سوال چهارم

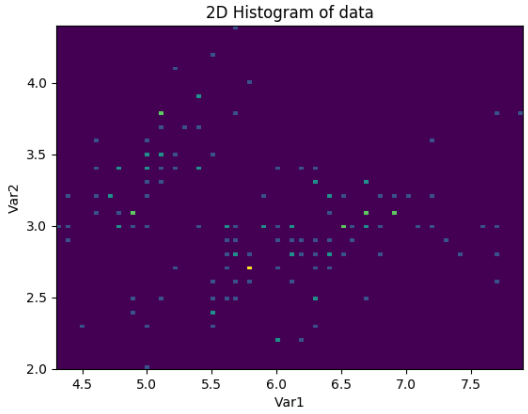
به کمک داده‌های آموزش IRIS که از طریق لینک زیر قاب دسترس است، هر یک از موارد زیر را تکمیل کنید.

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

(a

هیستوگرام داده‌ها را به کمک استفاده از کتابخانه Matplotlib در یک مختصات دوبعدی رسم نمایید و خروجی را به گزارش خود اضافه کنید. توجه کنید که تعداد ویژگی‌ها بیشتر از دو می‌باشد که در این صورت ناچار خواهید بود دو ویژگی را به دلخواه انتخاب کنید.

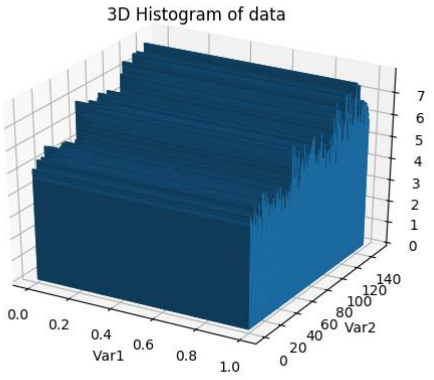
متغیرهای انتخابی برای این بخش از تمرین، ویژگی‌ها اول و دوم دیتا هستند. شکل زیر نیز هیستوگرام دوبعدی داده‌ها را نشان می‌دهد.



(b

مرحله قبل را با استفاده از کتابخانه Matplotlib و با کمک کلاس Axes3D در یک مختصات سه بعدی تکرار کنید.

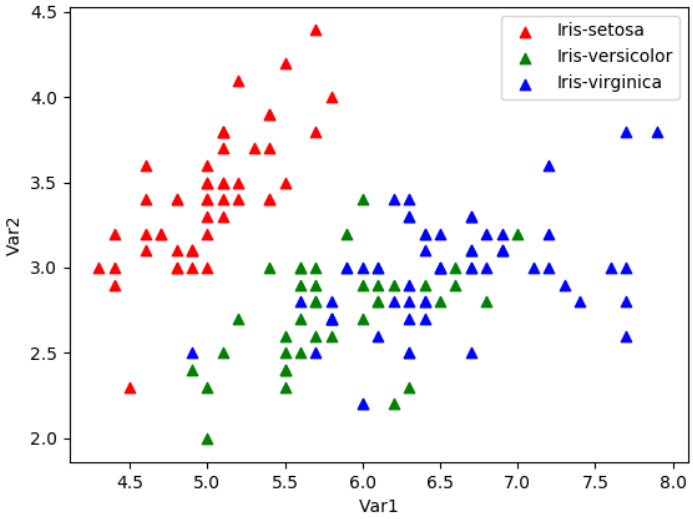
برای این بخش از تمرین، دو ویژگی اول دیتاست را انتخاب کرده‌ایم. شکل زیر نمودار هیستوگرام سه بعدی دو ویژگی انتخابی را نشان می‌دهد که از دو ویژگی برای تولید مش استفاده کرده‌ایم و سپس مقادیر دو ویژگی انتخابی را با دستور flatten در یک ویژگی مسطح کرده‌ایم و به بعد سوم نمودار اضافه نمودیم.



(c

داده‌ها 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.

