به‌نام خدا

دانشگاه صنعتی شریف

دانشکده مهندسی کامپیوتر

هوش مصنوعی

تمرین۴: ML, Decision Tree, Regression

مهراد میلانلو

۹۹۱۰۵۷۷۵

سوال ۱

الف)

می‌دانیم:

$$y_i = \beta_{\text{۰}} + \beta_{\text{۱}} x_i + \epsilon_i$$

$$\epsilon_i \sim \mathcal{N}(\text{۰}, \sigma^{\text{۲}})$$

$$\Rightarrow y_i = f(x_i; \beta) + \epsilon_i$$

$$\Rightarrow y_i = f(x_i; \beta) + \mathcal{N}(\text{۰}, \sigma^{\text{۲}})$$

اکنون از رابطه‌ی بالا امیدریاضی می‌گیریم. خواهیم داشت:

$$\mathbb{E}[y_i|x_i] = \mathbb{E}[f(x_i;\beta) + \mathcal{N}(\text{۰},\sigma^{\text{۲}})] = \mathbb{E}[f(x_i;\beta)] + \underbrace{\mathbb{E}[\mathcal{N}(\text{۰},\sigma^{\text{۲}})]}_{\text{۰}} = f(x_i;\beta)$$

$$\Rightarrow p(y_i|x_i,\beta,\sigma^{\text{۲}}) = \frac{\text{۱}}{\sqrt{\text{۲}\pi\sigma^{\text{۲}}}} exp\{-\frac{\text{۱}}{\text{۲}\sigma^{\text{۲}}}(y_i - f(x_i;\beta))^{\text{۲}}\}$$

اکنون برای به‌دست آوردن پارامترهای $\beta_{\text{۰}}, \beta_{\text{۱}}$ مقدار likelihood را با فرض این‌که (x_i, y_i) ها $i.i.d$ هستند تشکیل می‌دهیم:

$$p(\mathcal{D}|\beta,\sigma^{\text{۲}}) = p(Y|X,\beta,\sigma^{\text{۲}}) = \prod_{i=\text{۱}}^{N} p(y_i|x_i,\beta,\sigma^{\text{۲}})$$

$$\hat{\beta} = \arg\max_{\beta} = \arg\max_{\beta} \prod_{i=\text{۱}}^{N} p(y_i|x_i,\beta,\sigma^{\text{۲}})$$

برای راحتی کار، مقدار log احتمال likelihood را بیشینه می‌کنیم. چون تابع log صعودی‌است، جواب بهینه تغییری نمی‌کند. بنابراین خواهیم داشت:

$$\hat{\beta} = \arg\max_{\beta} \ln p(\mathcal{D}|\beta,\sigma) = \arg\max_{\beta} \ln \prod_{i=\text{۱}}^{N} p(y_i|x_i,\beta,\sigma^{\text{۲}})$$

$$= \arg\max_{\beta} \sum_{i=\text{۱}}^{N} \ln p(y_i|x_i,\beta,\sigma^{\text{۲}}) = \arg\max_{\beta} \sum_{i=\text{۱}}^{N} \ln(\frac{\text{۱}}{\sqrt{\text{۲}\pi\sigma^{\text{۲}}}} exp\{-\frac{\text{۱}}{\text{۲}\sigma^{\text{۲}}}(y_i - f(x_i;\beta))^{\text{۲}}\})$$

$$= \arg\max_{\beta}(-N\ln\sigma - \frac{N}{\text{۲}}\ln \text{۲}\pi - \frac{\text{۱}}{\text{۲}\sigma^{\text{۲}}} \underbrace{\sum_{i=\text{۱}}^{N}(y_i - f(x_i;\beta))^{\text{۲}}}_{sum\ of\ square\ errors})$$

همان‌طور که واضح است، بیشینه کردن مقدار likelihood پارامترهای $\beta_{\text{۰}}, \beta_{\text{۱}}$ معادل است با کمینه کردن مجموع مربعات خطای تخمین. بنابراین حکم موردنظر اثبات می‌شود.

