علی آراسته ۹۶۱۰۱۱۶۵

گزارش سری سوم تمرین کامپیوتری آمار و احتمال

سوال ۱:

فرض کنید متغیر تصادفی X از توزیع تجمعی دلخواه $F_X(x) = P\{X \le x\}$ پیروی می‌کند. از آن جایی که تابع cdf هر متغیر تصادفی پیوسته، تابعی اکیدا صعودی است، تابع $F_X(x)$ تابعی اکیدا صعودی است و در نتیجه $F_X^{-1}(x)$ قابل تعریف است.

حال اگر متغیر تصادفی U را به صورت $U = F_X(X)$ تعریف کنیم، متغیر تصادفی بدست آمده دارای توزیعی در بازه $[0,1]$ است؛ زیرا برد تابع $F_X(x)$، بازه $[0,1]$ است و در نتیجه احتمال این که متغیر تصادفی U خارج از بازه $[0,1]$ باشد، برابر صفر است.

حال cdf متغیر تصادفی U را در بازه $[0,1]$ محاسبه می‌کنیم.

$$F_U(u) = P\{U \le u\} = P\{F_X(X) \le u\}$$

چون تابع $F_X(x)$ تابعی اکیدا صعودی است، $F_X(X) \le u$ معادل با $X \le F_X^{-1}(u)$ است.

$$F_U(u) = P\{F_X(X) \le u\} = P\{X \le F_X^{-1}(u)\} = F_X\left(F_X^{-1}(u)\right) = u$$

در نتیجه متغیر تصادفی U دارای توزیع یکنواخت در بازه $[0,1]$ است.

فرض می‌کنیم متغیر تصادفی U دارای توزیع یکنواخت در بازه $[0,1]$ است و متغیر تصادفی X، متغیر تصادفی با cdf به صورت $F_X(x)$ باشد، که می‌خواهیم با استفاده تابع φ از متغیر تصادفی U بدست آوریم.

$$X = \varphi(U)$$

فرض می‌کنیم وجود دارد $x_0 < x_1$ به طوری که $F_X(x_0) = 0$ و $F_X(x_1) = 1$ باشد و برای هر a و b به صورت $x_0 \le a < b \le x_1$، $F_X(a) < F_X(b)$ است.

فرض می‌کنیم $x = \varphi(u)$ برای $0 \le u \le 1$ تابعی اکیدا صعودی باشد و $\varphi(0) = x_0$ و $\varphi(1) = x_1$ است.

بنابراین برای $x_0 \le x \le x_1$ تابع $u = \varphi^{-1}(x)$ اکیدا صعودی است.

$$P\{X \le x\} = P\{\varphi(U) \le x\} = P\{U \le \varphi^{-1}(x)\}$$
$$F_X(x) = F_U(\varphi^{-1}(x))$$

از آن جایی که $u = \varphi^{-1}(x)$، $0 \le \varphi^{-1}(x) \le 1$ و برای $0 \le u \le 1$، $F_U(u) = u$ است.

$$F_X(x) = \varphi^{-1}(x)$$
$$\varphi(u) = F_X^{-1}(u)$$

در نتیجه تابع φ مورد نظر به صورت $F_X^{-1}(U)$ است.

