به نام خدا

دانشگاه صنعتی شریف - دانشکده مهندسی برق

آمار و احتمال مهندسی – گروه ۴

دکتر میرمحسنی

نیمسال دوم ۹۷-۹۸

تمرین سری دو MATLAB

مهلت تحویل ۹۸/۳/۱۰

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 HW3_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 – تولید متغیر تصادفی

قضیه : فرض کنید متغیر تصادفی X از توزیع تجمعی دلخواه $F_X(x) = P\{X \leq x\}$ پیروی می کند. اگر متغیر تصادفی U را به صورت $U = F_X(x)$ تعریف کنیم، U توزیع یکنواخت بین صفر و یک دارد.

این قضیه را ثابت کنید و با توجه به آن روشی برای تولید متغیر تصادفی X با توزیع دلخواه $f_X(x)$ پیشنهاد دهید.

الف: تولید متغیر تصادفی نمایی

با کمک قضیه بالا یک تابع بنویسید که برداری به طول n تولید کند که هر درایه آن به طور مستقل توزیع نمایی داشته باشد:

$$f_X(x) = \frac{1}{\lambda} e^{-\frac{x}{\lambda}}$$

سپس نمودار pdf خروجی تابع را به ازای $n=10^5$ رسم کنید.

ب: تولید متغیر تصادفی رایلی

فرض کنید متغیر تصادفی X توزیع نمایی با پارامتر ۱ دارد. ثابت کنید متغیر تصادفی Y از توزیع رایلی با پارامتر σ پیروی می کند.

$$Y = \sigma\sqrt{2X} \ \rightarrow Y \sim Rayleigh(\sigma). f_Y(y) = \frac{y}{\sigma^2}\exp(-\frac{y^2}{2\sigma^2})$$

حال با کمک مطلب بالا تابعی بنویسی که برداری به طول n تولید کند که هر درایه آن مستقل از بقیه درایه ها از توزیه رایلی با پارامتر σ پیروی کند. سپس نمودار pdf خروجی را به ازای $n=10^5$ رسم کنید.

سوال دوم – آشنایی با کواریانس و PCA

PCA (Principal Component Analysis) روشی است که برای کاهش ابعاد داده‌های بزرگ استفاده می‌شود تا تعداد زیادی از ویژگی‌های یک مجموعه به تعداد کمتری تبدیل شود و در عین حال بخش قابل توجهی از اطلاعات را در خود جای داده باشد. به این ترتیب تحلیل اطلاعات ساده‌تر و سریع‌تر خواهد شد.

برای کاهش ابعاد مراحل زیر را طی می‌کنیم:

قدم اول : استاندارد کردن داده‌ها و یکی کردن محدوده متغیرهای مختلف است تا ویژگی‌های مختلف مقایسه‌پذیر باشند. (میانگین هر ویژگی را صفر و واریانس آن را یک قرار می‌دهیم.)

قدم دوم : محاسبه ماتریس کوواریانس

قدم سوم : پیدا کردن مولفه‌های اساسی (PC) ها که به ترتیب در راستای آن‌ها واریانس داده‌ها ماکسیمم شود. اثبات می‌شود که این راستاها بردارهای ویژه‌ی ماتریس کوواریانس هستند و مقدار ویژه‌ی متناظر با هر بردار ویژه بیانگر واریانس در آن راستا است.

بر اساس PC های به دست آمده می‌توانیم داده‌ها را در فضای جدیدی توصیف کنیم که محورهای مختصات جدید همان مولفه‌های خروجی PCA هستند. برای بردن داده‌ها به این فضای جدید باید ضرب داخلی داده و بردارهای ویژه (PCها) را حساب کنیم. (این کار معادل پیدا کردن تصویر هر یک از نقاط بر بردارهای ویژه است.)

برای جزییات بیشتر به فایل‌هایی که در CW قرار داده شده‌اند مراجعه کنید.

در این تمرین دیتاستی در اختیار شما قرار داده شده است. این دیتاست ارزش غذایی تعدادی ماده غذایی مختلف که متعلق به دسته‌های مختلف هستند به شما داده شده است. مراحل زیر را انجام دهید:

- **آشنایی با دیتاست**
  - دیتاست را در متلب لود کنید. تعداد نمونه‌ها، دسته‌ها و ویژگی‌ها را بررسی کنید.
  - برای ویژگی اول (Energy)، توزیع آن را رسم و سه گشتاور (moment) اول آن را حساب کنید.
  - دیتای ۴ دسته‌ی Vegetables and Vegetable Products، Beef Products، Nut and Seed Products، Baked Products را در فضای سه ویژگی اول رسم کنید. به هر دسته یک رنگ اختصاص دهید.

(توابعی که به شما برای این کار کمک می‌کند: unique، table2array، moment، scatter3 و ...)

- **پیاده‌سازی PCA**
  - ماتریس کوریلیشن داده‌ها را محاسبه کنید. از بین ویژگی‌هایی که دو به دو کوریلیشن بزرگتر از 0.9 دارند، فقط یک ویژگی را نگه دارید. (کوریلیشن بالا بین دو ویژگی، به معنی وابستگی زیاد بین آن‌هاست.)
  - بر روی دیتا تابع pca را اعمال کنید. در مورد خروجی‌های coeff، score، latent توضیح دهید.
  - خروجی explained چه چیزی را نشان می‌دهد؟ نمودار خروجی explained تابع pca را رسم کنید. توضیح دهید در مولفه‌های مختلف چه طور پراکنده شده است و کدام مولفه‌ها برای توصیف داده مناسب تر هستند.

- برای هر کدام از pcهای ۱ تا ۳، ۵ ویژگی اولی که بیشترین تاثیر را دارند را بیابید.
- برای ۳ ویژگی اول PCA داده‌های ۴ دسته‌ای که در بخش قبل تعیین شده‌اند را رسم کنید و با نمودار بخش قبل مقایسه کنید. چه تغییراتی مشاهده می‌کنید؟ توضیح دهید.