یادگیری ماشین

نیم‌سال اول ۰۱-۰۰

مدرس: دکتر عباس حسینی

دانشکده مهندسی کامپیوتر

تمرین چهارم

محمدجواد هزاره

شماره دانشجویی: ۹۸۱۰۱۰۷۴

سوال ۱

آ) بردار $\boldsymbol{\hat{n}}$ را بردار یکه در راستای متناظر با wای در نظر می‌گیریم که B را بدست می‌آورد. به عبارتی خواهیم داشت:

$$\forall i \in \{1, \cdots, m\} : y_i \langle B\boldsymbol{\hat{n}}, x_i \rangle \geq 1 \qquad (*)$$

فرض کنیم الگوریتم k بار اجرا می‌شود. هدف پیدا کردن کرانی برای k است. حال اگر در مرحله‌ی tام، داده‌ی jام اشتباه دسته‌بندی شده باشد، آنگاه $\boldsymbol{w^{t+1}} = \boldsymbol{w^t} + y_j \boldsymbol{x_j}$. بنابراین داریم:

$$\begin{aligned} \langle \boldsymbol{w^{t+1}}, B\boldsymbol{\hat{n}} \rangle &= B \langle \boldsymbol{w^t} + \boldsymbol{y_j x_j}, \boldsymbol{\hat{n}} \rangle \\ &= B \left(\langle \boldsymbol{w^t}, \boldsymbol{\hat{n}} \rangle + \langle y_j \boldsymbol{x_j}, \boldsymbol{\hat{n}} \rangle \right) \end{aligned}$$

$$\begin{aligned} \implies \langle \boldsymbol{w^{t+1}}, \boldsymbol{\hat{n}} \rangle &= \langle \boldsymbol{w^t}, \boldsymbol{\hat{n}} \rangle + \frac{1}{B} \langle y_j \boldsymbol{x_j}, B\boldsymbol{\hat{n}} \rangle \\ &\geq \langle \boldsymbol{w^t}, \boldsymbol{\hat{n}} \rangle + \frac{1}{B} \end{aligned}$$

که خط آخر با توجه به $(*)$ نتیجه شده است. حال با توجه به استقرا و اینکه $\boldsymbol{w^1} = \boldsymbol{0}$ ، برای هر t در بازه‌ی ۱ تا k خواهیم داشت: (تمامی نرم‌ها L_2 هستند.)

$$\langle \boldsymbol{w^{t+1}}, \boldsymbol{\hat{n}} \rangle \geq \frac{t}{B} \xRightarrow{\|\boldsymbol{\hat{n}}\|=1} \|\boldsymbol{w^{t+1}}\| \geq \frac{t}{B} \qquad (\star)$$

همچنین برای $\|\boldsymbol{w^{t+1}}\|$ داریم:

$$\begin{aligned} \|\boldsymbol{w^{t+1}}\|^2 &= \|\boldsymbol{w^t} + y_j \boldsymbol{x_j}\|^2 \\ &= \|\boldsymbol{w^t}\|^2 + y_j^2 \|\boldsymbol{x_j}\|^2 + 2\langle \boldsymbol{w^t}, y_j \boldsymbol{x_j} \rangle \quad (\diamond) \end{aligned}$$

و از آن‌جایی که فرض کرده بودیم داده‌ی jام اشتباه دسته‌بندی شده است، $\langle \boldsymbol{w^t}, y_j\boldsymbol{x_j}\rangle \leq 0$ خواهد بود و هم‌چنین مطابق تعریف الگوریتم برچسب‌ها یک یا منفی یک بودند که نتیجه می‌دهد $y_j^2 = 1$. مطابق تعریف R برای هر i، خواهیم داشت $\|\boldsymbol{x_i}\| \leq R$ ؛ با توجه به این موارد داریم:

$$(\diamond) \implies \|\boldsymbol{w^{t+1}}\|^2 \leq \|\boldsymbol{w^t}\|^2 + R^2$$

با استفاده از استقرا خواهیم داشت:

$$\|\boldsymbol{w^{t+1}}\|^2 \leq t\,R^2 \qquad (\star\star)$$

با کنار هم گذاشتن $(\star)$ و $(\star\star)$ و قرار دادن k به جای t خواهیم داشت:

$$\frac{k^2}{B^2} \leq \|\boldsymbol{w^{t+1}}\|^2 \leq k\,R^2 \implies \boxed{k \leq B^2R^2}$$

بنابراین حداکثر دفعات تکرار مراحل الگوریتم برابر B^2R^2 خواهد بود.