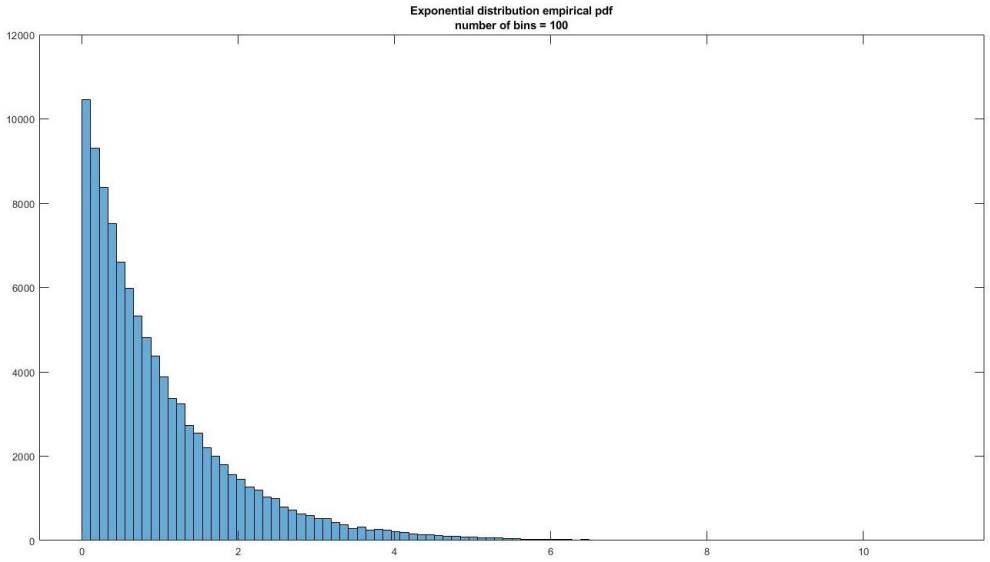
بنابراین روشی مناسب برای تولید متغیر تصادفی X با توزیع دلخواه $f_X(x)$، این است که ابتدا از روی تابع $f_X(x)$، تابع $F_X^{-1}(u)$ را بدست آوریم و سپس با تولید برداری به طول n با توزیع یکنواخت در بازه $[0,1]$ و اعمال تابع $F_X^{-1}(u)$ روی آن، برداری به طول n با توزیع دلخواه $f_X(x)$، بدست آوریم.

الف) تولید متغیر تصادفی نمایی:

متغیر تصادفی X دارای توزیع نمایی با پارامتر λ است:

$$X \sim \text{Exponential}(\lambda)$$
$$f_X(x) = \lambda e^{-\lambda x} \quad x > 0$$
$$F_X(x) = 1 - e^{-\lambda x} \quad x > 0$$
$$F_X^{-1}(u) = \frac{\ln\left(\frac{1}{1-u}\right)}{\lambda} \quad 0 \le u \le 1$$

با استفاده از روش گفته شده، تابعی برای تبدیل برداری به طول n با توزیع یکنواخت در بازه $[0,1]$ به برداری به طول n با توزیع رایلی می‌نویسیم و سپس با استفاده از دستور histogram متلب نمودار pdf آن را رسم می‌کنیم.



ب) تولید متغیر تصادفی رایلی:

متغیر تصادفی X دارای توزیع نمایی با پارامتر λ است:

$$X \sim \text{Exponential}(1)$$
$$f_X(x) = e^{-x} \quad x > 0$$

متغیر تصادفی Y تابعی از متغیر تصادفی X است:

$$Y = \sigma\sqrt{2X}$$

برای بدست آوردن $f_Y(y)$، ابتدا باید معادله $y = \sigma\sqrt{2x}$ را برای y ثابت حل می‌کنیم. از آن جایی که متغیر تصادفی X دارای توزیع نمایی و مقدار آن همواره مثبت است، معادله ذکر شده، برای y کوچکتر از صفر فاقد جواب است و برای y بزرگتر از صفر جواب یکتا دارد.

برای $y \geq 0$:

$$y = \sigma\sqrt{2x}$$
$$y^2 = \sigma^2(2x)$$
$$x = \frac{y^2}{2\sigma^2}$$

در نتیجه $f_Y(y)$ برای y کوچکتر از صفر، برابر صفر است.

