نمونه سوال درس داده‌کاوی

فصل ۱

۱. سوالات ۱ تا ۳ فصل اول کتاب لاروس

۲. داده‌کاوی چیست؟ مراحل فرایند کشف دانش را شرح دهید. (پاسخ: اسلایدهای ۸ و ۱۰ فصل اول)

۳. شش فاز فرایند استاندارد CRISP-DM را همراه با مثال شرح دهید.

۴. مشخص کنید کدامیک از موارد زیر مربوط به یک مسئله کلاس‌بندی، تخمین، پیشگویی، خوشه‌بندی یا کشف الگو (قواعد انجمنی) است. (اگر به بیش از یک مسئله ممکن است مربوط شود، توضیح دهید.)

- Determining the proportion of cases in which a new drug will exhibit dangerous side effects
- Segmentation of financial behaviors into safe and suspicious categories based on the similarities of them
- Dimension-reduction of a dataset that has hundreds of attributes
- Grouping very large quantities of genes that exhibit similar behavior
- Determining whether a particular credit card transaction is fraudulent
- Placing a new student into a particular track with regard to special needs
- The price of a stock three months into the future
- Grouping the clothes in a store based on the size
- The winner of this fall’s baseball World Series, based on a comparison of team statistics
- The percentage increase in traffic deaths next year if the speed limit is increased
- The number of points per game that Patrick Ewing will score.
- Assessing whether a mortgage application is a good or bad credit risk
- Diagnosing whether a particular disease is present
- The amount of money a randomly chosen family of four will spend for back-to-school shopping this fall.
- Finding out which items in a supermarket are purchased together and which items are never purchased together

- Determining whether a will was written by the actual deceased, or fraudulently by someone else
- Whether a particular molecule in drug discovery will lead to a profitable new drug for a pharmaceutical company
- The grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA.
- Examining the proportion of children whose parents read to them who are themselves good readers
- Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat
- Investigating the proportion of subscribers to a company's cell phone plan that respond positively to an offer of a service upgrade
- Finding degradation in telecommunications networks

۵. تفاوت خوشه‌بندی و کلاس‌بندی در چیست؟ شرح دهید و چند الگوریتم نام ببرید.

۶. تفاوت تخمین و کلاس‌بندی در چیست؟ شرح دهید و چند الگوریتم نام ببرید.

۷. تفاوت پیشگویی و تخمین در چیست؟ شرح دهید و چند الگوریتم نام ببرید.

فصل دوم:

۱. سوالات ۱، ۳، ۵ و ۶ فصل دوم کتاب لاروس
۲. روش‌های مدیریت داده‌های از بین رفته را نام ببرید. (پاسخ: اسلاید ۵ فصل دوم) توضیح دهید هر کدام چه مزیت و عیبی دارند.
۳. سه روش کشف داده‌های پرت را نام ببرید. (پاسخ: اسلاید ۱۵ فصل دوم)

فصل سوم:

۱. سوالات ۱ تا ۳ و ۷ فصل سوم کتاب لاروس
۲. نمایش side-by-side box plot، برای تحلیل چه نوع داده‌ای قابل استفاده است؟ عددی یا گسسته؟ شرح دهید. (پاسخ: اسلایدهای ۲۴ و ۲۵ فصل سوم)

فصل چهارم:

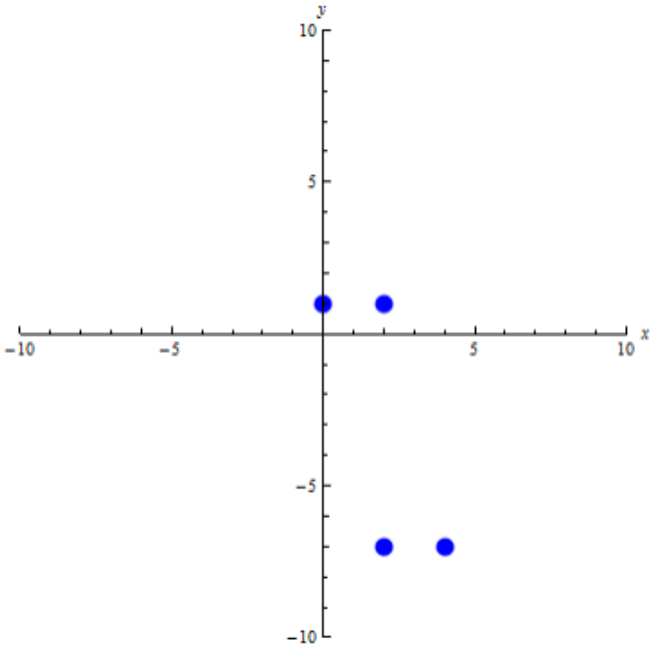
۱. سوالات ۱، ۲، ۵، ۶، ۷، ۸ و ۹ فصل چهارم کتاب لاروس