ب) کافیست مقیاس فضای ویژگی‌ها را در $\frac{1}{\eta}$ ضرب کنیم. داده‌های مسئله $(\tilde{\boldsymbol{x_i}}, y_i) = (\eta\boldsymbol{x_i}, y_i)$ خواهند بود و در مرحله‌ی آپدیت کردن وزن‌ها داریم $\boldsymbol{w^{t+1}} = \boldsymbol{w^t} + \eta y_i\boldsymbol{x_i} = \boldsymbol{w^t} + y_i\tilde{\boldsymbol{x_i}}$. بنابراین کافیست پارامترهای B و R را در فضای ویژگی‌های جدید که $\tilde{\boldsymbol{x_i}}$ است محاسبه کنیم. برای R در فضای جدید خواهیم داشت :

$$\tilde{R} = \max_i \|\tilde{\boldsymbol{x_i}}\| = \eta \max_i \|\boldsymbol{x_i}\| = \eta R$$

برای محاسبه‌ی $\tilde{B}$ داریم:

$$y_i\langle B\hat{\boldsymbol{n}}, \boldsymbol{x_i}\rangle \geq 1$$

$$y_i\langle B\hat{\boldsymbol{n}}, \eta\boldsymbol{x_i}\rangle \geq \eta$$

$$\frac{1}{\eta}\left(y_i\langle B\hat{\boldsymbol{n}}, \eta\boldsymbol{x_i}\rangle\right) \geq 1$$

$$y_i\langle \frac{B}{\eta}\hat{\boldsymbol{n}}, \eta\boldsymbol{x_i}\rangle \geq 1$$

بنابراین $\tilde{B} = \frac{B}{\eta}$ و اگر این‌گونه نباشد، یعنی بتوان $\tilde{B}$ دیگری یافت که در رابطه‌ی مورد نظر صدق کند و از مقدار داده شده کم‌تر باشد، فرض کم‌ترین بودن B نقض می‌شود. بنابراین مقدار ارائه شده کم‌ترین مقداری است که می‌توان برای $\tilde{B}$ پیدا کرد. بنابراین در این حالت مراحل الگوریتم حداکثر $\tilde{B}^2\tilde{R}^2 = \frac{B^2}{\eta^2}\eta^2R^2 = B^2R^2$ که مشابه همان حالت (آ) است تکرار خواهد شد.

سوال ۲

آ) ToDo

ب) همانطور که گفته شده، نشان می‌دهیم ضرب داخلی دو بردار صفر است. با توجه به روابطی که برای ضرایب داریم خواهیم داشت:

$$A = \sum_{i=1}^{m} y_i h_t(x_i) D_{t+1}(i)$$
$$= \sum_{i=1}^{m} \frac{y_i h_t(x_i) D_t(i)}{Z_t} \, e^{-\alpha_t y_i h_t(x_i)}$$

اگر $\mathcal{C}$ مجموعه‌ی اندیس داده‌هایی باشد که درست دسته‌بندی شده و $\mathcal{M}$ مجموعه‌ی اندیس داده‌هایی که غلط دسته‌بندی شده‌اند، آنگاه A را می‌توان به صورت زیر نوشت:

$$A = \frac{1}{Z_t} \left(\sum_{i \in \mathcal{C}} D_t(i) \, e^{-\alpha_t} + \sum_{j \in \mathcal{M}} -D_t(j) e^{\alpha_t} \right)$$

