Data mining
Homework #1

Dr Amirmazlaghani
Amirkabir University of Technology
(Tehran polytechnic)

توجه: پیش از شروع تمرین لطفا موارد زیر را با دقت مطالعه نمایید.

لطفا تمام فایل‌های تمرین را (از جمله فایل pdf گزارش و فایل‌های کد) در یک فایل zip/rar ذخیره کرده و نام آن را به HW1_XXXXXXXX.zip تغییر دهید. سپس آن را در مودل بارگذاری کنید.

سوال‌ها به دو بخش نظری و برنامه‌نویسی تقسیم شده‌اند. سوال‌های نظری را می‌توانید به جای فایل word در برگه کاغذ انجام داده و تصویر آن را در فایل word قرار دهید. دقت کنید که خوانایی تمرین شرط لازم برای دریافت نمره آن است. توجه کنید که فایل گزارش را حتما به صورت pdf شده تحویل دهید.

تمرین‌های برنامه‌نویسی را می‌توانید با یکی از زبان‌های Matlab یا Python انجام دهید.

برای هر سوال باید کدی جدا نوشته شود. برای مثال کدهای سوال ۳ بخش (a) را در فایل p3a.py ذخیره کنید.

مهلت تحویل تمرین تا تاریخ ۱۴ آبان است. در صورت تاخیر در بارگذاری، به مدت دو روز به ازای هر روز تاخیر ۱۰% از نمره اولیه تمرین کاسته خواهد شد.

سوال اول

از روشهای خوشه بندی ژن‌های درون سلول استفاده از میزان شباهت فعال بودن ژن ها است. فرض کنید در آزمایشگاهی مقدار فعالیت دو ژن g۱ و g۲ در زمان‌های c۱ تا c۱۶ اسکن شده است. جدول زیر این مقادیر را نشان می‌دهد. آیا میتوان این دو ژن را در یک خوشه قرار داد؟

	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10
g1	-5	-4	-3	-2	-1	1	2	3	4	5
g2	25	16	9	4	1	1	4	9	16	25

(a) **با استفاده از معیار شباهت همبستگی به این سوال پاسخ دهید.**

(b) **معیار شباهت MI را برای دو ژن محاسبه کنید. و به این سوال پاسخ دهید.**

(c) **ایا تفاوتی در نتیجه مشاهده میکنید؟ اگر بله چرا؟**

سوال دوم

برای وکتورهای داده شده موارد مطلوب را بدست آورید:

(a) $x=(1,1,1,1)$ و $y=(2,2,2,2)$

cosine, correlation, euclidean

(b) $x=(0,1,0,1)$ و $y=(1,0,1,0)$

cosine, correlation, euclidean, jaccard

(c) $x=(1,1,0,1,0,1)$ و $y=(1,1,1,0,0,1)$

manhattan, correlation, bhattacharyya

سوال سوم

ویژگی های زیر را به صورت دوتایی (binary)، گسسته (discrete) یا پیوسته (ontinuous) تعریف کنید. همچنین آن ها را به عنوان کیفی (اسمی یا ترتیبی) یا کمی (بازه یا نسبت) دسته بندی کنید. نکته: بعضی موارد ممکن است بیشتر از یک حالت داشته باشند. بنابراین، در صورتی که فکر می‌کنید چند حالت وجود دارد، دلیل انتخابتان را ذکر کنید.

مثال: سن به سال

پاسخ: گسسته، کمی، نسبت

(a) روشنایی اندازه‌گیری شده توسط نورسنج

(b) روشنایی از روی قضاوت ناظر انسانی

(c) زاویه اندازه‌گیزی شده بین ۰ و ۳۶۰ درجه

(d) ارتفاع از سطح اقیانوس

(e) درجه نظامی

(f) مدال‌های طلا و نقره و برنز در بازی المپیک

سوال چهارم

به کمک داده های آموزش Iris که از طریق لینک زیر، فایل csv. آنها قابل دسترسی است، هر یک از موارد زیر را تکمیل کنید.

https://archive.ics.uci.edu/ml/datasets/iris

(a) هیستوگرام داده ها را به کمک استفاده از کتابخانه Matplotlib در یک مختصات دو بعدی رسم نمایید و خروجی را به گزارش خود اضافه کنید. توجه کنید که تعداد ویژگی ها بیشتر از دو می باشد که در این صورت ناچار خواهید بود دو ویژگی را به دلخواه خود انتخاب کنید.

