HW04_Artificial Intelligence

Saeed Rezvi

Student ID: 98106542

---

۱. (۲۵ نمره) مساله‌ی رگرسیون خطی ساده را درنظر بگیرید. در تعریف احتمالاتی این مساله فرض می‌کنیم رابطه‌ی زیر بین x_i و y_i وجود دارد به طوری که $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ توزیع می‌شود.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

می‌دانیم که β_0، β_1 و σ مقادیر ثابت نامنفی هستند.

الف) اثبات کنید که تخمین بیشینه درست نمایی دو پارامتر β_0 و β_1 برابر با کمینه کردن مجموع مربعات خطا است.

$D = \{(x_i, y_i)\}$

Linear regression: $\beta_0 + \beta_1 x_i$

① minimize $\sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 x_i))^2 \triangleq \underset{\beta_0, \beta_1}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 + \beta_1 x_i)^2$ (I)

بیشترین شباهت

MLE → ② maximize $P([\underline{y}, \underline{X}] \mid \beta_0, \beta_1)$ → x_i are iid → maximize $\prod_{i=1}^{N} P\{y_i, x_i\} \mid \beta_0, \beta_1)$

$\underset{\beta_0, \beta_1}{\text{argmax}}$

Negative log → minimize $- \sum_{i=1}^{N} \log(P(\{y_i, x_i\} \mid \beta_0, \beta_1))$

$\Rightarrow P(y_i, x_i \mid \beta_0, \beta_1) = P(\varepsilon_i = y_i - \beta_0 - \beta_1 x_i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2}$

$\Rightarrow$ minimize $\sum_{i=1}^{N} \frac{1}{2\sigma^2} (y_i - \beta_0 - \beta_1 x_i)^2 - \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$

$\frac{1}{2\sigma^2}$: constant — $\log\left(\frac{1}{\sqrt{2\pi}\sigma}\right)$: constant

$\triangleq \underset{\beta_0, \beta_1}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$ (II)

→ (I) = (II) → اثبات شد

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})(y_i - \bar{y})} = \frac{S_{xy}}{S_{xx}}$$

ب)

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \xrightarrow{c_i = (x_i - \bar{x})} \beta_1 = \frac{1}{S_{xx}} \sum_{i=1}^{n} c_i (y_i - \bar{Y})$$

$$= \frac{1}{S_{xx}} \left\{ c_1 \left[Y_1 - \frac{1}{n}[Y_1 + \cdots Y_n] \right] + c_2 \left[Y_2 - \frac{1}{n}[Y_1 + \cdots Y_n] \right] + \cdots c_n \left[Y_n - \frac{1}{n}(Y_1 + \cdots Y_n) \right] \right\}$$

$$= Y_1 \left\{ \frac{1}{S_{xx}} \left[(x_1 - \bar{x}) - \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x}) \right] \right\} + \cdots + Y_n \left\{ \frac{1}{S_{xx}} \left[(x_n - \bar{x}) - \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x}) \right] \right\}$$

$$= \sum_{i=1}^{n} Y_i \left\{ \frac{1}{S_{xx}} \left[(x_i - \bar{x}) - \frac{1}{n} \sum_{j=1}^{n} (x_j - \bar{x}) \right] \right\} \Rightarrow \boxed{\hat{\beta}_1 = \sum_{i=1}^{n} \frac{Y_i}{S_{xx}} (x_i - \bar{x})}$$

$$\Rightarrow \hat{\beta}_1 \sim N\left(\frac{\sum_{i=1}^{n} (\beta_0 + \beta_1 x_i)(x_i - \bar{x})}{S_{xx}} , \sum_{i=1}^{n} \sigma^2 \frac{1}{S_{xx}^2} (x_i - \bar{x})^2 \right) \quad (I)$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2\bar{x}x_i + \bar{x}^2) = \sum_{i=1}^{n} x_i^2 - 2\bar{x} n \frac{1}{n} \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x} \frac{1}{n} \sum_{i=1}^{n} x_i = \boxed{\sum_{i=1}^{n} x_i (x_i - \bar{x})} \quad (II)$$

(II), (I) $\rightarrow$ $E(\hat{\beta}_1) = \frac{\beta_0 \overbrace{\sum (x_i - \bar{x})}^{0}}{S_{xx}} + \frac{\beta_1 \sum x_i (x_i - \bar{x})}{S_{xx}} = \beta_1$ (ادیب بودن)

$$\rightarrow \boxed{\hat{\beta}_1 \sim N\left(\beta_1 , \frac{\sigma^2}{S_{xx}} \right) \equiv N\left(\beta_1 , \frac{\sigma^2}{\sum (x_i - \bar{x})^2} \right)}$$

