

به نام خداوند بخشنده و مهربان



دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

درس داده کاوی

تمرین اول

استاد درس: دکتر ناظر فرد

نام دانشجو:

روزبه قاسمی ۹۵۳۱۴۲۴

زمستان ۱۳۹۸

سوال اول

A-Supervised:

یادگیری تحت نظارت، بخشی از Machine learning است که هدف آن ایجاد یک تابع از داده های آموزشی نظارت شده می باشد. داده های آموزشی شامل مجموعه ای از نمونه های آموزشی می باشد. در یادگیری تحت نظارت، هر نمونه یک جفت متشکل از یک داده ورودی (معمولاً یک بردار) و یک مقدار خروجی مورد نظر که به آن سیگنال تحت نظارت نیز گفته می شود می باشد. به زبان ساده تر یادگیری تحت نظر نوعی از یادگیری است که در آن سیستم بر اساس نمونه هایی که پیشتر برچسب گذاری شده اند یادگیری را انجام میدهد. به طور مثال می توان به روش های Regression و Classifier ها اشاره کرد.

B-UnSupervised:

یادگیری بدون نظارت روش یادگیری است که در آن سیستم می آموزد تا با الگوهای ورودی مشخص در داده های برچسب گذاری نشده آن را به بخش بزرگی از جامعه آماری تعمیم دهند. در مقایسه با روش یادگیری با نظارت در این روش خروجی مشخصی و یا متناظر با هر ورودی مورد انتظار نمی باشد، بلکه این روش خود کمک میکند تا پیش زمینه های مربوط به این امر را فراهم کنیم. به طور مثال می توان به روش Clustering اشاره کرد.

C-Semi Supervised:

یادگیری نیمه نظارت شده نیز نوعی از یادگیری است که هم از داده های طبقه بندی شده (برچسب خورده) و هم از داده های غیر طبقه بندی شده (برچسب نخورده) به صورت همزمان استفاده میشود تا دقت یادگیری مقداری بهبود یابد. به طور مثال می توان به Reinforcement learning اشاره کرد.

D-Outlier:

داده پرت یا outlier به داده ای می گویند که با دیگر داده های هم گروه فاصله بیشتری داشته باشد مثلاً فرض کنید داده های زیر، قد ۵ نفر از افراد کلاس داده کاوی می باشد:

189,183,192,39,178

داده پرت میان داده های بالا داده ۳۹ می باشد و تناسبی با داده های دیگر ندارد.

Missing Value -E

مقادیر گمشده یک اتفاق معمول است و برای درمان آنها باید استراتژی داشته باشید. مقدار از دست رفته می تواند تعدادی از موارد مختلف در داده های شما را نشان دهد. شاید داده ها در دسترس نبوده یا قابل استفاده نبوده یا واقعه اتفاق نیفتاده است. این ممکن است که شخصی که داده ها را وارد کرده است ، مقدار صحیح را ندانسته و یا تکمیل آن را از دست ندهد. روش های داده کاوی در نحوه برخورد با مقادیر از دست رفته متفاوت است. به طور معمول ، آنها مقادیر گمشده را نادیده می گیرند، یا سوابق حاوی مقادیر گمشده را حذف می کنند ، یا مقادیر گمشده را با میانگین جایگزین می کنند، یا مقادیر گمشده را از مقادیر موجود استنباط می کنند. در تئوری موج ، سر و صدا به عنوان یک سیگنال نامعتبر با هم تداخل داده های معتبر تعریف شده است و آن را مبهم می کند. در داده کاوی نیز فرق نمی کند .در داده کاوی و کیفیت داده ها ، مقدار پر سر و صدا داده ای است که دستگاه ها قادر به درک و تفسیر درست آن نیستند. مقدار پر سر و صدا به طور گسترده ای به عنوان سر و صدای کلاس یا سر و صدای ویژگی طبقه بندی می شود. نویز کلاس وقتی است که یک شیء داده به طور نادرست نشاندار شود. این امر می تواند به دلیل عدم کفایت داده یا خطای ورود داده ها اتفاق بیفتد. سرچشمه ها عمدتاً در مرز داده ها قرار دارند.

نویز ویژگی زمانی است که صفات شیء داده نادرست باشد. ممکن است از دست رفته، نادرست یا اشتباه باشد. آنها مقادیر کاملاً تصادفی هستند.