ب)

همان‌طور که در قسمت (الف) به‌دست آوردیم، برای بیشینه کردن مقدار likelihood پارامترهای $\beta_{\text{۰}}, \beta_{\text{۱}}$ کافیست مقدار

$$\sum_{i=\text{۱}}^{N}(y_i - f(x_i;\beta))^{\text{۲}}$$

را کمینه کنیم. به همین دلیل نسبت به β از آن مشتق می‌گیریم و برابر صفر قرار می‌دهیم:

$$Z = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\begin{cases} \frac{\partial Z}{\partial \beta_0} = -2\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Z}{\partial \beta_1} = -2x_i\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i) = 0 \end{cases} \tag{۱}$$

$$\xrightarrow{(1)} \sum_{i=1}^{N} y_i - N\beta_0 - \beta_1 \sum_{i=1}^{N} x_i = 0 \Rightarrow N\bar{y} - N\beta_0 - N\beta_1\bar{x} = 0$$

$$\Rightarrow \boxed{\beta_0 = \bar{y} - \beta_1\bar{x}}$$

$$\xrightarrow{(1)} -\sum_{i=1}^{N} y_i x_i + \beta_0 \sum_{i=1}^{N} x_i + \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \Rightarrow \sum_{i=1}^{N} y_i x_i - (\bar{y} - \beta_1\bar{x}) \sum_{i=1}^{N} x_i - \beta_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$\Rightarrow \underbrace{\sum_{i=1}^{N} y_i x_i - N\bar{y}\bar{x}}_{\sum_{i=1}^{N}(y_i-\bar{y})(x_i-\bar{x})} + N\beta_1\bar{x}^2 - \beta_1 \sum_{i=1}^{N} x_i^2 = 0 \Rightarrow \sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x}) = \beta_1 \underbrace{(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2)}_{\sum_{i=1}^{N}(x_i-\bar{x}^2)}$$

$$\Rightarrow \boxed{\beta_1 = \frac{\sum_i y_i(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}}$$

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} \\ \hat{\beta}_1 = \frac{\sum_i y_i(x_i-\bar{x})}{\sum_i (x_i-\bar{x})^2} \end{cases} \tag{۲}$$

$$\xrightarrow{(2)} \mathbb{E}[\hat{\beta}_1] = \frac{1}{\sum_i (x_i - \bar{x})^2} \cdot \mathbb{E}[\sum_{i=1}^{N} y_i(x_i - \bar{x})] = \frac{1}{\sum_i (x_i - \bar{x})^2} \cdot \sum_{i=1}^{N} (x_i - \bar{x})\mathbb{E}[y_i]$$

$$\xrightarrow[y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)]{y_i = \beta_0 + \beta_1 x_i + \epsilon_i} \mathbb{E}[\hat{\beta}_1] = \frac{1}{\sum_i (x_i - \bar{x})^2} \cdot (\underbrace{\sum_{i=1}^{N} \beta_0 (x_i - \bar{x})}_{=0} + \sum_{i=1}^{N} \beta_1 x_i (x_i - \bar{x})) = \frac{\beta_1}{\sum_i (x_i - \bar{x})^2} \cdot \sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})$$

$$\Rightarrow \boxed{\mathbb{E}[\hat{\beta}_1] = \beta_1}$$

$$\xrightarrow{(2)} \mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\bar{y} - \hat{\beta}_1\bar{x}] = \mathbb{E}[\bar{y}] - \bar{x}\mathbb{E}[\hat{\beta}_1] = \mathbb{E}[\frac{\sum_i y_i}{N}] - \beta_1\bar{x}$$

$$= \frac{1}{N}\mathbb{E}[\sum_{i=1}^{N} (\beta_0 + \beta_1 x_i)] - \beta_1\bar{x} = \beta_0 + \beta_1\bar{x} - \beta_1\bar{x}$$

$$\Rightarrow \boxed{\mathbb{E}[\hat{\beta}_0] = \beta_0}$$

بنابراین مقدار امیدریاضی تخمین‌گر متغیرهای β_0, β_1 برابر با خود متغیرهاست و ثابت می‌شود تخمین‌های به‌دست آمده نااریب (Unbiased) هستند. اکنون برای به‌دست آوردن واریانس تخمین‌گر ها داریم:

