به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

پاسخ تمرین سری اول داده کاوی

نیم‌سال اول ۰۰-۹۹

سوال ۱-

الف) پیوسته، کمی- نسبت

ب) پیوسته، کمی- بازه

ج) پیوسته، کمی – نسبت

د) گسسته، کیفی – ترتیبی

ح) پیوسته، کمی - نسبت

ت) گسسته، کیفی – ترتیبی

سوال ۲-

الف) خیر، در تعریف noise به این نکته اشاره شده است که بر اثر خطای اندازه گیری حاصل شده است. چیزی که خطا باشد مطلوب نیست.

outlier بله، آنها می توانند داده های مهمی باشند و گاه هدف اصلی داده کاوی هستند.

ب) بله، ممکن است noise تاثیر شدیدی بر بخشی از داده ها بگذارد به گونه ای که ویژگی های آن داده ها نسبت به بقیه داده ها تفاوت زیادی بکند و آن بخش از داده ها به عنوان outlier شناسایی شوند .

ج) خیر، زیرا مثلا توزیع تصادفی نیز می تواند در برخی موارد به مقادیری مانند مقادیر نرمال بیانجامد.

د) خیر، outlier ها می توانند داد ه های مفیدی باشند که تنها بنظر میرسد به مجموعه داد ه ها تعلق ندارند و لزوما نویز نیستند. برای مثال آی-کیو استیون-هاوکینگ ۱۶۰ است که مقدار آن صحیح اما دور از مقادیر نرمال آن یعنی حدود ۱۱۰ است.

سوال ۳-

الف) x=(0, -1, 0, 1), y=(1, 0, -1, 0) معیار های cosine, correlation, Euclidean

Euclidean:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2} = \sqrt{(0-1)^2 + (-1-0)^2 + \big(0-(-1)\big)^2 + (1-0)^2} = \sqrt{4} = 2$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{4}(0-1+0+1) = 0$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{4}(1+0-1+0) = 0$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2} = \sqrt{\frac{1}{3}((0-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{\frac{2}{3}}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2} = \sqrt{\frac{1}{3}((1-0)^2 + (0-0)^2 + (-1-0)^2 + (0-0)^2)} = \sqrt{\frac{2}{3}}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{3}\big((0-0)(1-0) + (-1-0)(0-0) + (0-0)(-1-0) + (1-0)(0-0)\big) = 0$$

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{0}{\sqrt{\frac{2}{3}} \times \sqrt{\frac{2}{3}}} = 0$$

Cosine:

$$< x,y > = 0*1 + (-1)*0 + 0*(-1) + 1*0 = 0$$

$$||x|| = \sqrt{0*0 + (-1)*(-1) + 0*0 + 1*1} = \sqrt{2}$$

$$||y|| = \sqrt{1*1 + 0*0 + (-1)*(-1) + 0*0} = \sqrt{2}$$

$$\cos(x,y) = \frac{< x,y >}{||x||\,||y||} = \frac{0}{\sqrt{2} \times \sqrt{2}} = 0$$

ب) x=(2, -1, 0, 2, 0, -3), y=(-1, 1, -1, 0, 0, -1) معیار های cosine, correlation, Euclidean

Euclidean:

$$d(x,y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

$$= \sqrt{(2-(-1))^2 + (-1-1)^2 + \big(0-(-1)\big)^2 + (2-0)^2 + (0-0)^2 + (-3-(-1))^2}$$

$$= \sqrt{9+4+1+4+0+4} = \sqrt{22}$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{6}(2-1+0+2+0-3) = 0$$

$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{6}(-1+1-1+0+0-1) = 0$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$

$$= \sqrt{\frac{1}{5}((2-0)^2 + (-1-0)^2 + (0-0)^2 + (2-0)^2 + (0-0)^2 + (-3-0)^2} = \frac{3\sqrt{10}}{5}$$