← برای اثبات توزیع $\hat{\beta}_0$، رابطه تقریبا مشابه با ما پیچیده‌تری با شد که توزیع آن را نمی‌نویسم.

$$\boxed{\hat{\beta}_0 \sim N\left(\beta_0 , \frac{\sigma^2}{n} \left(\frac{\sum x_i^2}{\sum (x_i - \bar{x})^2} \right) \right)}$$

ب)

$$\tilde{\beta}_1 = \frac{\sum \gamma_i y_i}{\sum \gamma_i x_i} \quad \text{such that} \quad \sum_i \gamma_i = 0$$

$\Rightarrow$ we derive from previous part that $\hat{\beta}_1 = \sum \frac{y_i}{S_{XX}} (x_i - \bar{x})$

$$\Rightarrow \hat{\beta}_1 = \frac{1}{S_{XX}} \sum Y_i (x_i - \bar{x}) \xrightarrow{S_{XX} = \sum x_i (x_i - \bar{x})}$$

$$\boxed{\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x}) x_i}}$$

$$\boxed{\gamma_i = x_i - \bar{x} \quad \text{، یکی از این خانواده‌هاست}}$$

پ)

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x}) y_i}{S_{XX}} = \frac{\sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \epsilon_i)}{S_{XX}}$$

نادرست می‌باشد.

$$\Rightarrow \hat{\beta}_1 \overset{E}{=} \left(\frac{\beta_1 S_{XX}}{S_{XX}}\right) + E\left(\frac{\beta_0 \cancel{\sum (x_i - \bar{x})}^{0}}{S_{XX}}\right) + E\left(\frac{\sum (x_i - \bar{x}) \epsilon_i}{S_{XX}}\right) = \beta_1$$

ت)

$$Var(\tilde{\beta}_1) = \frac{1}{S_{XX}^2} Var\left[\sum \gamma_i [\beta_0 + \beta_1 x_i + \epsilon_i]\right]$$

$$\xrightarrow[\text{variable}]{\text{independent}} Var(\tilde{\beta}_1) = \frac{1}{S_{XX}^2} \left[\underbrace{Var(\sum \gamma_i \beta_0)}_{\alpha} + \underbrace{Var(\sum \gamma_i \beta_1 x_i)}_{\beta} + \underbrace{Var(\sum \gamma_i \epsilon_i)}_{S_{XX}\sigma^2}\right]$$

$$\Rightarrow \frac{1}{S_{XX}^2} \left[\alpha + \beta + \sigma^2 S_{XX}\right], \text{ where } \alpha, \beta > 0$$

$$\Rightarrow \boxed{Var(\tilde{\beta}_1) \geq Var(\hat{\beta}_1)} \quad \checkmark$$

۲. (۲۰ نمره) فرض کنید قصد داشته باشیم مساله‌ی رگرسیون چند متغیره را درنظر بگیریم. تابع هزینه‌ای که باید کمینه شود به فرم زیر خواهد بود.

$$\min_W F(W) = \lambda W^T W + \|XW - Y\|_۲^۲$$

الف) اگر بخواهیم این مساله را با الگوریتم Stochastic Gradient Descent حل کنیم، شبه کد آن را بنویسید.

ب) حال فرض کنید تعریف کنیم

$$W_۱ = \operatorname*{argmin}_W L(W)$$

$$W_۲ = \operatorname*{argmin}_W L(W) + \lambda W^T W$$

که $L(W)$ یک تابع نامنفی است. اثبات کنید که $\|W_۲\|_۲ \leq \|W_۱\|_۲$ و ارتباط آن را با فرمول بندی مساله بیان کنید.