(دقت کنید مانند بخش الف، (x_i, y_i) ها را $i.i.d$ در نظر می‌گیریم که فرض معقولی‌است. همچنین x_i ها برخلاف y_i ها، متغیر Stochastic نیستند و در برخورد با آن‌ها مانند ضریب رفتار می‌کنیم. از آن‌جایی که این تخمین‌گرها برابر با جمع تعدادی متغیر نرمال و ضرایب هستند، واضح است که از توزیع نرمال پیروی می‌کنند.)

$$\xrightarrow{(2)} Var(\hat{\beta}_1) = Var(\frac{\sum_i y_i(x_i-\bar{x})}{\sum_i (x_i-\bar{x})^2}) = \frac{1}{(\sum_i (x_i-\bar{x})^2)^2} \cdot \sum_{i=1}^{N} (x_i-\bar{x})^2 Var(y_i)$$

$$\Rightarrow \boxed{Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2}}$$

$$\xrightarrow{(2)} \hat{\beta}_0 = \bar{y} - \frac{\bar{x}\sum_i y_i(x_i-\bar{x})}{\sum_i x_i(x_i-\bar{x})} = \frac{\bar{y}\sum_i x_i^2 - N\bar{y}\bar{x}^2 - \bar{x}\sum_i x_i y_i + N\bar{y}\bar{x}^2}{\sum_i x_i(x_i-\bar{x})} = \frac{\bar{y}\sum_i x_i^2 - \bar{x}\sum_i x_i y_i}{\sum_i x_i(x_i-\bar{x})}$$

$$= \frac{(\sum_i y_i)(\sum_i x_i^2) - N\bar{x}\sum_i x_i y_i}{N\sum_i x_i(x_i-\bar{x})}$$

$$\Rightarrow Var(\hat{\beta}_0) = \frac{1}{N^2(\sum_i x_i(x_i-\bar{x}))^2} \cdot (N(\sum_i x_i^2)^2 - N^2\bar{x}^2\sum_i x_i^2)\sigma^2 = \frac{N(\sum_i x_i^2)(\sum_i x_i(x_i-\bar{x}))}{N^2(\sum_{x_i(x_i-\bar{x})})^2}\sigma^2$$

$$\Rightarrow \boxed{Var(\hat{\beta}_0) = \frac{\sigma^2\sum_i x_i^2}{N\sum_i (x_i-\bar{x})^2}}$$

$$\Rightarrow \hat{\beta}_1 \sim \mathcal{N}(\beta_1, \frac{\sigma^2}{\sum_i (x_i-\bar{x})^2})\ ,\ \hat{\beta}_0 \sim \mathcal{N}(\beta_0, \frac{\sigma^2\sum_i x_i^2}{N\sum_i (x_i-\bar{x})^2})$$

پ)

قرار می‌دهیم: $\gamma_i = x_i - \bar{x}$

آن‌گاه خواهیم داشت:

$$\sum_i \gamma_i = \sum_i (x_i-\bar{x}) = \sum_i x_i - \sum_i \bar{x} = \sum_i x_i - n\bar{x} = 0$$

همچنین:

$$\hat{\beta}_1 = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} = \frac{\sum_i (x_i-\bar{x})y_i}{\sum_i (x_i-\bar{x})^2} = \frac{\sum_i (x_i-\bar{x})y_i}{\sum_i (x_i-\bar{x})x_i}$$

$$\xrightarrow{\gamma_i = x_i-\bar{x}} \boxed{\hat{\beta}_1 = \frac{\gamma_i y_i}{\gamma_i x_i}}$$

بنابراین تخمین‌گر MLE $\hat{\beta}_1$ عضوی از خانواده‌ی $\tilde{\beta}_1$ است.