سوال دوم

برای حل این مشکل چند راهکار موجود است که به ۴ تای آنها اشاره خواهم کرد:

(۱) مقادیر از دست رفته را به صورت دستی وارد کنیم.

(۲) برای پر کردن مقدار از دست رفته از مقدار میانگین ویژگی استفاده کنیم.

(۳) برای پر کردن مقدار از دست رفته از یک ثابت جهانی استفاده کنیم.

(۴) از میانگین ویژگی برای کلیه نمونه های متعلق به کلاس مشابه با Tuple داده شده استفاده کنیم.

سوال سوم

$$x \rightarrow y$$

$$\text{Support}(x \rightarrow y) = P(x \cap y)$$

$$\text{Confidence}(x \rightarrow y) = P(y|x) = \frac{P(x \cap y)}{P(x)}$$

$$\text{min-support} = 40\%$$

$$\text{min-conf} = 10\%$$

حال جستجوی آسان است

آسان به حل الگوریتم Apriori می پردازیم

item sets	support	frequent item sets (one)
{M}	$\frac{3}{5}$	✓
{O}	$\frac{3}{5}$	✓
{N}	$\frac{2}{5}$	X
{K}	$\frac{3}{5}$	✓
{E}	$\frac{3}{5}$	✓
{Y}	$\frac{3}{5}$	✓
{D}	$\frac{1}{5}$	X
{A}	$\frac{1}{5}$	X
{C}	$\frac{2}{5}$	X
{U}	$\frac{1}{5}$	X
{I}	$\frac{1}{5}$	X

حال برآ زنج اکتیم مایندیم :

itemsets	Support	frequent itemsets (two)
$\{u, o\}$	1/5	X
$\{u, k\}$	3/5	✓
$\{u, e\}$	2/5	X
$\{u, y\}$	1/5	X
$\{o, k\}$	3/5	✓
$\{o, e\}$	3/5	✓
$\{o, y\}$	1/5	X
$\{k, e\}$	4/5	✓
$\{k, y\}$	3/5	✓
$\{e, y\}$	2/5	X

حال بزرگ حالت نه تایی را باید برابر با k قبلی بدست بیاوریم

itemsets	support	Frequent itemsets (three)
$\{M, O, K\}$	$\frac{1}{5}$	X
$\{M, K, E\}$	$\frac{2}{5}$	X
$\{M, K, Y\}$	$\frac{2}{5}$	X
$\{O, E, K\}$	$\frac{3}{5}$	✓
$\{K, E, Y\}$	$\frac{2}{5}$	X
$\{O, K, Y\}$	$\frac{2}{5}$	X

پاسخ این سؤال از روش Apriori برابر است با $\{O, E, K\}$

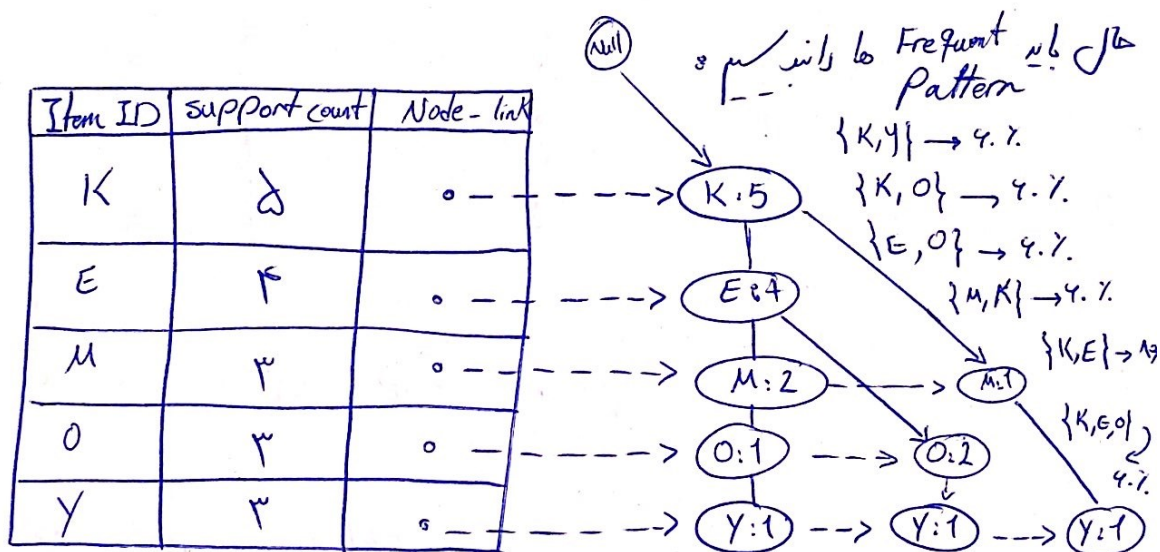
حال باید روش $FP\text{-}growth$ را انجام دهیم