A) pseudocode *W is augmented

1) initialize $W := 0^{n+1}$

2) for itteration $t \in [1, \dots T]$

2.1 draw random example with replacment (x_i)

2.2 compute loss by $F(W, x_i)$

2.3 compute gradient $\Delta W = -\frac{\partial F(W, x_i)}{\partial W}$

2.4 update parameter: $W := W + \gamma \Delta W$ (γ: coefficient)

B) $W_1 = \operatorname*{argmin}_W L(W)$ (I) $\quad W_2 = \operatorname*{argmin}_W L(W) + \lambda W^T W$ (II)

proof that $\|W_2\|_2 \leq \|W_1\|_2$:

→ we prove this, by contradiction, soppose $\|W_2\|_2 \geq \|W_1\|_2$

if $\|W_2\|_2 \geq \|W_1\|_2 \Rightarrow \lambda\|W_2\|_2 \geq \lambda\|W_1\|_2 \Rightarrow \lambda W_2^T W_2 \geq \lambda W_1^T W$ (1)

also we know from (I) that $\forall W_0 \in D(W) \quad L(W_1) \leq L(W_0)$

$\Rightarrow L(W_1) \leq L(W_2)$ (2)

using (1)(2) $\Rightarrow L(W_1) + \lambda W_1^T W \leq L(W_2) + \lambda W_2^T W_2 \Rightarrow W_2 \neq \operatorname*{argmin}_W L(W) + \lambda W^T W$ ← contradiction ✗

✓ $\|W_2\|_2 \leq \|W_1\|_2$

۳. (۱۰ نمره) یک دیتاست چند متغیره را درنظر بگیرید، به این معنی که $x_i \in \mathbb{R}^p$ که $p > 1$ است و $y_i \in \mathbb{R}$ است. فرض کنید مشاهده کرده‌ایم که یکی از ضرایب محاسبه شده یک مقدار خیلی بزرگ منفی نسبت به باقی متغیرها پیدا کرده است کدام یک از گزاره‌های زیر صحیح است؟ توضیح دهید.

- این ویژگی تاثیر زیادی روی مدل دارد و باید حفظ شود.
- این ویژگی تاثیر زیادی روی مدل ندارد و باید ایگنور شود.
- نمی‌توان بدون در دست داشتن اطلاعات بیشتر درمورد این ویژگی نظر داد. ✓

نمی‌توان نظر دقیقی در این رابطه داد:

۱) ممکن است این ضریب بزرگ ناشی از overfitting مدل باشد در نتیجه attribute ما مفید نباشد.

۲) ممکن است این attribute، کلید ویژگی‌ای حیاتی تعیین‌کننده باشد و عمل prediction ما بر پایه این ویژگی تعیین شود.

۴. (۱۵ نمره) با ارائه دلیل صحیح یا غلط بودن هر یک از گزاره‌های زیر را ثابت کنید.

- اگر bias زیاد است اضافه کردن تعداد داده‌های آموزش کمک زیادی به کم کردن بایاس نمی‌کند.
- کم کردن خطای مدل روی داده‌های آموزش منجر به کاهش خطای مدل روی داده‌های تست می‌شود.
- افزایش پیچیدگی مدل رگرسیون همواره منجر به کاهش خطای مدل روی داده‌ی آموزش و افزایش خطای مدل روی داده‌ی تست می‌شود.

① غلط. افزایش تعداد داده‌های آموزش باعث می‌شود خطا، روی تعداد بیشتری داده میانگین شود در نتیجه bias کمتری شود و جامعیت مدل افزایش می‌یابد.

② غلط. لزوماً نزدیکی درستی نمی‌باشد. ممکن است با کمترین خطای مدل روی داده‌ی آموزش، overfitting رخ دهد که منجر به بالاترین خطا روی داده تست می‌شود.

③ غلط! ممکن است در ابتدا، مدل رگرسیون بسیار ساده باشد. در این صورت با افزایش پیچیدگی، کاهش خطا روی داده‌ی تست را روی دارد.

۵. (۲۰ نمره) بر روی ۶ بیمار قلبی، مطالعاتی صورت گرفته است و جدول زیر از نتایج آن به دست آمده است.

HEART ATTACK	EXERCISES	SMOKES	MALE	CHEST PAIN	PATIENT ID
yes	yes	no	yes	yes	۱
yes	no	yes	yes	yes	۲
yes	no	yes	no	no	۳
no	yes	no	yes	no	۴
yes	yes	yes	no	yes	۵
no	yes	yes	yes	no	۶

الف) با استفاده از این داده‌ها درخت تصمیم گیری پیش بینی حمله قلبی را تشکیل دهید.

ب) درخت به دست آمده را به صورت تعدادی گزاره‌ی تصمیم گیری ترجمه کنید.