ت)

کافی‌ست ثابت کنیم امیدریاضی هر تخمین‌گر عضو این خانواده، برابر با خود متغیر است. بنابراین داریم:

$$\mathbb{E}[\tilde{\beta}_1] = \mathbb{E}[\frac{\sum_i \gamma_i y_i}{\sum_i \gamma_i x_i}] = \frac{1}{\sum_i \gamma_i x_i} \cdot \mathbb{E}[\sum_i \gamma_i y_i] = \frac{1}{\sum_i \gamma_i x_i} \cdot \sum_i \gamma_i \mathbb{E}[y_i]$$

$$\xrightarrow{y_i \sim \mathcal{N}(\beta_0+\beta_1 x_i, \sigma^2)} \mathbb{E}[\tilde{\beta}_1] = \frac{\beta_0 \overset{=0}{\cancel{\sum_i \gamma_i}} + \beta_1\sum_i \gamma_i x_i}{\sum_i \gamma_i x_i}$$

$$\Rightarrow \boxed{\mathbb{E}[\tilde{\beta}_1] = \tilde{\beta}_1}$$

ث)

ابتدا واریانس $\tilde{\beta}_۱$ را محاسبه می‌کنیم:

$$Var(\tilde{\beta}_۱) = Var(\frac{\sum_i \gamma_i y_i}{\sum_i \gamma_i x_i}) = \frac{۱}{(\sum_i \gamma_i x_i)^۲} \cdot Var(\sum_i \gamma_i y_i) = \frac{۱}{(\sum_i \gamma_i x_i)^۲} \cdot \sum_i \gamma_i^۲ Var(y_i)$$

$$\Rightarrow \boxed{Var(\tilde{\beta}_۱) = \frac{\sigma^۲ \sum_i \gamma_i^۲}{(\sum_i \gamma_i x_i)^۲}}$$

اکنون می‌خواهیم ثابت کنیم:

$$Var(\hat{\beta}_۱) \leqslant Var(\tilde{\beta}_۱)$$

$$\Leftrightarrow Var(\hat{\beta}_۱) = \frac{\sigma^۲}{\sum_i (x_i - \bar{x})^۲} \leqslant Var(\tilde{\beta}_۱) = \frac{\sigma^۲ \sum_i \gamma_i^۲}{(\sum_i \gamma_i x_i)^۲} \Leftrightarrow \frac{۱}{\sum_i (x_i - \bar{x})^۲} \leqslant \frac{\sum_i \gamma_i^۲}{(\sum_i \gamma_i x_i)^۲}$$

$$\Leftrightarrow (\sum_i \gamma_i x_i)^۲ \leqslant (\sum_i \gamma_i^۲)(\sum_i (x_i - \bar{x})^۲)$$

طبق نامساوی کوشی‌ـ‌شوارتز می‌دانیم:

$$(\sum_i \alpha_i \beta_i)^۲ \leqslant (\sum_i \alpha_i^۲)(\sum_i \beta_i^۲)$$

بنابراین تنها کافیست ثابت کنیم:

$$\sum_i (\gamma_i x_i) = \sum_i \gamma_i (x_i - \bar{x})$$

برای این هم داریم:

$$\sum_i \gamma_i (x_i - \bar{x}) = \sum_i \gamma_i x_i - \bar{x} \underbrace{\sum_i \gamma_i}_{=۰} = \sum_i \gamma_i x_i$$

پس حکم ثابت می‌شود زیرا:

$$(\sum_i \gamma_i x_i)^۲ = (\sum_i \gamma_i (x_i - \bar{x}))^۲ \leqslant (\sum_i \gamma_i^۲)(\sum_i (x_i - \bar{x})^۲)$$

خانواده‌ی $\tilde{\beta}_۱$ یک خانواده‌ی تخمین‌گر خطی نااریب است. همچنین نشان دادیم تخمین‌گر MLE $\hat{\beta}_۱$ عضوی از این خانواده‌است. کم‌تر بودن واریانس تخمین‌گر MLE نسبت به هر عضو دلخواه دیگری از این خانواده، نشان می‌دهد در بین تخمین‌گرهای خطی نااریب، تخمین‌گر MLE دقت بهتری دارد.