اولین مرحله مشابه الگوریتم Apriori با 1-itemsets در $min\text{-support}$ است

از قبل می دانیم که k itemset ها قبلی مورد قبول برابر است با

$\{M\}, \{K\}, \{O\}, \{E\}, \{Y\}$

TID	Itemsets - bought	ordered itemsets	items	Conditional Pattern base	Conditional FP Tree
T ₁₀₀	{M, O, N, K, E, Y}	K, E, M, O, Y	Y	{KEMO:1}, {KEO:1}, {KM:1}	{K:3}
T ₂₀₀	{D, O, N, K, E, Y}	K, E, O, Y	O	{KEM:1}, {KE:1}	{KE:3}
T ₃₀₀	{M, A, K, E}	K, E, M	M	{KE:1}, {K:1}	{K:3}
T ₄₀₀	{M, O, C, K, Y}	K, M, Y	E	{K:1}	{K:4}
T ₅₀₀	{C, O, O, K, Z, E}	K, E, O	K	—	—



مقایسه الگوریتم Apriori و FP-Growth

الگوریتم Apriori به فضای حافظه زیادی احتیاج دارد زیرا با تعداد زیادی از موارد تولید آیت‌های نامزد سروکار دارد. الگوریتم FP-Growth به حافظه کمتری احتیاج دارد که بخاطر ساختار فشرده آن، آیت‌های مکرر را بدون تولید آیت‌های مورد نظر پیدا می‌کند. الگوریتم Apriori از روش جستجو برای اولین بار استفاده می‌کند و FP-Growth از روش Divide and Conquer استفاده می‌کند و هر دو الگوریتم Apriori و FP-Growth برای استخراج الگوهای مکرر از پایگاه داده استفاده می‌شوند. هر دو

الگوریتم برای کشف الگوهای مکرر از برخی تکنیک ها استفاده می کنند. الگوریتم Apriori با بانک اطلاعاتی بزرگ خوب کار می کند اما الگوریتم FP-Growth با بانک اطلاعاتی بزرگ بد کار می کند.

سوال چهارم

الف) از آنجایی که ۲ جنس همرا داریم پس طبق فرمول داریم:

$$C^d - C^{d+1} + 1 = 3^4 - 2^7 + 1 = \frac{902}{7} \text{ association rules}$$

ب) بزرگترین Frequent itemset در این سؤال برابر ۴ می باشد که آن {milk, Diapers, Bread, Butter} این جنسیت جز ۴ این itemset دربار نگذاشته است.

پ) با فرض داشتن جابجست به نگار به پاسخ این قسمت می رسیم:

$$\frac{2}{4} \times \frac{5}{4} \times \frac{4}{4} \approx \frac{2 \times 5 \times 4}{32} = \frac{40}{32}$$

ت) با توجه به جدول {Bread, Butter} با تعداد support ۵ بیشترین support را دارد. itemset 2- ها دارد.

ث) می دانیم که فرمول Confidence برابر است با:

$$\text{Confidence}(X \rightarrow Y) = P(Y|X) = \frac{P(X \cap Y)}{P(X)}$$

از جدول می بینیم که در عمود قبلی وجود دارد، به آن دقت می شود که Butter

Bread از جایگاه اول به یک به خوردار می شود و در این Bread → Butter

$$\text{Confidence یک به یک} = \frac{Bread \rightarrow Butter}{5} = 1$$

$$Butter \rightarrow Bread = \frac{5}{5} = 1$$

سوال پنجم

به صورت کلی می‌توانیم بگوییم که:

(۱) به شما کمک می‌کند تا روی جالب‌ترین ویژگی‌های مجموعه آموزش تمرکز کنید. شما این کار را برای استفاده از نمونه‌های جدید گسترش می‌دهید، با فرض اینکه مجموعه آموزشی نماینده شما به طور دقیق به مهمترین ویژگی‌ها اشاره می‌کند.