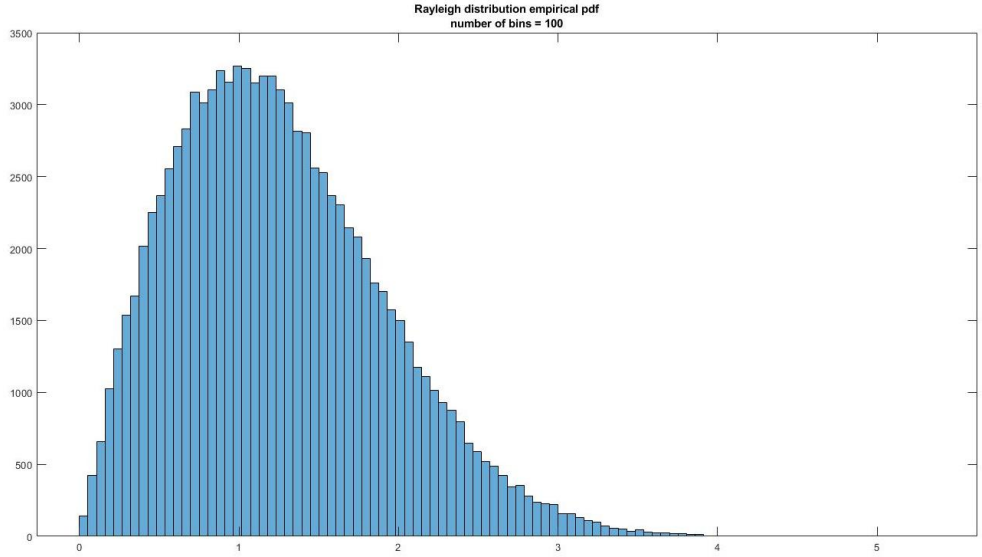
برای $y \geq 0$:

$$f_Y(y) = \sum_{i:\, \sigma\sqrt{2x_i}=y} \frac{f_X(x_i)}{\left|\frac{d}{dx}(\sigma\sqrt{2x})\right|} = \sum_{i:\, \sigma\sqrt{2x_i}=y} \frac{f_X(x_i)}{\frac{\sigma}{\sqrt{2x}}} = \frac{f_X\left(\frac{y^2}{2\sigma^2}\right)}{\frac{\sigma}{\sqrt{2(\frac{y^2}{2\sigma^2})}}} = \frac{e^{-\frac{y^2}{2\sigma^2}}}{\frac{\sigma^2}{y}} = \frac{y}{\sigma^2} e^{-\frac{y^2}{2\sigma^2}}$$

متغیر تصادفی Y دارای توزیع نمایی با پارامتر σ است:

$$Y \sim \text{Rayleigh}(\sigma)$$
$$f_Y(y) = \frac{y}{\sigma^2} e^{-\frac{y^2}{2\sigma^2}} \quad y > 0$$
$$F_Y(y) = 1 - e^{-\frac{y^2}{2\sigma^2}} \quad y > 0$$
$$F_X^{-1}(u) = \sigma\sqrt{2\ln(\frac{1}{1-u})} \quad 0 \leq u \leq 1$$

با استفاده از روش گفته شده، تابعی برای تبدیل برداری به طول n با توزیع یکنواخت در بازه $[0,1]$ به برداری به طول n با توزیع رایلی می‌نویسیم و سپس با استفاده از دستور histogram متلب نمودار pdf آن را رسم می‌کنیم.