۲. منظور از معیار پراکندگی چیست؟ چند مورد نام ببرید. (پاسخ: اسلاید ۴ و ۵ فصل چهارم)

۳. منظور از برون‌یابی (extrapolation) چیست؟ مخاطرات آن را شرح دهید. (پاسخ: اسلاید ۱۹ فصل چهارم)

۴. مخاطرات وجود ویژگی‌های همبسته (multicollinearity) را شرح دهید. چه راه حلی را برای برطرف شدن این مشکل پیشنهاد می‌کنید؟ (پاسخ: اسلاید ۲۴ فصل چهارم)

۵. نقاط (0,1), (2,1), (2,-7), (4,-7) را در نظر بگیرید (نمودار زیر).



الف) توضیح دهید خروجی رگراسیون خطی برای این نقاط، چه خطی خواهد بود؟

ب) معادله تقریبی این خط را بدست آورید.

ج) مقدار میانگین مربعات خطا (MSE) و خطای استاندارد را برای این رگراسیون محاسبه کنید. یادآوری:

$$\mathrm{MSE} = \frac{\sum_i (y_i - \hat{y}_i)^2}{n} \qquad s = \sqrt{\mathrm{MSE}}$$

ج) برای یک مختصات جدید x=5، مقدار y، چه مقدار تخمین زده می‌شود؟

فصل پنجم:

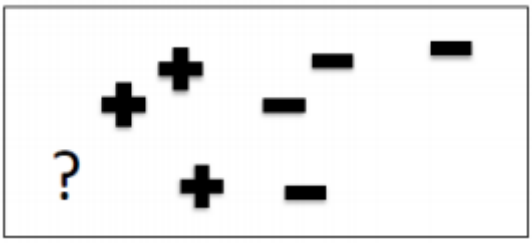
۱. سوالات ۱ تا ۱۲ ، کتاب لاروس

۲. چه زمانی در آموزش، بیش‌پوشش (overfitting) رخ می‌دهد؟ چه روشی را برای مقابله با آن پیشنهاد می‌کنید؟ (پاسخ: اسلایدهای ۴ و ۵ فصل پنجم)

۳. منظور از معیار فاصله چیست؟ بهتر است معیار فاصله چه ویژگی‌هایی داشته باشد؟ برای هر کدام از داده‌های عددی و گسسته، یک معیار فاصله مثال بزنید. (پاسخ: اسلاید ۱۶ و ۱۷ فصل پنجم)

۴. توضیح دهید اگر داده‌ها نرمالایز نشده باشند، چه مخاطراتی در محاسبه معیار فاصله خواهند داشت؟ (پاسخ: اسلاید ۱۸ فصل پنجم)

۵. داده‌های زیر را در نظر بگیرید. برای چه مقدار k، الگوریتم K-NN، نمونه‌ی ? را در کلاس – دسته‌بندی می‌کند؟



فصل یازدهم:

۱. در مورد نحوه‌ی ارزیابی تحلیل‌های تکاملی (evolutionary data analysis) شرح دهید. (پاسخ: اسلاید ۴ فصل یازدهم)

۲. در مورد نحوه‌ی ارزیابی مسئله‌های پیش‌بینی و تخمین، شرح دهید و چند معیار معرفی کنید. (پاسخ: اسلاید ۵ فصل یازدهم)

۳. داده‌های آموزشی جدول زیر را در نظر بگیرید.

instance	x	y	Class
1	2	4	-
2	3	1	+
3	3	2	+
4	0	4	-
5	2	4	-
6	2	1	+

الف) داده‌های تست زیر را با کمک روش k-NN یکبار برای k=1 و یکبار برای k=3 کلاس‌بندی کنید.

instance	x	y	Predicted 1-NN	Predicted 3-NN
6	1	3		
7	2	3		
8	2	2		
9	3	3		
10	3	2		

حال فرض کنید خروجی ایده‌آل نمونه‌های ۶،۷ و ۹ منفی و نمونه‌های ۸ و ۱۰ مثبت هستند.

ب) ماتریس ابهام (confusion matrix) هر دو حالت کلاس‌بندی فوق را ترسیم نمایید.

ج) مقدار error rate، false postive و false negative را در هر دو حالت کلاس‌بندی قسمت (الف) محاسبه کنید و نتایج را با هم مقایسه نمایید.

د) مقدار lift را برای هر دو ماتریس ابهام محاسبه نمایید. کدام روش از نظر lift بهتر است؟

موفق و موید باشید - ربیعی