$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$

$$= \sqrt{\frac{1}{5}((-1-0)^2 + (1-0)^2 + (-1-0)^2 + (0-0)^2 + (0-0)^2 + (-1-0)^2)}$$

$$= \frac{2\sqrt{5}}{5}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})(y_k - \bar{y})$$

$$= \frac{1}{5}\big((2-0)(-1-0) + (-1-0)(1-0) + (0-0)(-1-0) + (2-0)(0-0)$$

$$+ (0-0)(0-0) + (-3-0)(-1-0)\big) = \frac{1}{5}(-2-1+0+0+0+3) = 0$$

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{0}{\frac{3\sqrt{10}}{5} \times \frac{2\sqrt{5}}{5}} = 0$$

Cosine:

$$< x, y > = 2 * (-1) + (-1) * 1 + 0 * (-1) + 2 * 0 + 0 * 0 + (-3) * (-1) = -2 - 1 + 3 = 0$$
$$||x|| = \sqrt{2 * 2 + (-1) * (-1) + 0 * 0 + 2 * 2 + 0 * 0 + (-3) * (-3)} = \sqrt{4 + 1 + 0 + 4 + 0 + 9}$$
$$= 3\sqrt{2}$$
$$||y|| = \sqrt{(-1) * (-1) + 1 * 1 + (-1) * (-1) + 0 * 0 + 0 * 0 + (-1) * (-1)} = 2$$
$$\cos(x, y) = \frac{< x, y >}{||x||||y||} = \frac{0}{3\sqrt{2} \times 2} = 0$$

ج) x=(1, 1, 0, 1, 0, 1), y=(1, 1, 1, 0, 0, 1) معیار های cosine, correlation, Jaccard

Jaccard:

$$J = \frac{f_{11}}{f_{01} + f_{01} + f_{11}} = \frac{3}{1 + 1 + 3} = \frac{3}{5}$$

Correlation:

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = \frac{1}{6}(1 + 1 + 0 + 1 + 0 + 1) = \frac{2}{3}$$
$$\bar{y} = \frac{1}{n}\sum_{k=1}^{n} y_k = \frac{1}{6}(1 + 1 + 1 + 0 + 0 + 1) = \frac{2}{3}$$
$$s_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \bar{x})^2}$$
$$= \sqrt{\frac{1}{5}\left(\left(1 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2 + \left(0 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2 + \left(0 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2\right)}$$
$$= \sqrt{\frac{1}{5}\left(\frac{1}{9} + \frac{1}{9} + \frac{4}{9} + \frac{1}{9} + \frac{4}{9} + \frac{1}{9}\right)} = \frac{2\sqrt{15}}{15}$$
$$s_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \bar{y})^2}$$
$$= \sqrt{\frac{1}{5}\left(\left(1 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2 + \left(0 - \frac{2}{3}\right)^2 + \left(0 - \frac{2}{3}\right)^2 + \left(1 - \frac{2}{3}\right)^2\right)} =$$

$$= \sqrt{\frac{1}{5}\left(\frac{1}{9}+\frac{1}{9}+\frac{1}{9}+\frac{4}{9}+\frac{4}{9}+\frac{1}{9}\right)} = \frac{2\sqrt{15}}{15}$$

$$s_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})$$

$$= \frac{1}{5}\left(\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)+\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)+\left(0-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)+\left(1-\frac{2}{3}\right)\left(0-\frac{2}{3}\right)\right.$$
$$\left.+\left(0-\frac{2}{3}\right)\left(0-\frac{2}{3}\right)+\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)\right) = \frac{1}{5}\left(\frac{1}{9}+\frac{1}{9}-\frac{2}{9}-\frac{2}{9}+\frac{4}{9}+\frac{1}{9}\right) = \frac{1}{15}$$

$$corr(x,y) = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{15}}{\frac{2\sqrt{15}}{15}\times\frac{2\sqrt{15}}{15}} = \frac{1}{4}$$

Cosine:

$$< x,y > = 1*1+1*1+0*1+1*0+0*0+1*1 = 3$$
$$||x|| = \sqrt{1*1+1*1+0*0+1*1+0*0+1*1} = 2$$
$$||y|| = \sqrt{1*1+1*1+1*1+0*0+0*0+1*1} = 2$$
$$\cos(x,y) = \frac{< x,y >}{||x||\,||y||} = \frac{3}{2\times 2} = \frac{3}{4}$$

سوال۴-

الف)

۱. تخمین missing value ها با توجه به سایر داده ها.
۲. حذف آن ها در صورتی که تعداد قابل چشم پوشی باشد.