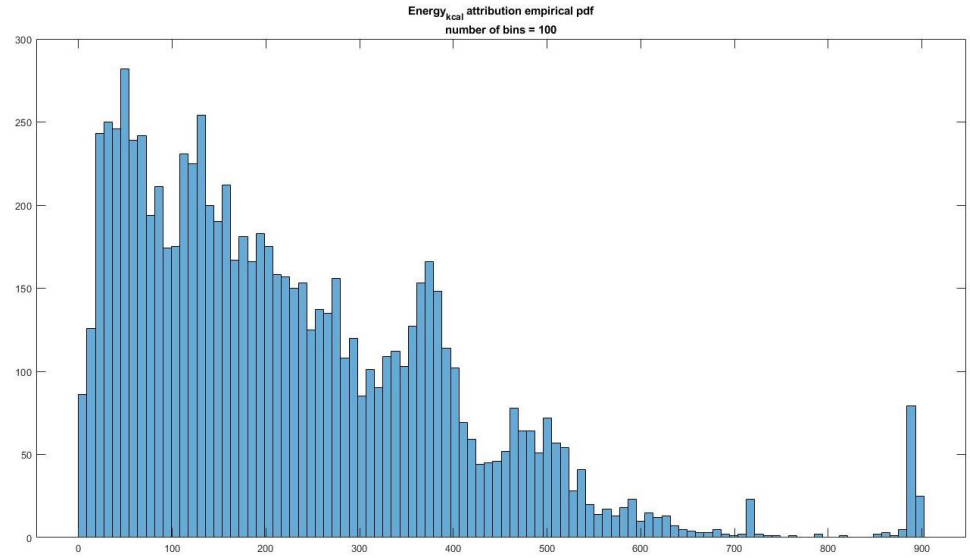


سوال ۲:

آشنایی با دیتاست

دیتاست حاوی اطلاعات مربوط به ۸۶۱۸ است. در این دیتاست نمونه ها در ۲۵ دسته، طبقه بندی شده اند و در مورد هر یک ۳۸ ویژگی اندازه‌گیری شده است.

با استفاده از دستور histogram متلب نمودار توزیع ویژگی اول (Energy) را رسم می‌کنیم.



$$\text{گشتاور مرتبه اول} = 226.4386$$

$$\text{گشتاور مرتبه دوم} = 7.9964 * 10^4$$

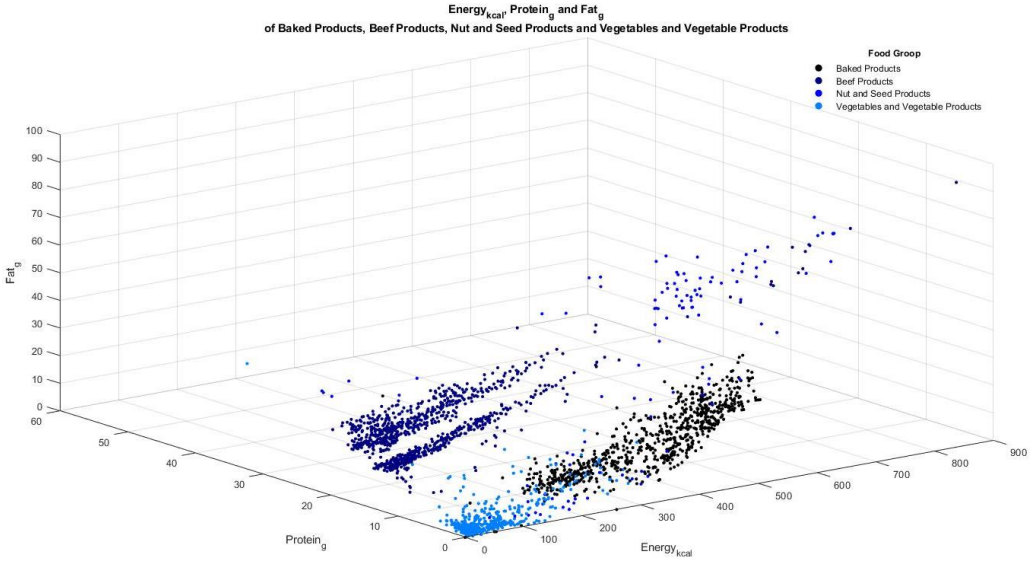
$$\text{گشتاور مرتبه سوم} = 3.6776 * 10^7$$

$$\text{گشتاور مرکزی مرتبه اول} = 0$$

$$\text{گشتاور مرکزی مرتبه دوم} = 2.8689 * 10^4$$

$$\text{گشتاور مرکزی مرتبه سوم} = 5.6759 * 10^6$$

با استفاده از دستور scatter3 متلب دیتای ۴ دسته‌ی Baked Products، Beef Products، Nut and Seed Products و Vegetables and Vegetable Products را در فضای سه ویژگی اول رسم می‌کنیم.