(b) مرحله ی قبل را با استفاده از کتابخانه Matplotlib و با کمک کلاس Axes3D در یک مختصات سه بعدی تکرار کنید. برای راهنمایی می توانید به اسلاید پانزدهم لکچر چهارم مراجعه کنید تا نمونه پلات سه بعدی را رویت کنید.

(c) داده ها را با در نظر داشتن دو ویژگی دلخواه خود و با استفاده از کلاس Axes3D در یک نمودار پراکندگی یا Scatter Plot رسم نمایید. نتیجه را در گزارش قرار دهید.

(d) به کمک کتابخانه Numpy یا Pandas، مقدار میانگین (mean) و واریانس (variance) داده ها را بدست آورید و نتیجه را به گزارش خود اضافه کنید. دقت کنید که با بیشتر از یک ویژگی در این داده ها سرو کار دارید، در نتیجه مقدار میانه و واریانس در هر بعد باید محاسبه شود و به صورت یک بردار در خروجی نمایش داده شود.

(e) به دلخواه خود دو کلاس را انتخاب کرده و ماتریس کواریانس آن دو را بدست آورید و خروجی را به گزارش خود اضافه کنید. دقت کنید که ماتریس کواریانس در این حالت یک ماتریس مربعی دو در دو خواهد بود.

(f) به دلخواه خود دو کلاس را انتخاب کرده و ماتریس همبستگی آن دو را بدست آورید و خروجی را به گزارش خود اضافه کنید. تفاوت این ماتریس با ماتریس مرحله قبل در چیست؟ لطفا پاسخ این سوال را نیز به گزارش اضافه فرمایید.

(g) در گیاه ویرجینیکا کدام دو ویژگی شباهت بیشتری دارند؟ چرا؟

(h) فرض کنید بخواهید با دانستن یک ویزگی نوع گیاه را حدس بزنید. دانستن کدام ویژگی بهتر به رسیدن جواب کمک میکند؟ چرا؟

سوال پنجم

در این سوال قرار است انواع رگرسیون را پیاده سازی کنید. داده های این سوال در فایل‌های data.mat و data.npz قرار دارند. این داده‌ها براساس رابطه‌ی زیر تولید شده اند :

$$y = 4\,x_2^2\,x_1 + 2\,x_2^2 + 3\,x_1 + 1$$

بنابراین نمونه‌ها به صورت $[x_1, x_2]$ و خروجی متناظرشان y است.

در فایل data.npz و data.mat شش آرایه‌ی یک بعدی، $y, x_1, x_2, y_{test}, x_{1,test}, x_{2,test}$ قرار دارند.

یک مثال از دسترسی به آرایه‌ی y فایل در زبان پایتون، چنین است :

```python
#python
a = np.load('data.npz')
print(a['y'])
```

آرایه‌ی y_{test} ، خروجی صحیح روی داده‌های $(x_{1,test}, x_{2,test})$ است و شما باید خروجی کد خود را با این آرایه مقایسه کنید.

در هر سه حالت تابع هزینه‌ی ما تابع SSE (Sum of squared errors) خواهد بود.

سه حالت را با استفاده از Gradient decent و Stochastic gradient decent پیاده‌سازی کرده و نتایج را روی داده های تست با هم مقایسه کنید.فرض کنید فرم کلی رابطه‌ی y و xها را می‌دانیم از فرمول بسته‌ی رگرسیون خطی/خطی تعمیم یافته برای محاسبه‌ی ضرایب (W) استفاده کنید.

امتیازی: برای داده های تست، نتیجه‌ی کد خودتان و مقادیر صحیح خروجی که در آرایه ی y_{test} است را بر حسب $x_{1,test}$ و $x_{2,test}$ به صورت نمودار سه بعدی نمایش دهید.

مقدار تابع خطا روی داده های آموزش و داده های تست را برای هر یک از سه حالت گزارش نمایید.