

1 - فرض کنید که مجموعه داده فروش محصولات یک فروشگاه به شکل زیر دخیره شده است و قصد طراحی یک سیستم برای پیش‌بینی تعداد فروش آیتم‌های جدید براساس سایر ویژگی‌های آن‌ها را داریم.

ProdID	Name	Desc	Price	Category	Units Sold
13	Errorplane	"A truly uncaught exception ..."	404.00	Toy	9
42	Rock Kit	"Launch into minerology with ..."	123.45	Toys	1
54	Punative Jokes	"Jokes that will get you fined ..."	1.00	Books	30

...

الف- فرض کنید که به دنبال استفاده از یک سیستم خطی برای حل این مسئله هستیم. برای هر کدام از ویژگی‌های موجود چه تکنیک‌های مهندسی ویژگی را پیشنهاد می‌دهید و چرا؟

ب- فرض کنید که رابطه‌ی بین قیمت و تعداد آیتم‌های فروخته شده برای هر دسته (category) از آیتم‌ها متفاوت باشد. برای این منظور کدامیک از ویژگی‌های ترکیبی زیر مناسب‌تر است و چرا؟

- A.  $\phi(\text{category}, \text{price}) = \text{category} + \text{price}$
- B.  $\phi(\text{category}, \text{price}) = \text{price} \times \text{category}$
- C.  $\phi(\text{category}, \text{price}) = \text{OneHot}(\text{category}) + \text{price}$
- D.  $\phi(\text{category}, \text{price}) = \text{price} \times \text{OneHot}(\text{category})$
- E.  $\phi(\text{category}, \text{price}) = \text{category} \times \text{OneHot}(\text{price})$
- F.  $\phi(\text{category}, \text{price}) = \text{Concatenate}(\text{OneHot}(\text{category}), \text{price})$

2 - فرض کنید در یک مسئله طبقه‌بندی تعداد نمونه‌های یکی از کلاس‌ها بسیار بیشتر از کلاس دیگر باشد، برای حل این مشکل چه راهکارهایی ارائه می‌دهید؟

3 - مشکل overfitting در شبکه‌های عصبی چیست؟ راه تشخیص آن چیست؟ چه راهکارهایی برای حل آن وجود دارد (نام برده و شرح دهید) ؟

4 - نمودارهای زیر را در نظر بگیرید:

- A. histogram
- B. pie chart
- C. bar plot
- D. line plot
- E. side-by-side boxplots
- F. scatter plot
- G. stacked bar plot
- H. overlaid line plots
- I. mosaic plot

برای هر یک از مسائل زیر کدام نمودار را پیشنهاد می‌دهید و چرا؟

الف- تعداد اتاق هر واحد مسکونی و قیمت فروش مسکن در شهر تهران

ب- قیمت فروش و تاریخ فروش مسکن در شهر تهران از سال ۱۳۸۰ تا ۱۳۹۰

ج- وزن تولد نوزادان در بیمارستان‌های شهر تهران

د- سطح تحصیلات والدین دانش آموزان دبیرستان دانش

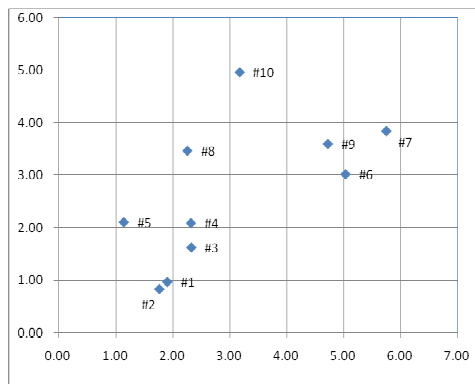
ه- معدل کارشناسی و نمره GRE دانشجویان متقاضی ادامه تحصیل در Toronto Message of Light University

و- نژاد و جنسیت دانشجویان دانشگاه شهید بهشتی

5 - فرض کنید که یک مجموعه داده از اطلاعات فردی و وضعیت تحصیلی دانش‌آموزان یک مدرسه شامل جنسیت، سن، نژاد، وضعیت اقتصادی، و نمرات دروس ریاضی، ادبیات فارسی، علوم تجربی و املاء را جمع‌آوری کرده‌ایم. پنج سوال در مورد این مجموعه داده طرح و آزمون فرض مناسب برای آن پیشنهاد دهید .

6 - آیا می‌توان برای یک طبقه بند نزدیک‌ترین همسایه، یک طبقه بند درخت تصمیم متناظر ساخت ؟ چرا؟

7 - الگوریتم خوشه بندی k-means را برای مجموعه داده‌های زیر تا دو مرحله اجرا کنید. فرض کنید که  $k=2$  و نقاط شروع اولیه برای دو خوشه برابر با نقطه داده شماره ۱ و داده شماره ۱۰ باشد.



Data #	x	y
1	1.90	0.97
2	1.76	0.84
3	2.32	1.63
4	2.31	2.09
5	1.14	2.11
6	5.02	3.02
7	5.74	3.84
8	2.25	3.47
9	4.71	3.60
10	3.17	4.96

8 - الگوریتم کاهش ابعاد ویژگی PCA را شرح دهید.

9 - پنج نفر از کارمندان دانشکده (به غیر از اساتید) و مسئولیتشان را نام ببرید؟