که باتوجه به اینکه $\alpha_t = \ln(\sqrt{\frac{1-\epsilon_t}{\epsilon_t}})$، ادامه می‌دهیم:

$$A = \frac{1}{Z_t} \left(\sum_{i \in \mathcal{C}} D_t(i) \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} - \sum_{j \in \mathcal{M}} D_t(j) \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \right)$$
$$= \frac{1}{Z_t \sqrt{\epsilon_t (1-\epsilon_t)}} \left(\sum_{i \in \mathcal{C}} \epsilon_t D_t(i) + \sum_{j \in \mathcal{M}} \epsilon_t D_t(j) - \sum_{j \in \mathcal{M}} D_t(j) \right)$$
$$= \frac{1}{Z_t \sqrt{\epsilon_t (1-\epsilon_t)}} \left(\epsilon_t \sum_{i=1}^{m} D_t(i) - \sum_{j \in \mathcal{M}} D_t(j) \right)$$

با توجه به اینکه ضرایب نرمالایز شده‌اند داریم $\sum_{i=1}^{m} D_t(i) = 1$، همچنین برای خطا داریم:

$$\epsilon_t = \sum_{i=1}^{m} D_t(i) \, \boldsymbol{I}(h(x_i) \neq y_i) = \sum_{i \in \mathcal{M}} D_t(i)$$

بنابراین:

$$A = \frac{1}{Z_t \sqrt{\epsilon_t (1-\epsilon_t)}} \left(\epsilon_t - \sum_{j \in \mathcal{M}} D_t(j) \right) = 0$$

بنابراین بردار ضرایب $\boldsymbol{D}_{t+1}$ بر بردار شامل $y_i h_t(x_i)$ عمود بوده یا به عبارتی uncorrelated هستند.

سوال ۳

آ) نرخ خطا را به صورت زیر بازنویسی می‌کنیم:

$$R(h) = \mathbb{P}\{Y \neq h(X)\} = 1 - \mathbb{P}\{Y = h(X)\}$$

بنابراین برای کمینه کردن خطا کافیست احتمال برابر شدن Y با $h(X)$ را بیشینه کنیم. این عبارت را نیز به صورت زیر می‌توان باز کرد که $D = \{x \in \mathrm{R}^n \,|\, h(x) = 0\}$ و $D' = \mathrm{R}^n - D$ و n بعد بردار ویژگی‌ها است.

$$\begin{aligned}\mathbb{P}\{Y = h(x)\} &= \mathbb{P}\{Y = 1, h(X) = 1\} + \mathbb{P}\{Y = 0, h(X) = 0\} \\ &= \mathbb{P}\{Y = 1, X \in D'\} + \mathbb{P}\{Y = 0, X \in D\} \\ &= \int_{D'} \mathbb{P}\{Y = 1 \,|\, X = x\} p_X(x)\, dx + \int_D \mathbb{P}\{Y = 0 \,|\, X = x\} p_X(x)\, dx \quad (*)\end{aligned}$$

که اگر تعریف کنیم $m(x) = \mathbb{P}\{Y = 1 \,|\, X = x\}$، آنگاه خواهیم داشت $\mathbb{P}\{Y = 0 \,|\, X = x\} = 1 - m(x)$. با جایگذاری این تعاریف در رابطه‌ی بالا ادامه می‌دهیم:

$$\begin{aligned}\mathbb{P}\{Y = h(x)\} &\overset{(*)}{=} \int_{D'} m(x) p_X(x)\, dx + \int_D \big(1 - m(x)\big) p_X(x)\, dx \\ &= \int_{D'} m(x) p_X(x)\, dx + \int_D \big(1 - m(x)\big) p_X(x)\, dx + \left(\int_D m(x) p_X(x)\, dx - \int_D m(x) p_X(x)\, dx\right) \\ &= \int_{\mathrm{R}^n} m(x) p_X(x)\, dx + \int_D \big(1 - 2m(x)\big) p_X(x)\, dx \\ &= C + \int_D \big(1 - 2m(x)\big) p_X(x)\, dx\end{aligned}$$