(۲) به شما کمک می‌کند تا روی نکات مهم داده‌ها، بهبود زمان آموزش (همیشه، در صورت کاهش) و بعضی اوقات بهبود تعمیم مدل، تمرکز کنید (زیرا مدل‌ها توسط جمعیت بیش از حد در بعضی از کلاس‌ها و محیط‌های خاص کمتر اشتباه می‌شوند).

(۳) کاهش داده می‌تواند باعث افزایش راندمان ذخیره‌سازی و کاهش هزینه‌ها شود. فروشندگان ذخیره‌سازی غالباً ظرفیت ذخیره‌سازی را از نظر ظرفیت خام و ظرفیت مؤثر توصیف می‌کنند، که به داده‌ها پس از کاهش اشاره دارد.

حال می‌توان به استراتژی‌های موجود نیز اشاره کرد:

- جمع‌آوری داده‌های مکعب
- فشردن داده‌ها
- تفسیر و تولید سلسله‌مراتب مفهومی
- کاهش ابعاد

سوال ششم

اگر رابطه بین دو متغیر صفر باشد، می‌توان نتیجه گرفت که هیچ رابطه خطی بین متغیرها وجود ندارد، زیرا همبستگی رابطه خطی بین دو متغیر را اندازه‌گیری می‌کند. بنابراین این دو متغیر ممکن است به روش‌های غیر خطی دیگری مرتبط باشند. بنابراین، ما نمی‌توانیم اطمینان حاصل کنیم که متغیرهای ما مستقل هستند؛ شاید آن‌ها دارای برخی روابط قوی اما غیر خطی باشند.

سوال هفتم

$A_1(2, 1)$, $A_2(2, 5)$, $A_3(8, 4)$, $B_1(5, 1)$, $B_2(7, 5)$, $B_3(6, 4)$,
 $C_1(1, 2)$, $C_2(4, 1)$

الف) از طریق روش Euclidean metric ، فاصله هر کدام از point با دیگرش را محاسبه کنید. نتایج آن به شکل زیر است :

	A1	A2	A3	B1	B2	B3	C1	C2
A1	0	5	8, 4	3, 6	7, 5	7, 1	1, 0	2, 2
B1	3, 6	4, 2	5	0	3, 4	4, 1	7, 1	1, 4
C1	1, 0	3, 2	7, 1	7, 1	4, 7	5, 3	0	3, 4

بسیار از در این مورد داریم :

cluster K2 = {A3, B1, B2, B3, C2}

cluster K3 = {A2, C1}

ب) مشابه قسمت قبل اما clustering را با دو دسته کنیم :

Centroid 1 : (2, 1)

Centroid 2 : $\left(\frac{1+5+7+6+4}{5}, \frac{4+1+5+1+9}{5} \right) = (4, 4)$

Centroid 3 : (1, 5 , 2, 5)

	A1	A2	A3	B1	B2	B3	C1	C2
Centroid 1	0	0	1, 2, 1	3, 6	4, 0, 5	4, 2, 1	1, 0, 6	2, 1, 2
" 2	4, 1, 2	4, 1, 2	2, 1, 2	2, 2, 2	1, 2, 1	2	4, 2	2, 2
" 3	4, 2, 1	1, 2, 1	4, 2, 1	0, 2, 5	0, 2, 5	4, 2, 2	2, 2, 1	4, 0, 2

\Rightarrow Cluster $K' 1 = \{A1, C2\}$
 $" K' 2 = \{A3, B1, B2, B3\}$
 $" K' 3 = \{A2, C1\}$

حالا نسبت به مرکز جدید آفراسات

Centroid 1 $\rightarrow (4, 9, 0)$, Centroid 2 $\rightarrow (1, 2, 3)$, Centroid 3 $\rightarrow (4, 0, 2)$