ب)

۱. کاهش ابعاد با الگوریتم هایی همچون PCA
۲. حذف ویژگی های غیرمرتبط با عملیات هدف.
۳. حذف ویژگی هایی که به نوعی به هم وابستگی دارند و با داشتن یکی می توان بقیه را حدس زد.

ج)

۱. نمونه برداری و جمع آوری مجدد داده های با برچسب بخصوص.
۲. وزن دهی متفاوت به هر یک از دسته ها هنگام انجام عملیات آموزش.

د)

۱. ساده سازی مدل یادگیری مورد استفاده با کاهش پارامتر یا مرتبه
۲. کم کردن زمان یادگیری مدل آموزشی

۳. استفاده از روش های Ensembling

ه)

۱. استفاده از یک مدل یادگیری پیچیده تر یا بزرگ تر
۲. افزایش تعداد گام ها و زمان آموزش
۳. جمع آوری داده های آموزشی بیشتر

سوال۵-

الف) outlier ها می توانند تاثیر بالایی داشته باشند. از آنجا که تابع هزینه $||\boldsymbol{X}\beta - y||_2^2$ بوده و هدف یافتن $\boldsymbol{\beta}$ هایی است که این تابع خطا را کمینه کنند داده های پرت به راحتی توان دوم فاصله بین نقاط و تابع پیشنهادی را بالا می برند.

ب) تابع کمترین مربعات خطاها یا MSE در نظر گرفته می شود زیرا میانگین کل نقاط در نظر گرفته میشود و می توان معیار درستی از فاصله تمامی نمونه داده ها از خط مورد نظر تخمین زد. علاوه بر آن مشتق پذیر بودن این تابع از نکات مورد توجه می باشد.

ج)

$$X\beta = y \quad , \quad y = \beta_0 + \beta_1 x$$

$$X = \begin{bmatrix} 1 & 1.58 \\ 1 & 1.6 \\ 1 & 1.62 \\ 1 & 1.65 \\ 1 & 1.68 \\ 1 & 1.7 \\ 1 & 1.74 \\ 1 & 1.75 \\ 1 & 1.77 \\ 1 & 1.8 \end{bmatrix} \qquad y = \begin{bmatrix} 57.5 \\ 58.2 \\ 59.5 \\ 62.1 \\ 63.4 \\ 64.5 \\ 66.2 \\ 67.7 \\ 69.4 \\ 71.3 \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

صف: min توابع مربع $\Rightarrow \quad \beta = (X^T X)^{-1} X^T y$

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1.58 & 1.6 & \cdots & 1.8 \end{bmatrix} \begin{bmatrix} 1 & 1.58 \\ 1 & 1.6 \\ \vdots & \vdots \\ 1 & 1.8 \end{bmatrix} = \begin{bmatrix} 10 & 16.89 \\ 16.89 & 28.57 \end{bmatrix}$$

$$\Rightarrow (X^T X)^{-1} = \frac{1}{0.5149} \begin{bmatrix} 28.57 & -16.89 \\ -16.89 & 10 \end{bmatrix} = \begin{bmatrix} 55.48 & -32.80 \\ -32.80 & 19.42 \end{bmatrix}$$

$$X^T y = \begin{bmatrix} 639.8 \\ 1083.824 \end{bmatrix} \Rightarrow \beta = \begin{bmatrix} 55.48 & -32.8 \\ -32.8 & 19.42 \end{bmatrix} \begin{bmatrix} 639.8 \\ 1083.824 \end{bmatrix} = \begin{bmatrix} -53.3 \\ 62.42 \end{bmatrix}$$