بنابراین فقط عبارت دوم در رابطه‌ی بالا به h بستگی دارد که اگر بخواهیم احتمال مورد نظر را بیشینه کنیم، باید X هایی عضو D باشند که عبارت زیر انتگرال برای آن‌ها نامنفی شود. بنابراین برای h^* داریم:

$$\left.\begin{aligned}\forall x \in D : h^*(x) = 0 \\ \forall x \in D : 1 - 2m(x) \geq 0\end{aligned}\right\} \implies h^*(x) = 0 \iff m(x) \leq \frac{1}{2}$$

بنابراین تابع h^* به صورت زیر خواهد بود:

$$h^*(x) = \begin{cases} 1 & m(x) > \frac{1}{2} \\ 0 & m(x) \leq \frac{1}{2} \end{cases}$$

ب) با توجه به قسمت (آ) برای خطای توابع داده شده داریم:

$$\begin{cases} R(h^*) = 1 - C - \int_D \big(1 - 2m(x)\big) p_X(x)\, dx & D = \{x \,|\, m(x) \leq \frac{1}{2}\} \\ R(\hat{h}) = 1 - C - \int_B \big(1 - 2m(x)\big) p_X(x)\, dx & B = \{x \,|\, \hat{m}(x) \leq \frac{1}{2}\} \end{cases}$$

بنابراین:

$$E = R(\hat{h}) - R(h^*) = \int_D \big(1 - 2m(x)\big) p_X(x)\, dx - \int_B \big(1 - 2m(x)\big) p_X(x)\, dx$$

با اضافه و کم کردن $2\hat{m}(x)$ به انتگرال‌ده‌ها ادامه می‌دهیم:

$$\begin{aligned} E &= \int_D \big(1 - 2m(x) + 2\hat{m}(x) - 2\hat{m}(x)\big) p_X(x)\, dx - \int_B \big(1 - 2m(x) + 2\hat{m}(x) - 2\hat{m}(x)\big) p_X(x)\, dx \\ &= \left(\int_D \big(1 - 2\hat{m}(x)\big) p_X(x)\, dx - \int_B \big(1 - 2\hat{m}(x)\big) p_X(x)\, dx \right) + 2 \int_D \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx \\ &\quad - 2 \int_B \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx \end{aligned}$$

حاصل داخل پرانتز در عبارت بالا منفی خواهد بود؛ چرا که B بهترین مجموعه‌ای است که می‌توان با توجه به $\hat{m}$ انتخاب کرد. اگر در استدلال‌های قسمت (آ) تابع m را با $\hat{m}$ جایگذاری کنیم، به این نتیجه می‌رسیم که B مجموعه‌ای است که حاصل انتگرال داخل پرانتز روی B را بیشینه می‌کند. بنابراین همین انتگرال روی هر مجموعه‌ی دیگری مانند D مقداری کم‌تر یا مساوی حاصل انتگرال روی B خواهد داشت. بنابراین:

$$E \leq 2 \int_D \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx - 2 \int_B \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx \quad (\star)$$

برای ادامه فرض کنیم $F = \{x \,|\, x \in D \wedge x \notin B\}$ و $G = \{x \,|\, x \in B \wedge x \notin D\}$. انتگرال‌های بالا به ازای xهای مشترک در B و D یکدیگر را خنثی می‌کنند و فقط xهایی که در F و G هستند باقی خواهند ماند. از

طرفی داریم:

$$\begin{cases} \forall x \in F : m(x) \leq \dfrac{1}{2},\ \hat{m}(x) > \dfrac{1}{2} & \implies |\hat{m}(x) - m(x)| = \hat{m}(x) - m(x) \\ \forall x \in G : m(x) > \dfrac{1}{2},\ \hat{m}(x) \leq \dfrac{1}{2} & \implies |\hat{m}(x) - m(x)| = -\big(\hat{m}(x) - m(x)\big) \end{cases}$$

بنابراین در ادامه‌ی $(\star)$ خواهیم داشت:

$$\begin{aligned} E &\leq 2\int_F \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx - 2\int_G \big(\hat{m}(x) - m(x)\big) p_X(x)\, dx \\ &\leq 2\int_F |\hat{m}(x) - m(x)|\, p_X(x)\, dx + 2\int_G |\hat{m}(x) - m(x)|\, p_X(x)\, dx \\ &\leq 2\int_{F\cup G} |\hat{m}(x) - m(x)|\, p_X(x)\, dx \\ &\leq 2\int_{R^n} |\hat{m}(x) - m(x)|\, p_X(x)\, dx \end{aligned}$$