سوال ۲

الف)

در این سوال می‌خواهیم تابع $F(W)$ را کمینه کنیم. شبه‌کد الگوریتم Stochastic Gradient Descent این‌گونه است:

```
stochastic_gradient_descent(F: Cost function, η: Learning rate, num_iter: Number of Iterations, D
    : Data):
    Initialize: W ← W.
    for iteration in range(num_iter):
        S ← a random subset from Data
        for sample in S:
            W = W − η∇_W F_sample(W)
    return W
```

Listing ۱: Stochastic Gradient Descent

دقت کنید تفاوت الگوریتم SGD با GD عادی این است که در هر مرحله از الگوریتم، به‌جای این‌که مقدار خروجی را روی تمام داده‌ها آپدیت کنیم، یک زیرمجموعه‌ی رندوم از داده‌ها را گرفته و روی آن این عمل را انجام می‌دهیم.

در این‌جا تابع هزینه‌ی ما برابر است با:

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

بنابراین گرادیان این تابع که در شبه‌کد بالا نیز استفاده شده است برابر است با:

$$\boxed{\nabla F(W) = ۲\lambda W + ۲X^T(XW - Y)}$$

ب)

با استفاده از برهان خلف فرض می‌کنیم این‌گونه نباشد و داشته باشیم:

$$\|W_۱\| < \|W_۲\|$$

با توجه به این فرض، طبق تعریف واضح است که:

$$L(W_۱) \leqslant L(W_۲)$$

پس خواهیم داشت:

$$\begin{rcases} W_۱^T W_۱ < W_۲^T W_۲ \\ L(W_۱) \leqslant L(W_۲) \end{rcases} \xRightarrow{\lambda > ۰} L(W_۱) + \lambda W_۱^T W_۱ < L(W_۲) + \lambda W_۲^T W_۲ \xRightarrow{W_۲ = \arg\min L(W) + \lambda W^T W} Impossible.$$

بنابراین با رسیدن به تناقض حکم ثابت می‌شود.

در اصل عملی که انجام می‌دهیم، Regularization برای جلوگیری از مشکل Overfit است. به‌طوری که با اضافه کردن جمله‌ی $\lambda W^T W$ و کوچک‌تر کردن پارامترها، اثر آن‌ها را کم‌تر می‌کنیم. که این موضوع را در بالا نشان دادیم.

سوال ۳

نمی‌توان بدون در دست داشتن اطلاعات بیشتر در مورد این ویژگی نظر داد.

در واقع مقایسه‌ی بزرگی یک ضریب نسبت به ضرایب دیگر معیار مناسبی برای بررسی Overfit نیست. زیرا ممکن است یک ضریب به دلیل تاثیر و ارزش بالای feature مربوطه بزرگ باشد و یا از طرفی ممکن است برعکس آن رخ دهد. ممکن است واحد یک feature در مقایسه با feature های دیگر طوری باشد که ضریب آن بسیار بزرگ شود. بنابراین با داده‌ی محدود این سوال نمی‌توان نظر داد و اطلاعات بیشتری نیاز است.

سوال ۴

- **اگر bias زیاد است اضافه کردن تعداد داده‌های آموزش کمک زیادی به کم کردن بایاس نمی‌کند. : درست**

افزایش تعداد داده‌های آموزش در صورتی ممکن است به کم کردن بایاس کمک کند که واریانس مدل بالا باشد. در واقع bias به این معنی‌است که مدل در حالت Underfit قرار دارد. این مشکل نیز با اضافه کردن داده حل نمی‌شود و احتمالا مدل بیش از حد ساده است و باید مدل را تغییر بدهیم.

