به نام خدا

دانشگاه صنعتی شریف - دانشکده مهندسی برق

آمار و احتمال مهندسی – گروه 4

دکتر میرمحسنی

نیمسال دوم 97-98

تمرین سری چهار MATLAB

مهلت تحویل 98/4/8

به نکات زیر توجه کنید:

- فایل تحویلی باید به فرمت zip. یا rar. و شامل یک فایل m. (شامل کدهای تمام سوالات) و توابع نوشته شده (در صورت وجود!) و گزارش به فرمت pdf باشد. گزارش باید شامل نمودارها و نتایج خواسته شده و پاسخ به تمامی سوالات و محاسبات دستی و اثبات‌های لازم باشد. اسم فایل را به فرم HW4_Student_Number قرار دهید و در سامانه CW آپلود کنید.
- دقت کنید که کدهای شما debug نخواهد شد!
- نمودارها باید دارای عناوین مشخص باشند.
- کد های خود را در گزارش کار نیاورید و از publish کردن و livescript بپرهیزید.
- برای راحتی ابتدای کد از دستورهای clc، clear all و close all استفاده کنید.
- کدهای بخش‌های مختلف را به وسیله %% از هم جدا کنید و کامنت گذاری مناسب انجام دهید.
- از کپی کردن هم جدا بپرهیزید. :)
- در صورت داشتن هرگونه سوال، به psmatlab98spring@gmail.com ایمیل بزنید.

سوال اول : آزمون آماری t-test

برای رد یا تایید فرضیه‌ای درباره‌ی توزیع یک داده و یا تفاوت بین چند دسته داده، از آزمون فرض‌های آماری (statistiacal hypothesis tests) استفاده می‌کنیم. یکی از این آزمون‌های آماری t-test است.

در آزمون‌های آماری دو فرضیه در نظر می‌گیریم : null hypothesis (H_0) و alternative hypothesis (H_1) . در H_0 فرض می‌کنیم ادعا ما درمورد توزیع نمونه‌ها صحیح است و در H_1 خلاف آن را در نظر می‌گیریم و در طول آزمون بررسی می‌کنیم که آیا شواهد کافی برای رد H_0 داریم یا نه.

فرض کنید نمونه‌های تصادفی $X_1.X_2.\dots.X_n$ از یک توزیع را داریم و می‌خواهیم درمورد میانگین این توزیع (μ) فرضیه‌ای را ثابت کنیم. به طور مثال فرض کنید می‌خواهیم ثابت کنیم میانگین توزیع مورد نظر برابر با μ_0 است. در این صورت :

$$H_0: \mu = \mu_0$$

$$H_1: \mu \neq \mu_0$$

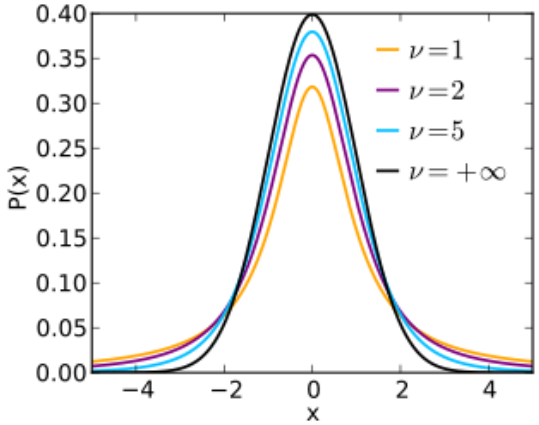
برای این کار باید با توجه به میانگین و واریانس نمونه‌هایی که در اختیار داریم، بررسی کنیم H_0 با چه احتمالی رد می‌شود.
آماره‌ای که می‌توانیم برای بررسی ادعا خود استفاده کنیم را به صورت زیر تعریف می‌کنیم:

$$W_1 = \frac{\bar{X} - \mu_0}{\sigma\sqrt{n}}$$

که در آن $\bar{X}$ میانگین نمونه ها و σ واریانس توزیع است. در صورتی که واریانس توزیع برای ما مشخص نباشد به جای σ از انحراف معیار استاندار نمونه‌ها استفاده می‌کنیم:

$$W_2 = \frac{\bar{X} - \mu_0}{S\sqrt{n}}$$

با فرض این که توزیع مورد نظر گاوسی یا تقریبا گاوسی (bell-shaped) است، W_1 از توزیع $N(0.1)$ و W_2 از توزیع t با درجه آزادی $n-1$ پیروی می‌کند.