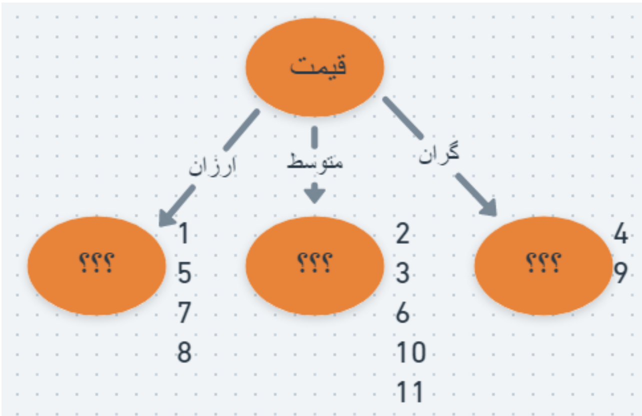
که در مرحله‌ی آخر از مثبت بودن انتگرال‌ده استفاده شده و مقادیری مثبت به حاصل اضافه شده است که تغییری در جهت نابرابری ایجاد نمی‌کند.

سوال ۴

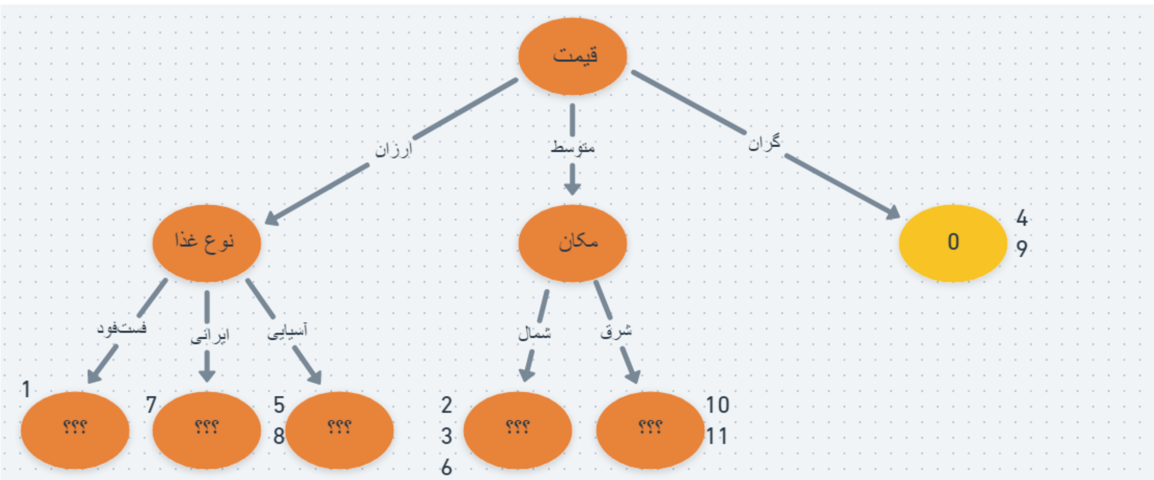
آ) برای ساخت درخت از هیوریستیک Information Gain استفاده می‌کنیم؛ به این صورت که در هر راس، ویژگی‌ای را انتخاب می‌کنیم که بین داده‌هایی که به آن راس رسیده‌اند این پارامتر برای آن بیشینه باشد. در مرحله‌ی اول و برای راس داریم: [۱]

$$IG(\text{نوع غذا}) \approx 0.052, \quad IG(\text{قیمت}) \approx 0.189, \quad IG(\text{محل}) \approx 0.016, \quad IG(\text{محدودیت}) \approx 0.016$$

بنابراین در ریشه ویژگی **قیمت** انتخاب خواهد شد و درخت به صورت زیر تقسیم‌بندی می‌شود:



