{\partial w_1} = \left(-\frac{A}{e}\right)\times(-2) = \frac{2A}{e}$$

$$\frac{\partial g}{\partial x_1} = \frac{\partial g}{\partial v_1}\cdot\frac{\partial v_1}{\partial x_1} = \left(-\frac{A}{e}\right)\times(-3) = \frac{3A}{e}$$

$$\frac{\partial g}{\partial w_0} = \frac{\partial g}{\partial v_0}\cdot\frac{\partial v_0}{\partial w_0} = \left(-\frac{A}{e}\right)\times(-1) = \frac{A}{e}$$

$$\frac{\partial g}{\partial x_0} = \frac{\partial g}{\partial v_0}\cdot\frac{\partial v_0}{\partial x_0} = \left(-\frac{A}{e}\right)\times(2) = -\frac{2A}{e}$$

DAT

Subject: Date:

۵- الف) Binary cross entropy از SSE برای مسائل classification

به دلیل محدب بودن بهتر است و در local minima گیر نمی افتد

ب)

$$J(w) = \sum_{i=1}^{n} \left[-y^{(i)} \log(\hat{y}^{(i)}) - (1-y^{(i)}) \log(1-\hat{y}^{(i)}) \right]$$

مبنای log را e می گیریم

$$\frac{\partial J}{\partial \hat{y}^{(i)}} = \sum_{i=1}^{n} -\frac{y^{(i)}}{\hat{y}^{(i)}} + \frac{1-y^{(i)}}{1-\hat{y}^{(i)}} = \frac{\hat{y}^{(i)} - y^{(i)}}{\hat{y}^{(i)}(1-\hat{y}^{(i)})} \quad \text{I}$$

$$\hat{y}^{(i)} = \sigma(z) \Longrightarrow \frac{\partial \hat{y}^{(i)}}{\partial z} = -\frac{-e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{(1+e^{-z})^2} \quad \text{II}$$

$$z = w^T x^{(i)} \Longrightarrow \frac{\partial z}{\partial w} = x^{(i)} \quad \text{III}$$

$$\text{I} \times \text{II} \times \text{III} \Longrightarrow \frac{\hat{y}^{(i)} - y^{(i)}}{\hat{y}^{(i)}(1-\hat{y}^{(i)})} \times \frac{e^{-z}}{(1+e^{-z})^2} \times x^{(i)}$$

$$= \frac{\hat{y}^{(i)} - y^{(i)}}{\cancel{\hat{y}^{(i)}(1-\hat{y}^{(i)})}} \times \cancel{\hat{y}^{(i)}(1-\hat{y}^{(i)})} \times x^{(i)} = \boxed{(\hat{y}^{(i)} - y^{(i)}) x^{(i)}}$$

$$\boxed{\frac{\partial J}{\partial w} = \sum_{i=1}^{n} \left(\sigma(w^T x^{(i)}) - y^{(i)} \right) x^{(i)}}$$

ج)

$$w^{(i)} = w^{(i)} - \alpha \nabla_w J$$

$$\boxed{\Longrightarrow w^{(i)} = w^{(i)} - \alpha \left(\sigma(w^T x^{(i)}) - y^{(i)} \right) x^{(i)}}$$

DAT

CamScanner

Subject: Date:

ج) اگر multi-class نبود از تابع softmax به جای sigmoid استفاده می‌کنیم

$$\text{softmax} \Longrightarrow \hat{y}^{(i)} = \frac{e^{w^{T(i)} \cdot x^{(i)}}}{\sum_{j=1}^{3} e^{w_j^{T(i)} \cdot x^{(i)}}}$$

برای تابع loss هم داریم:

$$J(w) = \sum_{i=1}^{n} \sum_{j=1}^{3} y_j^{(i)} \log(\hat{y}_j^{(i)})$$

DAT

Subject: Date:

۶ـ الف) اگر $y \neq \hat{y} \Longrightarrow W = W + \eta y X$

$W^T X_A = [0\ 0\ 0]\begin{bmatrix}1\\2\\1\end{bmatrix} = 0 \quad sign(0) = +1 \Rightarrow y = \hat{y}$ ✓

$W^T X_B = [0\ 0\ 0]\begin{bmatrix}1\\-1\\2\end{bmatrix} = 0 \quad sign(0) = +1 \Rightarrow y = \hat{y}$ ✓

$W^T X_C = [0\ 0\ 0]\begin{bmatrix}1\\1\\-2\end{bmatrix} = 0 \quad sign(0) = +1 \Rightarrow y \neq \hat{y} \Rightarrow W = \begin{bmatrix}0\\0\\0\end{bmatrix} - \begin{bmatrix}1\\1\\-2\end{bmatrix} = \begin{bmatrix}-1\\-1\\2\end{bmatrix}$

$W^T X_D = [-1\ -1\ 2]\begin{bmatrix}1\\-2\\-1\end{bmatrix} = -1 \quad sign(-1) = -1 \Rightarrow y = \hat{y}$ ✓

بردار وزن آپدیت شده: $W = \begin{bmatrix}-1\\-1\\2\end{bmatrix}$

ب) $W^T X_A = [-1\ -1\ 2]\begin{bmatrix}1\\2\\1\end{bmatrix} = -1 \quad sign(-1) = -1 \Rightarrow y \neq \hat{y}$

$W^T X_B = [-1\ -1\ 2]\begin{bmatrix}1\\-1\\2\end{bmatrix} = 4 \quad sign(4) = +1 \Rightarrow y = \hat{y}$ ✓