- **کم کردن خطای مدل روی داده‌های آموزش منجر به کاهش خطای مدل روی داده‌های تست می‌شود. : نادرست**

ممکن است مدل ما در حالت Overfit قرار بگیرد. همان‌طور که در «سوال ۳» نیز به‌نوعی این موضوع را بررسی کردیم، ممکن است Overfit باعث بشود خطای مدل روی داده‌های آموزش بسیار کم باشد اما روی داده‌های تست عملکرد فاجعه‌باری داشته باشد. بنابراین با پیچیدگی زیاد مدل و کاهش خطای کارکرد آن روی داده‌های آموزش لزومی ندارد خطای آن روی داده‌های تست کاهش پیدا کند.

- **افزایش پیچیدگی مدل رگرسیون همواره منجر به کاهش خطای مدل روی داده‌ی آموزش و افزایش خطای مدل روی داده‌ی تست می‌شود. : نادرست**

اگر مدل بیش از حد ساده باشد و در حالت Underfit قرار داشته باشد، باید برای کاهش خطای آن روی داده‌های تست پیچیدگی آن را افزایش بدهیم و لزومی ندارد که خطای آن روی داده‌ی تست افزایش پیدا کند.

سوال ۵

الف)

می‌خواهیم درخت تصمیم‌گیری پیش‌بینی حمله‌ی قلبی را بسازیم. بنابراین ریشه‌ی درخت Attack Heart است. برای سادگی نوشتن علائم در رابطه‌ها، اختصارهای زیر را استفاده می‌کنیم:

$$\begin{cases} HeartAttack \rightarrow HA \\ Exercises \rightarrow E \\ Smokes \rightarrow S \\ Male \rightarrow M \\ ChestPain \rightarrow CP \end{cases}$$

برای محاسبه‌ی آنتروپی داریم:

$$H(X) = -\sum_{i=1}^{k} P(X = x_i) \log_2 P(X = x_i)$$

ابتدا آنتروپی Attack Heart را محاسبه می‌کنیم:

$$H(HA) = -\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6}) \approx 0.92$$

اکنون برای یافتن node Decision بعدی، برای تمام node ها، مقدار Gain Information یا IG را حساب می‌کنیم. برای محاسبه‌ی IG داریم:

$$IG(X) = H(Y) - H(Y|X)$$
$$H(Y|X) = -\sum_{j=1}^{v} P(X = x_j) \sum_{i=1}^{k} P(Y = y_i|X = x_j) \log_2 P(Y = y_i|X = x_j)$$

بنابراین با محاسبه‌ی IG ها در این مرحله داریم:

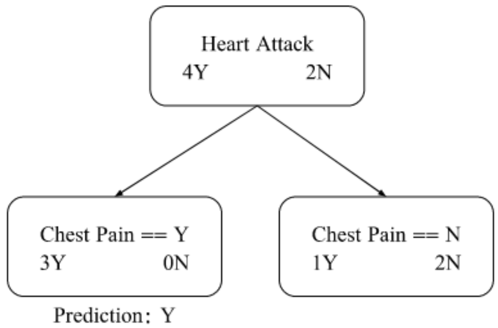
$$IG(E) = H(HA) - H(HA|E) = H(HA) + \frac{2}{3}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) + \frac{1}{3}(1\log_2(1) + 0\log_2(0)) \approx 0.92 - 0.66 = 0.26$$

$$IG(S) = H(HA) - H(HA|S) = H(HA) + \frac{2}{3}(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4})) + \frac{1}{3}(\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})) \approx 0.92 - 0.87 = 0.05$$

$$IG(M) = H(HA) - H(HA|M) = H(HA) + \frac{۲}{۳}(\frac{۱}{۲}\log_{۲}(\frac{۱}{۲}) + \frac{۱}{۲}\log_{۲}(\frac{۱}{۲})) + \frac{۱}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) \approx ۰/۹۲ - ۰/۶۶ = ۰/۲۶$$

$$IG(CP) = H(HA) - H(HA|CP) = H(HA) + \frac{۱}{۲}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) + \frac{۱}{۲}(\frac{۱}{۳}\log_{۲}(\frac{۱}{۳}) + \frac{۲}{۳}\log_{۲}(\frac{۲}{۳})) \approx ۰/۹۲ - ۰/۴۶ = ۰/۴۶$$