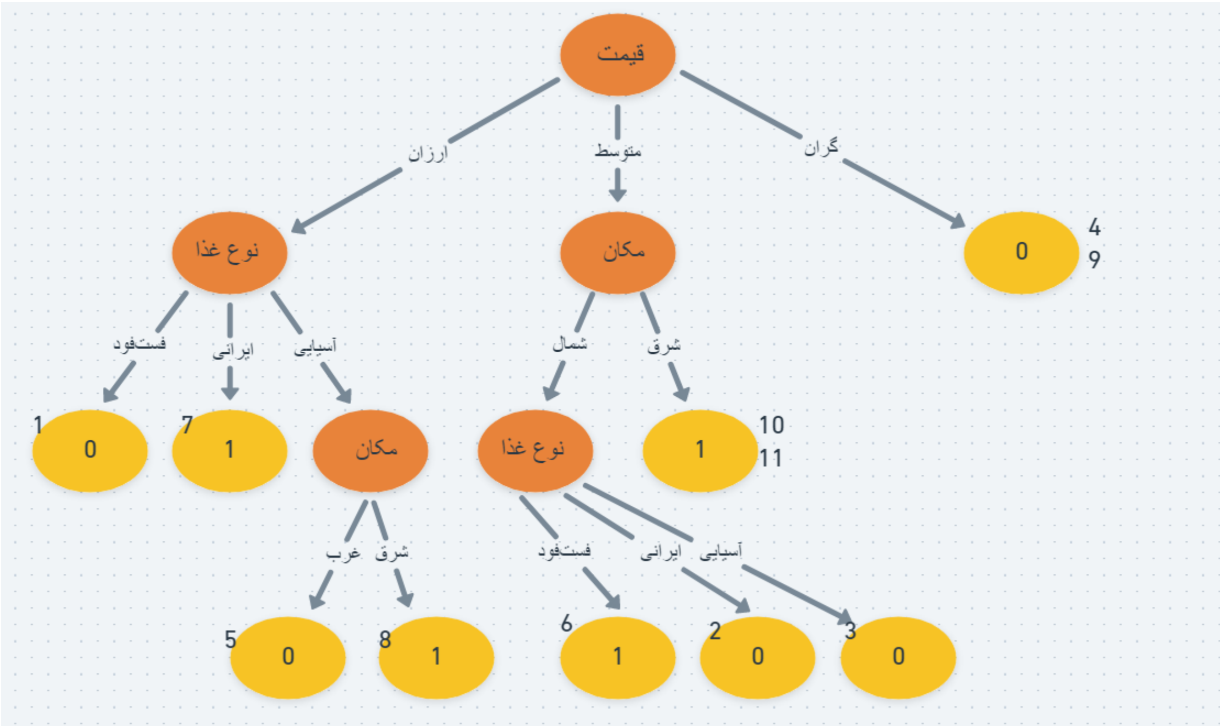
در مرحله‌ی بعد داده‌هایی که به سمت گران دسته‌بندی می‌شوند همه یک برچسب خورده‌اند بنابراین این راس به برگ تبدیل شده و مقدار برچسب 0 می‌گیرد. برای راس ارزان، داده‌ها IG برابری دارند بنابراین یکی از آن‌ها را به صورت تصادفی انتخاب کرده‌ایم که ویژگی **نوع غذا** بوده است. برای راس متوسط نیز IG ویژگی **مکان** بیش‌تر بوده و این ویژگی انتخاب شده است. درخت بعد از این مرحله به شکل زیر در می‌آید:



در این مرحله راس‌هایی که یک داده برای آن‌ها مانده برچسب همان داده را می‌گیرند. داده‌های ۱۰ و ۱۱ نیز

[۱] برای محاسبه‌ی Information Gain از ماشین حساب آنلاین استفاده شده است.