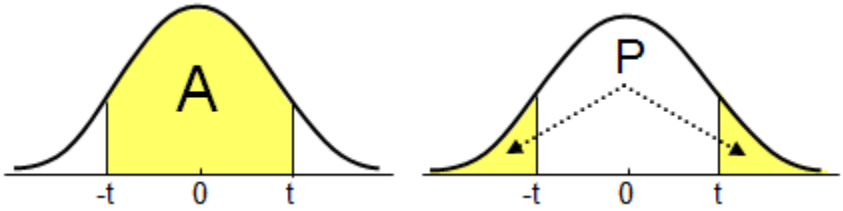
t-distribution

برای بررسی فرضیه مورد نظر خطا E را به صورت زیر تعریف می‌کنیم:

$$E = P(type\ I\ error) = P(reject\ H_0 | H_0\ is\ true) \leq \alpha$$

که یعنی احتمال این که با وجود درست بود فرض H_0 آن را رد کنیم. مقدار α را significance level می‌نامیم و حدی از این خطاست که برای ما قابل قبول است.

در مثالی که زده شد، باید حدی برای W_2 (W_1) در نظر بگیریم (t) که اگر $|W_2| > t$ بود، H_0 را رد کنیم و E کمتر از α باشد. و در غیر این صورت دلیلی برای رد H_0 نداشته باشیم. همانطور که گفتیم آماره مورد استفاده از توزیع گاوسی یا t (مانند شکل زیر) پیروی می‌کند و باید t را طوری تعیین کنیم که در شکل زیر P کمتر از α باشد.



به این ترتیب برای انجام آزمون بر روی نمونه‌ها باید اول میانگین و انحراف معیار نمونه‌ها و سپس مقدار W را به دست بیاوریم (w) . سپس مقدار p-value را به این صورت تعریف می‌کنیم: مقدار خطا در صورتی که حد (threshold)، t برابر با w باشد. اگر مقدار p-value کوچک باشد (کوچک‌تر از α) شواهد قوی برای رد H_0 داریم و در غیر این صورت دلیل محکمی برای رد کردن فرضیه نداریم و آن را با احتمال خوبی درست در نظر می‌گیریم.

در این تمرین قصد داریم به کمک MATLAB آزمون t-test را برای تایید یا رد چند فرضیه پیاده کنیم.

در این مسئله ما با اطلاعات تولد در ایالت کالیفرنیا کار میکنیم. فایل California_birth_1 شامل دویست نمونه تصادفی از فایل California_birth است. در این مسئله مهم است که با فایل California_birth_1 کار کنید.

نام متغیر	توضیح
mage	سن مادر
tgrams	وزن بچه بر حسب گرم
smoke	اگر مادر در حین حاملگی سیگار می کشید یک و در غیر این صورت صفر.
weeks	تعداد هفته های بارداری
mage	سن مادر

1) ابتدا میانگین، میانه، واریانس، مینیمم و ماکزیمم متغیر های weeks و tgrams را بیابید.
2) در مورد t test تحقیق کنید و کاربرد ها و نوع محاسبات در آن را به صورت خلاصه بنویسید.

3) به کمک t test نشان دهید که آیا هر کدام از دو گزاره زیر صحت دارند یا خیر:

الف) متوسط سن مادر هایی که بچه به دنیا می آورند در کالیفرنیا بالای 25 سال است با سطح اهمیت(significance level) 0.05.

ب) متوسط تعداد هفته های بارداری در کالیفرنیا زیر 39 هفته است.

4) در یک نمودار وزن بچه های مادر های سیگاری و غیر سیگاری را در کنار هم نشان دهید. نشان دهید که شما فرضیه صفر را برای گزاره زیر رد می کنید یا می پذیرید.

متوسط وزن بچه های سیگاری کمتر از متوسط وزن بچه های غیرسیگاری است.

- برای همه ی تست های بالا null hypothesis و P-value را باید محاسبه کنید.

سوال دوم : آشنایی با دسته بندی (Classification)

شما در درس با علم آمار و احتمال آشنا شده اید و احتمالا عبارت یادگیری ماشین را تا به حال شنیده اید، در این قسمت با استفاده از ابزار آمار و احتمالی که در درس آموخته اید به حل مسله‌ای در زمینه یادگیری ماشین خواهیم پرداخت.

در ابتدا بهتر است قدم به قدم با مسله آشنا شویم.

در تمرین سری سوم شما با دیتاست آشنا شدید ، به زبان ساده می توان گفت دیتاست ماتریسی است که هر ستون آن یک ویژگی (feature) یا متغییر تصادفی مورد بررسی و هر سطر متناظر با یک نمونه (sample) یا تحقق از وقوع آن متغییر هاست . معمولا هر سطر با یک دسته (class) در تناظر است ودر اغلب موارد نمونه های یک دیتاست متعلق به دو یا چند دسته متفاوتند.