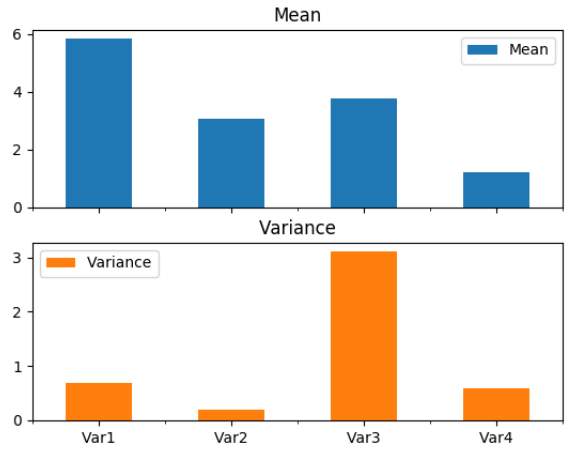
برای این قسمت از scatterplot و از دو ویژگی ابتدایی دیتاست استفاده کرده‌ایم. از طرفی در این پلات سعی کردیم تا داده‌های هر کلاس را بر اساس برچسب کلاس، به صورت رنگ‌های مختلف نشان دهیم تا پراکندگی داده‌ها در این ابعاد مشخص باشند. با توجه به شکل زیر متوجه می‌شویم که داده‌های کلاس قرمز رنگ در این دو بعد، به صورت بسیار خوبی از سایر کلاس‌ها متمایز هستند، اما دو کلاس دیگر دارای همپوشانی هستند که برای دسته‌بندی آن‌ها باید سایر ابعاد مسئله را نیز در نظر داشت.



(d

به کمک کتابخانه Numpy یا Pandas ، مقدار میانگین و واریانس داده‌ها را به دست آورید و نتیجه را به گزارش خود اضافه کنید. دقت کنید که با بیشتر از یک ویژگی در این داده‌ها سروکار دارید، در نتیجه مقدار میانه و واریانس در هر بعد باید محاسبه شود و به صورت یک بردار در خروجی نمایش داده شود.

با استفاده از کتابخانه Pandas ابتدا اقدام به محاسبه میانگین و واریانس کرده و سپس در دو پلات این موارد را نشان می‌دهیم. چون تعداد ویژگی‌ها 4 عدد است، در نتیجه 4 مقدار به ازای هر شاخص خواهیم داشت.



(e

به دلخواه خود دو کلاس را انتخاب کرده و ماتریس کوواریانس آن دو را به دست آورید و خروجی را به گزارش خود اضافه کنید. دقت کنید که ماتریس کوواریانس در این حالت یک ماتریس مربعی دو در دو خواهد بود.

نتیجه خروجی ماتریس کوواریانس به صورت زیر است.

[[3.4135788944723644, 2.3860713567839187]

, [2.3860713567839187, 3.106001005025126]]

برای محاسبه ماتریس کوواریانس از رابطه زیر استفاده کرده‌ایم. مسلم است که کوواریانس $\text{cov}(A,B)$ و $\text{cov}(B,A)$ باید یکی شود.

$$\text{cov}(A,B) = \frac{1}{N-1}\sum_{i=1}^{N}(A_i-\mu_A)^*(B_i-\mu_B)$$

$$C = \begin{pmatrix} \text{cov}(A,A) & \text{cov}(A,B) \\ \text{cov}(B,A) & \text{cov}(B,B) \end{pmatrix}.$$

(f

به دلخواه خود دو کلاس را انتخاب کرده و ماتریس همبستگی آن دو را به دست آوردید 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 نمایید.

برای محاسبه مقادیر همبستگی کافی است تا ماتریس کوواریانس را تقسیم بر انحرافات معیار دو کلاس مورد نظر کنیم

$$\rho(A,B) = \frac{\text{cov}(A,B)}{\sigma_A\sigma_B}.$$

خروجی به دست آمده به صورت زیر خواهد بود.

[[1. 0.73646918]

[0.73646918 1.]]

تحلیل:

ماتریس همبستگی همان ماتریس کوواریانس است ولی ماتریس همبستگی به نوعی استاندارد شده ماتریس کوواریانس است. ماتریس کوواریانس هر مقداری بین منفی بی‌نهایت و مثبت بی‌نهایت می‌تواند داشته باشد اما ماتریس همبستگی حداکثر مقدار 1 و حداقل مقدار -1 را خواهد داشت. و به نوعی ماتریس کوواریانس میزان تغییر دو متغیر تصادفی را بیان می‌کند ولی ماتریس همبستگی ارتباط بین آن دو متغیر را نشان می‌دهد. شاید در ماتریس کوواریانس تنها می‌توانستیم بگوییم که ارتباط بین دو کلاس به صورت خطی مستقیم است اما نمی‌توانستیم میزان این ارتباط را بیان کنیم. اما در ماتریس همبستگی با توجه به اینکه مقدار حدود 0.7 را برای ارتباط بین این دو کلاس داریم، پس می‌توان گفت که ارتباط خطی بین این دو کلاس ، مستقیم و ارتباط تقریبا قوی بین این دو کلاس برقرار است زیرا مقدار 0.7 به مقدار 1 نزدیک‌تر است تا مقدار صفر. مقدار صفر در ماتریس همبستگی نشانگر مستقل بودن دو متغیر تصادفی است.

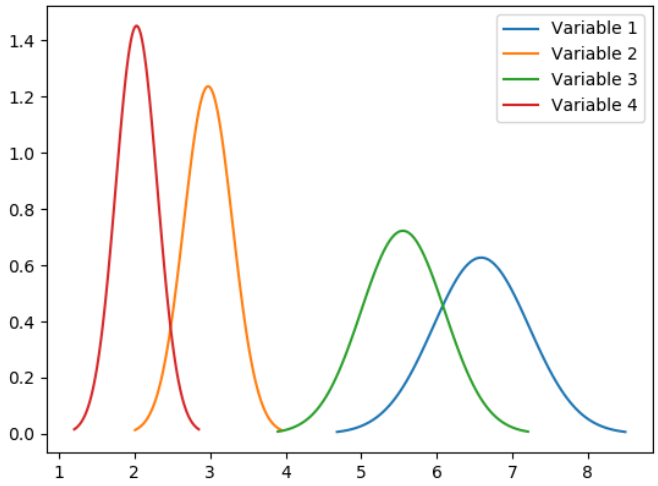
(g

در گیاه ویرجینیکا کدام دو ویژگی شباهت بیشتری دارند؟ چرا ؟

برای این سوال می‌توان از موضوعات مختلفی برای جواب استفاده کرد.

مثلا می‌توان از شکل توزیع ویژگی‌ها استفاده کرد. با فرض اینکه ویژگی‌ها دارای توزیع گاوسی باشند، می‌توان با ترسیم نمودار آن‌ها بر حسب میانگین و واریانس هر ویژگی ، مشابهت ویژگی‌ها را با استفاده از این توزیع‌ها تشخیص داد.

به عنوان مثال در شکل زیر که توزیع چهار ویژگی رسم شده است، مشخص است که توزیع متغیرهای 1 و3 نسبت به یکدیگر مشابه هستند و توزیع متغیرهای 2و 4 نیز مشابه به هم.



از طرفی می‌توان از ماتریس همبستگی نیز استفاده کرد. ماتریس همبستگی بین ویژگی‌ها را به دست می‌آوریم

[[1. 0.45722782 0.86422473 0.28110771]
[0.45722782 1. 0.40104458 0.53772803]
[0.86422473 0.40104458 1. 0.32210822]
[0.28110771 0.53772803 0.32210822 1.]]

طبق این ماتریس، ویژگی 1 و 3 با مقدار0.86 نسبت به هم بیشترین مشابهت را دارند که در شکل توزیع‌ها نیز مشاهده کردیم. سپس ویژگی 2 و4 با مقدار 0.53 بیشترین شباهت را دارند که شباهت این دو متغیر نیز در شکل توزیع‌ها ذکر شد.

(h

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

برای این کار بهتر است که از شاخص بهره اطلاعات یا اطلاعات متقابل استفاده کنیم. ابزار mutual Information در کتابخانه‌های یادگیری ماشین پایتون موجود است و می‌توان از آن استفاده کرد. این ابزار به ما می‌گوید که ارتباط بین ویژگی‌ها و هدف (برچسب کلاس‌ها) تا چه حد است.

خروجی این تابع به صورت زیر است

[0.48467537 0.19982519 0.99096098 0.99312168]

با توجه به خروجی به دست آمده، مشاهده می‌کنیم که برای دسته‌بندی داده‌ها بهتر است از ویژگی‌ها سوم و چهارم استفاده کرد که تقریبا برای دسته‌بندی کلاس‌ها مناسب‌تر هستند. البته با اختلاف بسیار کمی، ویژگی چهارم می‌تواند بهترین ویژگی باشد.