برچسب یکسانی دارند بنابراین راس آن‌ها نیز همان برچسب آن‌ها یعنی 1 را خواهد گرفت. برای مابقی راس‌ها از میان ویژگی‌های باقی مانده، آن ویژگی که IG بیش‌تری داشته باشد را انتخاب می‌کنیم. برای داده‌های ۵ و ۸، ویژگی‌های باقی‌مانده «مکان» و «محدودیت» هستند که هر دو نیز IG یکسانی دارند بنابراین به صورت تصادفی **مکان** انتخاب شده است. برای داده‌های ۲،۳ و ۶، **نوع غذا** بیش‌ترین IG را دارد. پس از این تقسیم‌بندی راس‌های باقی مانده یک داده خواهند داشت و در نتیجه به برگ تبدیل می‌شوند. درخت تصمیم در نهایت به صورت زیر در خواهد آمد:



ب) با استفاده از درختی که در قسمت (آ) بدست آوردیم داریم:

$$\begin{cases} \text{رضایت‌مندی}(12) = 0 \\ \text{رضایت‌مندی}(13) = 1 \\ \text{رضایت‌مندی}(14) = 1 \\ \text{رضایت‌مندی}(15) = 1 \\ \text{رضایت‌مندی}(16) = 1 \end{cases}$$

ج) برای بدست آوردن F_1Score نیاز به $recall$ و $precision$ داریم. برای این پارامترها نیز خواهیم داشت:

$$\begin{cases} recall = \dfrac{TP}{TP+FN} = \dfrac{2}{2+0} = 1 \\ precision = \dfrac{TP}{TP+FP} = \dfrac{2}{2+2} = 0.5 \end{cases}$$

بنابراین برای F_1Score داریم:

$$F_1\text{Score} = \frac{2 \times precision \times recall}{precision + recall} = \frac{2 \times 0.5 \times 1}{0.5+1} = \frac{2}{3}$$

سوال ۵

آ) با توجه به استقلال داده‌ها از یکدیگر، درست‌نمایی را می‌توان به صورت زیر نوشت:

$$\mathbb{P}\{\mathcal{D}\,|\,\pi_1,\cdots,\pi_K\} = \prod_{i=1}^{N} \mathbb{P}\{\mathcal{D}_i\,|\,\pi_1,\cdots,\pi_K\} \quad (*)$$

که با توجه به این‌که $\mathcal{D} = \{\phi_n, t_n\}_{i=1}^{N}$ ، برای احتمال دیدن یک داده خواهیم داشت:

$$\mathbb{P}\{t_i\,|\,\pi_1,\cdots,\pi_K\} = \pi_j, \quad (t_i)_j = 1 \quad (\star)$$

بنابراین احتمال دیدن هر داده برابر با خواهد بود با π_k که k کلاسی‌ست که به آن تعلق دارد. حال اگر N_i تعداد داده‌هایی باشد که به کلاس iام تعلق دارند، با توجه به $(*)$ و $(\star)$ برای احتمال دیدن $\mathcal{D}$ خواهیم داشت:

$$\mathbb{P}\{\mathcal{D}\,|\,\pi_1,\cdots,\pi_K\} = \prod_{i=1}^{K} \pi_i^{N_i}$$

برای پیدا کردن تخمین‌گر π_i از لگاریتم تابع درست‌نمایی استفاده می‌کنیم. بنابراین داریم:

$$\ln\left(\mathbb{P}\{\mathcal{D}\,|\,\pi_1,\cdots,\pi_K\}\right) = \sum_{i=1}^{K} N_j \ln(\pi_j)$$

که N_j تعداد داده‌هایی‌ست که در کلاس jام دسته‌بندی شده‌اند. برای پیداکردن π_iهایی که عبارت بالا را بیشینه می‌کنند، شرط دیگری نیز داریم و آن هم یک شدن جمع همه‌ی آن‌ها یا به عبارتی $\sum_{j=1}^{K} \pi_j = 1$ است. برای پیدا کردن بیشینه از روش ضرایب لاگرانژ استفاده می‌کنیم. تابع لاگرانژ به صورت زیر خواهد بود:

$$L(\pi_1,\cdots,\pi_K,\lambda) = \sum_{j=1}^{K} N_j \ln(\pi_j) - \lambda\Big(\sum_{j=1}^{K} \pi_j - 1\Big)$$

بنابراین برای مشتقات داریم:

$$\begin{cases} \dfrac{\partial L}{\partial \pi_j} = \dfrac{N_j}{\pi_j} - \lambda \\ \dfrac{\partial L}{\partial \lambda} = -\displaystyle\sum_{j=1}^{K} \pi_j + 1 \end{cases}$$