مسله‌ای که میخواهیم برای یک دیتاست خاص آن را حل کنیم مسله‌ی دسته بندی (classification) داده ها است، در واقع ما دسته های مختلف داده ها را می شناسیم و با بررسی ویژگی های یک نمونه می‌خواهیم تعیین کنیم این نمونه متعلق به چه دسته ای بوده است.

یکی از روش‌هایی که به حل این مسله میپردازد Naïve Bayesian Classifier نام دارد که برای هر نمونه با بیشینه کردن احتمال دسته به شرط وقوع آن نمونه، محتمل ترین دسته را می یابد و نمونه را به آن نسبت می دهد، شما با این classifier در درس تحت عنوان یک قاعده تصمیم‌گیری آشنایی دارید.

ترجمه ی ریاضی جملات بالا بدین صورت است :

Xi	ith Column of dataset , ith feature
$X = [X_1, X_2, \dots, X_n]$	Dataset
$x = [x_1, x_2, \dots, x_n]$	A sample
n	$number\ of\ features$
C_k	$kth\ Class\ label$
K	Number of Classes
$y(x)$	Detected Class label for sample x

الگوریتم برای پیدا کردن دسته هر نمونه :

$$y(x) = \underset{k=1:K}{\operatorname{argmax}} P(C_k|x)$$

باوجود امکان اما محاسبه ی احتمال $P(c_k|x)$ با استفاده از تعریف فراوانی احتمال به دلیل مختلف بود حالات x و امکان زیاد بود تعداد ویژگی ها کار سختی است، با استفاده از قاعده ی بیز آن را بر حسب احتمالاتی می نویسیم که می توان آن را با استفاده از دیتاست آسان تر محاسبه کرد پس داریم :

$$y(x) = \underset{k=1:K}{\operatorname{argmax}} P(C_k|x) = \underset{k=1:K}{\operatorname{argmax}} \frac{P(C_k)P(x|C_k)}{P(x)}$$

احتمال $P(x)$ مستقل از از دسته ها است و در ماکسیمم سازی نقشی ندارد پس می تواند حذف شود بنابراین مساله بهینه سازی به صورت زیر می شود :

$$\underset{k=1:K}{\operatorname{argmax}} P(C_k)P(x|C_k) = \underset{k=1:K}{\operatorname{argmax}} P(x, C_k) = \underset{k=1:K}{\operatorname{argmax}} P(x_1, x_2, .. x_n, C_k)$$

حال محاسبه ی این احتمال joint نیز آسان نمی باشد و مستلزم ساختارهای شرطی طولانی و صرف زمان است که بهینه نمیباشد، برای رفع این مشکل یک **فرض اضافی** در Naïve Bayesian Classifier در نظر گرفته می شود و آن استقلال متغییر های تصادفی یا ویژگی ها از هم است، با این فرض می توان نوشت : (با استفاده از قانون زنجیری احتمال)

$$\begin{aligned} p(C_k, x_1, \ldots, x_n) &= p(x_1, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2, \ldots, x_n, C_k) \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k)\, p(x_3, \ldots, x_n, C_k) \\ &= \ldots \\ &= p(x_1 \mid x_2, \ldots, x_n, C_k)\, p(x_2 \mid x_3, \ldots, x_n, C_k) \ldots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

با فرض استقلال گفته شده :

$$p(x_i \mid x_{i+1}, \ldots, x_n, C_k) = p(x_i \mid C_k)$$

پس می توان گفت :

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

پس دسته تخصیص یافته به نمونه x به صورت زیر تعیین می‌شود :

$$y(x) = \underset{k=1:K}{\operatorname{argmax}} P(C_k) \prod_{i=1}^{n} P(x_i | C_K) \qquad (I)$$

که محاسبه احتمالات دخیل در آن با کمک دیتاست و تعریف فراوانی احتمال کار به مراتب ساده‌تری است و حجم محاسبات کمتری دارد.

توضیح درباره‌ی دیتاستی که قرار است با آن کار کنید :

در این تمرین با دیتاستی به نام Nursery کار خواهید کرد که اطلاعات متقاضیان ثبت نام در مهدکودک ها و نتیجه پذیرش آنهاست برای دیدن جزییات به فایل txt. همراه دیتاست توجه کنید.