الگوریتم پیدا کردن درخت تصمیم گیری:

① محاسبه‌ی entropy اولیه

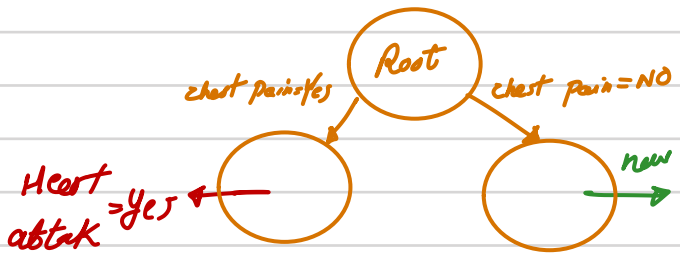
② محاسبه‌ی IG برای هر attribute

انتخاب attribute با بیشترین IG

Repeat step2 ③

$$H(initial) = -\left[\frac{4}{6}\log_2\frac{4}{6} + \frac{2}{6}\log_2\frac{2}{6}\right] = 0.918$$

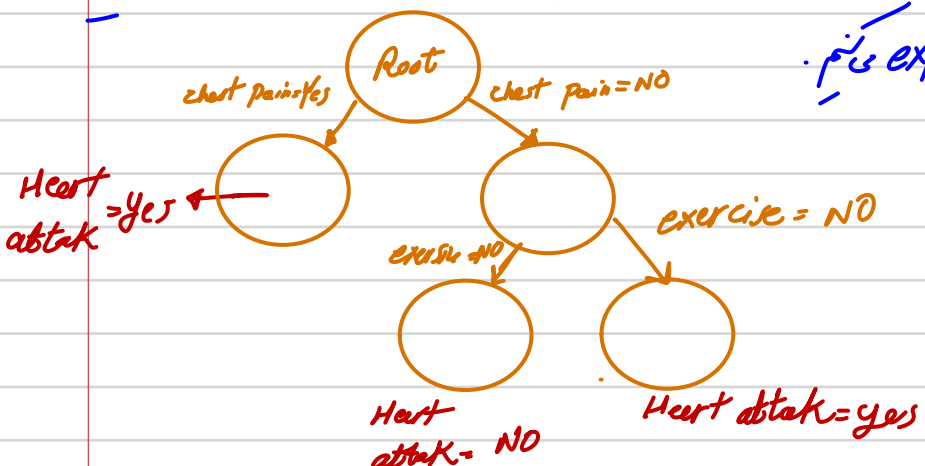
← برای مرحله‌ی ② راه ضعیف این است که expand کردن کدام attribute، پراکندگی کمتری را ایجاد می‌کند. که در این سوال واضحاً با expand کردن chest pain، کمترین پراکندگی اطلاعات ← بیشترین information gain را خواهیم داشت.



new table

HA	exercise	Smoke	Male
yes	NO	yes	NO
NO	yes	NO	yes
NO	yes	yes	yes

دوباره مرحله‌ی ② را اجرا می‌کنیم. چه Male را expand کنیم، چه exercise هردو بیشترین IG را خواهد داشت. در اینجا ما exercise را expand می‌کنیم.



→ final decision Tree

ب) ① اگر فرد، chest pain داشته باشد، قطعاً مبتلا به حمله‌ی قلبی می‌شود.

② اگر فرد chest pain نداشته باشد و مقایت ورزشی نداشته باشد مبتلا به حمله‌ی قلبی می‌شود.

③ اگر فرد chest pain نداشته باشد و مقایت ورزشی داشته باشد مبتلا به حمله‌ی قلبی نمی‌شود.

۶. (۱۰ نمره) نشان دهید هر دسته بند دودویی به فرم $h : \{۰,۱\}^d \mapsto \{۰,۱\}$ می‌تواند به صورت یک درخت تصمیم گیری به عمق حداکثر $d+۱$ با گره‌های به فرم $(x_i = ۰?)$ برای یک $i \in \{۱,\dots,d\}$ پیاده سازی شود.

← بدترین حالت زمانی رخ می‌دهد که entropy در هر گره، تمام attributeها را برابر ۱ نشود.
(یعنی با احتمال $\frac{۱}{۲}$، $y_i = ۰$ و به احتمال $\frac{۱}{۲}$، ۱ = y_i شود ← در این صورت به d گره نیاز داریم تا به leaf node برسیم. پس یک عمق دیگر که بیش از این می‌رسیم می‌دانیم "قطعاً" به یک جواب برای prediction خودی رسیم. ← با عمق $d+۱$ می‌توان decision tree را ساخت.
حداکثر