ک ضرایب خط معادله

سوال ۶-

$$H(X) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

طبقه اقتصادی:

$$-\frac{70,000}{200,000}\log_2\frac{70,000}{200,000} \quad -\frac{55000}{200,000}\log_2\frac{55000}{200,000} \quad -\frac{75,000}{200,000}\log_2\frac{75,000}{200,000}$$

فقیر — متوسط — مرفه

$$+0.53 \quad +0.51 \quad +0.53 \quad = 1.57$$

خرید مسکن دارد:

$$-\frac{1}{2}\log\frac{1}{2} \quad -\frac{1}{2}\log\frac{1}{2} \quad = 1$$

$$-\frac{20,000}{200,000}\log\frac{20,000}{200,000} = +0.33$$

$$-\frac{35,000}{200,000}\log\frac{35,000}{200,000} = +0.44$$

$$-\frac{45,000}{200,000}\log\frac{45,000}{200,000} = +0.48$$

$$-\frac{50,000}{200,000}\log\frac{50,000}{200,000} = 0.5$$

$$-\frac{20,000}{200,000}\log\frac{20,000}{200,000} = +0.33$$

$$-\frac{30,000}{200,000}\log\frac{30,000}{200,000} = +0.41$$

overal = 2.49

Mutual I = H(طبقه) + H(خرید) − H(خرید، طبقه)

$= 1.57 + 1 - 2.49 = 0.08$

ج) چون MI بیشتر صفر است، پس بین دو متغیر خرید و طبقه اقتصادی از هم مستقل اند.

سوال ۷-

الف) normalization: نگاشت داده‌ها به اعداد جدید است و در مواردی که بازه‌ی ویژگی‌ها تفاوت چشمگیری دارد استفاده می‌شود. (به طور مثال بازه‌ی یک ویژگی ۰ تا ۱ باشد و بازه‌ی ویژگی دیگر ۱۰ تا ۱۰۰۰ باشد)

discretization: برای تبدیل داده‌های پیوسته به مقادیر گسسته کاربرد دارد و در مواقعی که دسته‌های مشخصی در بازه‌ی یک ویژگی برایمان مهم هستند از آن استفاده می‌کنیم.

feature creation and feature selection: در feature selection تعداد مشخصی ویژگی که ارزش کاویدن دارند را انتخاب می‌کنیم. برای کاهش بعد استفاده می‌شود و می‌توان برای کم کردن ویژگی‌هایی که به هم وابسته هستند از آن استفاده کرد. در feature creation هدف ایجاد ویژگی‌های جدیدی است که اطلاعات جدید و ارزشمند بتوانند به داده‌ها اضافه کنند.

Sampling: نمونه‌برداری از داده برای پردازش‌های بعدی است و در مواقعی که حجم داده بسیار زیاد و یا توان پردازشی محدود است از آن می‌توان استفاده کرد.

Aggregation: ترکیب چند ویژگی برای تبدیل به یک ویژگی جدید. می‌توان برای کاهش ویژگی‌هایی که correlation بالایی با هم دارند از آن استفاده کرد.

ب)

$$v_i' = \frac{v_i - min}{max - min}(new\_max - new\_min) + new\_min \quad : min\_max$$

$$v_{200} = \frac{200-200}{1000-200} \times 1 + 0 = 0$$

$$v_{300} = \frac{300-200}{800} = \frac{1}{8} = 0.125$$

$$v_{400} = \frac{200}{800} = \frac{1}{4} = 0.250$$

$$v_{600} = \frac{400}{800} = \frac{1}{2} = 0.5$$

$$v_{1000} = \frac{800}{800} = 1$$

$$v_i' = \frac{v_i - \mu}{6} \qquad \mu = 500 \qquad : Z\_score$$

$$6 = 282.84$$

$$v_{200} = \frac{-300}{282.84} = -1.06 \qquad v_{1000} = \frac{500}{282.84} = 1.76$$

$$v_{300} = \frac{-200}{282.84} = -0.70$$

$$v_{400} = \frac{-100}{282,84} = -0.35$$

$$v_{600} = \frac{100}{282,84} = 0.35$$