لینک زیر نیز راجب دیتاست می تواند مفید باشد :

http://archive.ics.uci.edu/ml/datasets/Nursery

خواسته ها :

1) دیتاست به دو صورت فایل csv. و هم mat. در اختیار شما قرار گرفته است، دقت کنید این دو فایل یکی هستند و مجاز به استفاده از هرکدام هستید.(فایل csv. جهت برخورد شما با فرمت دیتاست های واقعی است و با ابزار ImportData متلب به راحتی به mat. تبدیل میشود.) دیتاست همانند تمرین قبل به صورت یک table در اختیار شما قرار داده شده است، در این مرحله برای سهولت کار در مراحل بعدی آن را به وسیله MATLAB به یک ماتریس عددی تبدیل کنید، در واقع به هر عضو فضای حالت یک ویژگی و اسم کلاس ها یک عدد طبیعی نسبت دهید. (تناظر ها را بدانید چون در ادامه به اسم کلاس ها نیاز خواهید داشت.)

   راهنمایی : استفاده از ابزارهایی مانند find ، unique و دسترسی به اعضای جدول توسط اپراتور {} بسیار به شما کمک خواهد کرد.

2) سپس لازم است دیتاست را به دو بخش مساوی آموزش (train) و تست تقسیم کنید، برای اینکار تابع randperm مفید است. هر sample باید فقط به یکی از بخش های آموزش و یا تست تعلق داشته باشد. از دیتاست آموزش برای ساختن مدل و از دیتاست تست برای ارزیابی نتایج استفاده خواهید کرد. دقت کنید باید به صورت تصادفی این تقسیم رخ دهد تا داده ها نامتوازن تقسیم نشوند.
3) حال لازم است به صورت مناسب احتمال های لازم (به معادله I دقت کنید) را با استفاده از داده های آموزش استخراج کنید و درماتریس های مناسب ذخیره کنید.
   مطالعه لینک زیر مفید است :

   https://www.geeksforgeeks.org/naive-bayes-classifiers/

4) در این قسمت با استفاده از توضیحات داده شده، احتمالات محاسبه شده در قسمت ج و داده های تست برای داده های تست کلاس پیش بینی کنید و با کلاس های اصلی مقایسه کنید و درصدهای صحت دسته بندی خود را در Confusion Matrix بیان کنید.

5) (اختیاری) آیا با ایده کاهش ابعاد و PCA میتوانید نتایج کلاسیفیکشن خود را بهبود دهید؟ اگر بله الگوریتم پیشنهادی خود را پیاده سازی کرده و با نتایج بخش قبل مقایسه کنید.

سوال سوم : آشنایی با رگرسیون (Regression)

در مسله قبلی میخواستیم با مشاهده‌ی نمونه های متغییرهای تصادفی هر نمونه را به یک کلاس نسبت دهیم، حال اگر بخواهیم یک ویژگی پیوسته راجب داده ها از ویژگی ها استخراج کنیم با مسله رگرسیون روبه‌رو خواهیم بود؛ برای مثال اگر بخواهیم میزان درآمد افراد را براساس شرایط زندگی آنها تخمین بزنیم، درآمد افراد یک متغییر پیوسته است و باید از رگرسیون استفاده کنیم.

مطالعه ی لینک زیر می‌تواند مفید باشد :

https://www.statisticssolutions.com/how-to-conduct-multiple-linear-regression/

در این سوال با دیتاست forest fires کار خواهید کرد، با استفاده از ویژگی های داده شده مساحت جنگل های سوخته شده درآتش‌سوزی را تخمین می زنید. اطلاعات درباره ی دیتاست در فایل txt. پیوست شده داده شده است.

خواسته ها :

أ) دیتاست را مشابه مسله قبلی به مقادیر عددی مناسب تبدیل کنید.
ب) داده ها را به دو بخش train و test تقسیم کنید.
ج) با استفاده از تابع fitlm روی داده های آموزش (train) مدل خطی فیت کنید تا متغییر area را برحسب سایر متغییرها تخمین بزنید. با استفاده از نتیجه تحلیل کنید کدام ویژگی ها تاثیر بیشتری در پیش بینی مساحت سوخته شده دارند. درباره pvalue های خروجی تحقیق کنید.
د) درباره ی Rsquared تحقیق کنید و بیان کنید این پارامتر چه چیزی را درباره ی مدل فیت شده بیان میکند.
ه) برای داده های آموزش با استفاده از متد Fitted. مدل خطی ساخته شده مساحت آتش سوزی پیش یینی شده توسط مدل برای هر سمپل را به دست آورده و با استفاده از مقادیر اصلی خطا (MSE) را محاسبه کنید.
و) برای داده های تست با استفاده از متد predict. مدل خطی ساخته شده مساحت آتش سوزی شده را تخمین بزنید و با استفاده از مقادیر اصلی خطا (MSE) را محاسبه کنید.