	A1	A2	A3	B1	B2	B3	C1	C2
Centroid 1	4, 1, 2	4, 2, 1	4, 2, 2	2, 2	4, 0, 2	4, 2, 2	4, 0, 2	1, 1, 9
" 2	4, 2, 1	1, 2, 1	4, 2, 2	0, 2, 5	0, 2, 5	4, 2, 2	1, 2, 1	4, 0, 2
" 3	4, 2, 2	4, 2	1, 2, 2	2, 1, 2	0, 2, 6	1, 2, 2	4, 2, 2	4, 0, 1

و نسبت به مرکز جدید آفراسات

Cluster $K'' 1 = \{A1, B1, C2\}$

" $K'' 2 = \{A2, C1\}$

" $K'' 3 = \{A3, B2, B3\}$

سوال هشتم

از آنجایی که جواب K -means وابسته به initialization است ممکن است همه به جواب $global\ optimum$ نرسیم. پس ممکن است با انتخاب نادرست initialization در نهایت به $local\ optimum$ برسیم. با مثال زیر بهترین موضوع را نشان می‌دهیم:

$$M = \{1, 2, 3, 4\} \Rightarrow \begin{cases} \text{Centroid 1} = 1 \rightarrow \text{cluster 1} = \{1\} \\ \text{" 2} = 3 \rightarrow \text{" 2} = \{2, 3, 4\} \end{cases}$$

این جوابی که بدست آوردیم ممکن است $local\ optimum$ باشد زیرا ممکن است centroid نادرست آن به شکل زیر باشد:

$$\begin{cases} \text{Centroid 1} = 1.5 \\ \text{" 2} = 3.5 \end{cases} \Rightarrow \begin{cases} \text{cluster 1} = \{1, 2\} \\ \text{" 2} = \{3, 4\} \end{cases}$$

قسمت A

Section A

Read data from dataset file

```
[2] import pandas as pd
data = pd.read_csv('/content/drive/My Drive/Data Mining/HW1/data.csv')
```

Show dataset in one table

```
[3] data
```

	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state
0	1	female	1984.0	China	filtered at airport	visit to Wuhan	NaN	1/20/2020	released
1	2	male	1964.0	Korea	filtered at airport	visit to Wuhan	NaN	1/24/2020	released
2	3	male	1966.0	Korea	capital area	visit to Wuhan	NaN	1/26/2020	released
3	4	male	1964.0	Korea	capital area	visit to Wuhan	NaN	1/27/2020	released
4	5	male	1987.0	Korea	capital area	visit to Wuhan	NaN	1/30/2020	released
...
171	172	female	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
172	173	male	1949.0	Korea	Daegu	NaN	NaN	2/24/2020	deceased
173	174	female	1958.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
174	175	male	1997.0	Korea	Gyeongsangbuk-do	NaN	NaN	2/24/2020	isolated
175	176	female	1950.0	Korea	capital area	NaN	NaN	2/24/2020	isolated

176 rows × 9 columns

در این قسمت ابتدا کتابخانه pandas را import میکنیم تا از آن استفاده کنیم. سپس از طریق تابع read_csv فایل دیتاست را از drive می‌خوانیم. سپس با صدا کردن دوباره variable ای که مجموعه داده را در آن لود کردیم، آنرا چاپ می‌کنیم. در ادامه نیز ابتدا و انتها مجموعه داده را نیز چاپ کرده ام. این برای آن است که نشان دهم یکی از function های مرتبط به چاپ جدول صدا کردن Head و Tail است.


Show the head of dataset

```
[4] data.head()
```

	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state
0	1	female	1984.0	China	filtered at airport	visit to Wuhan	NaN	1/20/2020	released
1	2	male	1964.0	Korea	filtered at airport	visit to Wuhan	NaN	1/24/2020	released
2	3	male	1966.0	Korea	capital area	visit to Wuhan	NaN	1/26/2020	released
3	4	male	1964.0	Korea	capital area	visit to Wuhan	NaN	1/27/2020	released
4	5	male	1987.0	Korea	capital area	visit to Wuhan	NaN	1/30/2020	released

Show the end of dataset

```
[5] data.tail()
```



	id	sex	birth_year	country	region	infection_reason	infected_by	confirmed_date	state	
171	172	female	1997.0	Korea	Gyeongsangbuk-do		NaN	NaN	2/24/2020	isolated
172	173	male	1949.0	Korea	Daegu		NaN	NaN	2/24/2020	deceased
173	174	female	1958.0	Korea	Gyeongsangbuk-do		NaN	NaN	2/24/2020	isolated
174	175	male	1997.0	Korea	Gyeongsangbuk-do		NaN	NaN	2/24/2020	isolated
175	176	female	1950.0	Korea	capital area		NaN	NaN	2/24/2020	isolated