C و D ⟵ در الف بررسی شد و درست پیش بینی شده

با اینکه در یک مرحله نتوانستیم تمامی داده‌ها را درست دسته‌بندی کنیم ولی به این معنا نیست که خطی جدایی پذیر نیستند. در واقع بعد از اجرای یک مرحله دیگر به وزن $\begin{bmatrix}0\\1\\2\end{bmatrix}$ می‌رسیم که تمامی داده‌ها را درست دسته‌بندی می‌کند

DAT

Subject: Date:

ج) در ب گفته شد که با وزن $\begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ داده ها خطی جدایی پذیر هستند پس

با استفاده از این وزن یک مرز تعیین می کنیم

$$\Rightarrow 0\times f_0 + 1\times f_1 + 3\times f_2 = f_1 + 3f_2 = 0$$

معادله $f_1 + 3f_2 = 0$ یک مرز تصمیم است

A: $2+3>0$ ✓ B: $-1+6>0$ ✓

C: $1-6<0$ ✓ D: $-2-3<0$ ✓

DAT

۷- الف) برای توابع محدب: $0 \leq \theta \leq 1$ $\quad f(\theta x+(1-\theta)y) \leq \theta f(x)+(1-\theta)f(y)$

• $f(x)=\sqrt{x} \Rightarrow \sqrt{\theta x+(1-\theta)y} \overset{?}{\leq} \theta\sqrt{x}+(1-\theta)\sqrt{y}$

$\xrightarrow{()^2} \theta x+(1-\theta)y \overset{?}{\leq} \theta^2 x+(1-\theta)^2 y+2\theta(1-\theta)\sqrt{xy}$

$\Rightarrow \theta(1-\theta)x+\theta(1-\theta)y \overset{?}{\leq} 2\theta(1-\theta)\sqrt{xy}$

$\xrightarrow[\div\,\theta(1-\theta)]{} \frac{x+y}{2} \overset{?}{\leq} \sqrt{xy}$

اما می‌دانیم که میانگین حسابی دو عدد از میانگین هندسی دو عدد بزرگتر است

$\Rightarrow x^2+y^2+2xy \overset{?}{\leq} 4xy \Rightarrow (x-y)^2 \leq 0$ ✗

پس $f(x)=\sqrt{x}$ محدب نیست

• $f(x)=x^2 \Rightarrow (\theta x+(1-\theta)y)^2 \overset{?}{\leq} \theta x^2+(1-\theta)y^2$

$\Rightarrow \theta^2 x^2+(1-\theta)^2 y^2+2\theta(1-\theta)xy \overset{?}{\leq} \theta x^2+(1-\theta)y^2$

$\Rightarrow \theta(1-\theta)x^2+\theta(1-\theta)y^2-2\theta(1-\theta)xy \overset{?}{\geq} 0$

$\Rightarrow \theta(1-\theta)(x-y)^2 \geq 0$ ✓ $\quad \theta \geq 0 \quad 1-\theta \geq 0 \quad (x-y)^2 \geq 0 \Rightarrow$ $f(x)=x^2$ محدب است

DAT

CamScanner

Subject: Date:

ب)

- $f(x) = -\log(x) \quad 0 < x$

$f'(x) = \frac{-1}{x} \Rightarrow f''(x) = \frac{1}{x^2}$

$f''(x)$ در بازه $(0, \infty)$ مثبت است پس $f(x)$ محدب است

- $f(x) = x^4 - 2x^2 \quad x \in R$

$f'(x) = 4x^3 - 4x \Rightarrow f''(x) = 12x^2 - 4$

$f''(x)$ به ازای $x=0$ منفی است پس $f(x)$ محدب نیست

DAT

CamScanner

Subject: Date:

۱- الف) X و K مربعی هستند پس بُعد ماتریس خروجی:

طول و عرض برابر

$$\frac{W/H - f_{W/H} + 2p}{S_{W/H}} + 1 = \frac{4-2+2}{1} + 1 = 5$$

$$X = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 & 0 \\ 0 & 3 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 3 & 1 & 0 \\ 0 & 2 & 3 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}_{6\times 6}$$

X با padding ⟸

$$K = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}$$

برای درایه اول ماتریس خروجی محاسبات را انجام می‌دهیم. باقی درایه‌ها به طور مشابه محاسبه می‌شوند

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} * \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} = 0\times 1 + 0\times(-1) + 0\times 0 + 1\times 1 = 1$$

$$O = \begin{bmatrix} 1 & 2 & 0 & 1 & 0 \\ 2 & 0 & 4 & 0 & 1 \\ -3 & 3 & 2 & 2 & 1 \\ 2 & 2 & -1 & 4 & 1 \\ -2 & -1 & 2 & -1 & 2 \end{bmatrix}_{5\times 5}$$

DAT

Subject: Date:

ب) i- attention score می‌شود ضرب q ها و k ها ($q_i \cdot k_j$) یا به صورت ماتریسی QK^T

$$QK^T = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}$$

$$\xrightarrow[\div\sqrt{2}]{\text{scale}} \begin{bmatrix} 1.41 & 2.12 \\ 0.7 & 0.7 \end{bmatrix} = S$$

ii-

$$\text{softmax}(S) \Longrightarrow \begin{bmatrix} 4.11 & 8.34 \\ 0.5 & 0.5 \end{bmatrix}$$

$$\xrightarrow{\div \text{sum}} \begin{bmatrix} \frac{4.11}{12.45} & \frac{8.34}{12.45} \\ 0.5 & 0.5 \end{bmatrix} = \begin{bmatrix} 0.33 & 0.67 \\ 0.5 & 0.5 \end{bmatrix} = S'$$

iii-

$$O = S'V \Rightarrow O = \begin{bmatrix} 0.33 & 0.67 \\ 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix} = \begin{bmatrix} 1.33 & 2.34 \\ 1.5 & 2 \end{bmatrix}$$

DAT

CamScanner