معادلات بالا را برابر صفر قرار داده و با بازنویسی رابطه‌ی اول خواهیم داشت:

$$N_j = \lambda \pi_j^* \xRightarrow{\sum_{j=0}^{K}} \lambda = N$$
$$\Longrightarrow \boxed{\pi_j^* = \frac{N_j}{N}}$$

ب) فرض کنیم $\mathcal{C}_i$ مجموعه‌ی شامل اندیس داده‌هایی باشد که کلاس مربوط به آن‌ها i است. هم‌چنین $\boldsymbol{\mu}$ بردار شامل μ_kها بوده و $\boldsymbol{\pi}$ بردار شامل π_kها باشد. با توجه به استقلال داده‌ها برای درست‌نمایی خواهیم داشت:

$$\mathbb{P}\{\mathcal{D} \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \Phi\} = \prod_{i=1}^{N} \mathbb{P}\{t_i \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \phi_i\}$$

که با توجه به تعریف $\mathcal{C}_i$ها خواهیم داشت:

$$\mathbb{P}\{\mathcal{D} \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \Phi\} = \prod_{j=1}^{K} \left(\prod_{i \in \mathcal{C}_j} \mathbb{P}\{\mathcal{C}_j \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \phi_i\} \right)$$

که احتمال داخل پرانتز بالا را نیز می‌توان به صورت زیر نوشت که d بعد ویژگی‌هاست:

$$\begin{aligned}
\mathbb{P}\{\mathcal{C}_j \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \phi_i\} &= \mathbb{P}(\mathrm{C}_j)\mathbb{P}\{\phi_i \mid \mathcal{C}_j, \boldsymbol{\mu}, \Sigma\} \\
&= \pi_j (2\pi)^{-\frac{d}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\phi_i - \mu_j)^T \Sigma^{-1} (\phi_i - \mu_j) \right)
\end{aligned}$$

از بیشینه کردن لگاریتم درست‌نمایی استفاده می‌کنیم:

$$\begin{aligned}
f(\boldsymbol{\mu}, \Sigma) &= \ln\left(\mathbb{P}\{\mathcal{D} \mid \boldsymbol{\mu}, \Sigma, \boldsymbol{\pi}, \Phi\}\right) \\
&= \sum_{j=1}^{k} N_j \ln(\pi_j) - \frac{Nd}{2}\ln(2\pi) - \frac{N}{2}\ln(|\Sigma|) - \frac{1}{2}\sum_{j=1}^{K} \left(\sum_{i \in \mathcal{C}_j} (\phi_i - \mu_j)^T \Sigma^{-1} (\phi_i - \mu_j) \right)
\end{aligned}$$

برای مشتق تابع بالا نسبت به $\boldsymbol{\mu}$ و Σ خواهیم داشت:

$$\begin{cases}
\dfrac{\partial f}{\mu_j} = -\dfrac{1}{2} \displaystyle\sum_{i \in \mathcal{C}_j} 2\Sigma^{-1}(\phi_i - \mu_j) \\
\dfrac{\partial f}{\partial \Sigma} = -\dfrac{N}{2}\Sigma^{-1} + \dfrac{1}{2} \displaystyle\sum_{j=1}^{K} \left(\sum_{i \in \mathcal{C}_j} (\phi_i - \mu_j)(\phi_i - \mu_j)^T \Sigma^{-1} \Sigma^{-1} \right)
\end{cases}$$

با صفر قرار دادن معادلات بالا برای μ_j^* و Σ^* خواهیم داشت:

$$N_j \mu_j^* = \sum_{i \in \mathcal{C}_j} \phi_i \implies \boxed{\mu_j^* = \frac{\sum_{i \in \mathcal{C}_j} \phi_i}{N_j}}$$

$$\boxed{\Sigma^* = \frac{\sum_{j=1}^{K} \left(\sum_{i \in \mathcal{C}_j} (\phi_i - \mu_j)(\phi_i - \mu_j)^T \right)}{N}}$$