قسمت B

Section B

Show the shape of dataset

This dataset contains 176 data which have nine attributes for each data

```
[6] data.shape
```

```
(176, 9)
```

With this instruction we could find the information of dataset

it shows the name of each attribute with the type of them and etc.

```
[7] data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 176 entries, 0 to 175
Data columns (total 9 columns):
id                176 non-null int64
sex               176 non-null object
birth_year        166 non-null float64
country           176 non-null object
region            166 non-null object
infection_reason  95 non-null object
infected_by       42 non-null float64
confirmed_date    176 non-null object
state             176 non-null object
dtypes: float64(2), int64(1), object(6)
memory usage: 12.5+ KB
```

در این قسمت نیز ابعاد آنرا چاپ میکنیم و در ادامه نشان دادم که از طریق info اطلاعات هر attribute جدول را نشان دادم.

قسمت C

Section C

در این قسمت سه معیار max و mean و std برای سال های تولد را بدست می آوریم.

```
[8] data[['birth_year']].max()
```

```
birth_year    2009.0
dtype: float64
```

```
[9] data[['birth_year']].mean()
```

```
birth_year    1973.385542
dtype: float64
```

```
[10] data[['birth_year']].std()
```

```
birth_year    17.032825
dtype: float64
```

قسمت D

Section D

Show the number of filed in 'region' column

در این قسمت تعداد field های موجود در ستون Region را بدست می آوریم.

```
[11] data.region.count()
```

```
166
```

Section E

As we know in section B, we have 176 entries and based on info about some columns like birth year and region and infection reason, etc. have less than 176 non-null data. So we have null cells. There are 5 known ways to handle null cells. one way is deleting the whole row and another is replacing it with mean so that it doesn't affect the results in a wrong way or using methods that support missing value or predict the missing values etc. we implement two of these strategies.

```
[12] data['birth_year'].isnull().sum()
```

```
10
```

```
[13] data['birth_year'].mean()
```

```
1973.3855421686746
```

```
[14] import numpy as np
data['birth_year'].replace(np.NaN, data['birth_year'].mean()).head(20)
```

```
0    1964.0
1    1964.0
2    1966.0
3    1964.0
4    1967.0
5    1964.0
6    1991.0
7    1957.0
8    1992.0
9    1966.0
10   1995.0
11   1971.0
12   1992.0
13   1980.0
14   1977.0
15   1977.0
16   1982.0
17   1999.0
18   1983.0
19   1978.0
Name: birth_year, dtype: float64
```

قسمت E

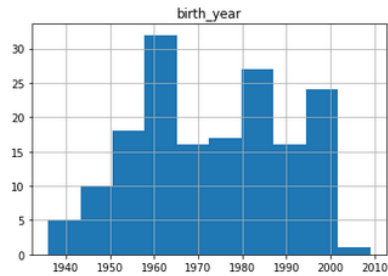
توضیحات این قسمت در متن به صورت انگلیسی نوشته شده است 😊

قسمت F

Section F

```
[15] import matplotlib.pyplot as plt
      data.hist(column='birth_year')

array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f91be27be48>]],
      dtype=object)
```



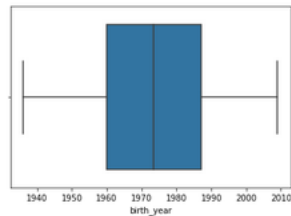
قسمت G

Section G

In this data set, only column birth_year is a numeric pillar and we only need to check that outlier has taken place.

```
import seaborn as sns
sns.boxplot(x=data['birth_year'])

<matplotlib.axes._subplots.AxesSubplot at 0x7f91b33d2470>
```



However, based on section F, we may suppose people born between 2000 and 2010 or 1940 and 1950 are outliers because they are so few. In larger scales, we should delete or replace these from the dataset so that we get better performance.

```
[17] import matplotlib.pyplot as plt
      from collections import Counter

      data_dict = dict(Counter(data['birth_year']))
      data_list = sorted(data_dict.items())
      x, y = zip(*data_list)
      plt.scatter(x, y)
      plt.show()
```



پایان