از آن‌جایی که فیچر Pain Chest بین بقیه‌ی فیچرها، IG ماکسیمم را دارد، آن را قرار می‌دهیم. تا به‌این‌جا شکل Tree Decision این‌گونه است:



اکنون برای پیدا کردن فیچر بعدی، میان نمونه‌هایی با مقدار $ChestPain == No$ مجددا IG ها را حساب می‌کنیم:

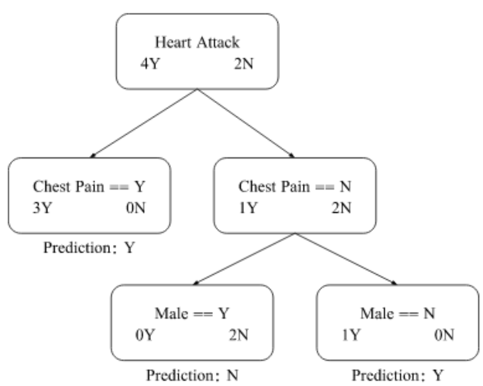
$$H(HA) = -\frac{۴}{۶}\log_{۲}(\frac{۴}{۶}) - \frac{۲}{۶}\log_{۲}(\frac{۲}{۶}) \approx ۰/۹۲$$

$$IG(E) = H(HA) - H(HA|E) = H(HA) + \frac{۲}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) + \frac{۱}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) \approx ۰/۹۲ - ۰/۰۰ = ۰/۹۲$$

$$IG(S) = H(HA) - H(HA|S) = H(HA) + \frac{۲}{۳}(\frac{۱}{۲}\log_{۲}(\frac{۱}{۲}) + \frac{۱}{۲}\log_{۲}(\frac{۱}{۲})) + \frac{۱}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) \approx ۰/۹۲ - ۰/۶۶ = ۰/۲۶$$

$$IG(M) = H(HA) - H(HA|M) = H(HA) + \frac{۲}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) + \frac{۱}{۳}(۱\log_{۲}(۱) + ۰\log_{۲}(۰)) \approx ۰/۹۲ - ۰/۰۰ = ۰/۹۲$$

بنابراین هر دو فیچر Male, Exercises می‌توانند فیچر بعدی ما باشند. ما فیچر Male را انتخاب می‌کنیم و در نهایت درخت تصمیم به شکل زیر می‌شود:



ب)

از ریشه به‌ترتیب شروع به نوشتن گزاره‌های تصمیم‌گیری می‌کنیم:

```
if ChestPain == True:
    HeartAttack = False
if ChestPain == False:
```

```
    if Exercises == True:
        HeartAttack = False
    if Exercises == False:
        HeartAttack = True
```

سوال ۶

طبق صورت سوال، فرض کنید فیچرها با لیبل‌های $\{x_1, ..., x_d\}$ باشند. در بدترین حالت درختی که تشکیل می‌دهیم یک درخت دودویی کامل است که در عمق iام تصمیم‌گیری بر اساس فیچر x_i انجام می‌شود. بنابراین عمق درخت دقیقا $d + 1$ خواهد بود. در عمقی که node ها لیبل x_d دارند، برای بچه‌هایشان می‌توان دقیقا مقدار خروجی را قرار داد. اکنون یک دسته‌بندی دلخواه h را در نظر بگیرید. برای هر n تایی مرتب از ۰ و ۱ها مانند d به‌طوری که $d \in D_h$ باشد، واضح است که d متناظر با یک مسیر از ریشه به یکی از برگ‌های درخت ساخته‌شده است. در عمق آخر نیز با توجه به مقدار $x_1, ..., x_d$ ، خروجی برگ متناظر با مسیر، خروجی تابع را تعیین می‌کند.