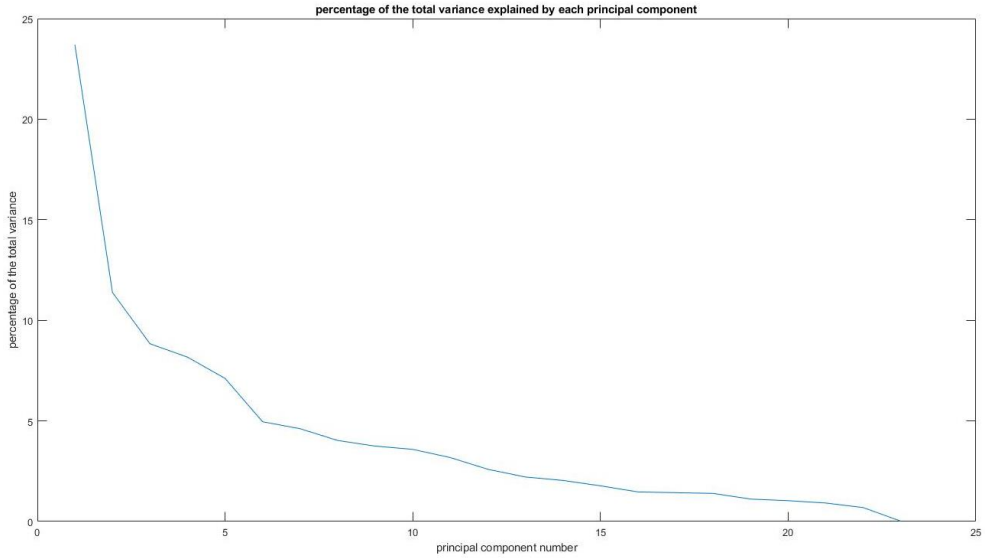
پیاده‌سازی PCA

خروجی coeff: این خروجی ضرایب مولفه‌های اساسی را بر می‌گرداند. هر ستون ماتریس coeff ضرایب یکی از مولفه‌های اساسی را تعیین می‌کند. ستون‌ها در ماتریس coeff از چپ به راست به ترتیب کاهش واریانس مولفه‌هایی اساسی قرار گرفته اند.

خروجی score: این خروجی ضرایب تصویر مقادیر ماتریس ورودی را بر فضایی که پایه‌های آن مولفه‌های اساسی هستند، بر می‌گرداند.

خروجی latent: این خروجی واریانس مولفه‌های اساسی را بر می‌گرداند. هر یک از مولفه‌های این خروجی یکی از مقدار ویژه‌های ماتریس کوواریانس متغیرهای ماتریس ورودی است. سطرها در بردار ستونی latent از بالا به پایین به ترتیب کاهش واریانس مولفه‌هایی اساسی قرار گرفته اند.

خروجی explained: این خروجی درصدی از واریانس کل را که توسط هر یک از مولفه‌های اساسی بیان می‌شود، را بر می‌گرداند.



با توجه به نمودار اکثر واریانس در مولفه‌های ابتدایی پراکنده شده است. با افزایش شماره مولفه اساسی، درصدی از واریانس کل که توسط آن مولفه اساسی بیان می‌شود، به سرعت کاهش می‌یابد. بنابراین مولفه‌های ابتدایی برای توصیف داده‌ها مناسب تر هستند.

با توجه به خروجی coeff داریم:

برای PC اول به ترتیب ویژگی‌های Riboflavin، Niacin، VitB6، Iron و Folate بیشترین تاثیر را دارند.

برای PC دوم به ترتیب ویژگی‌های Carb، Sugar، VitB12، Protein و Energy بیشترین تاثیر را دارند.

برای PC سوم به ترتیب ویژگی‌های Fat، Energy، Phosphorus، Folate و Protein بیشترین تاثیر را دارند.

با استفاده از دستور scatter3 متلب برای سه ویژگی اول PCA داده‌های ۴ دسته‌ی Baked Products، Beef Products، Nut and Seed Products و Vegetables and Vegetable Products را رسم می‌کنیم.