به نام خدا

دانشگاه صنعتی امیرکبیر

(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گزارش پروژه سوم (مدلهای احتمالاتی گرافی)

درس یادگیری ماشین آماری

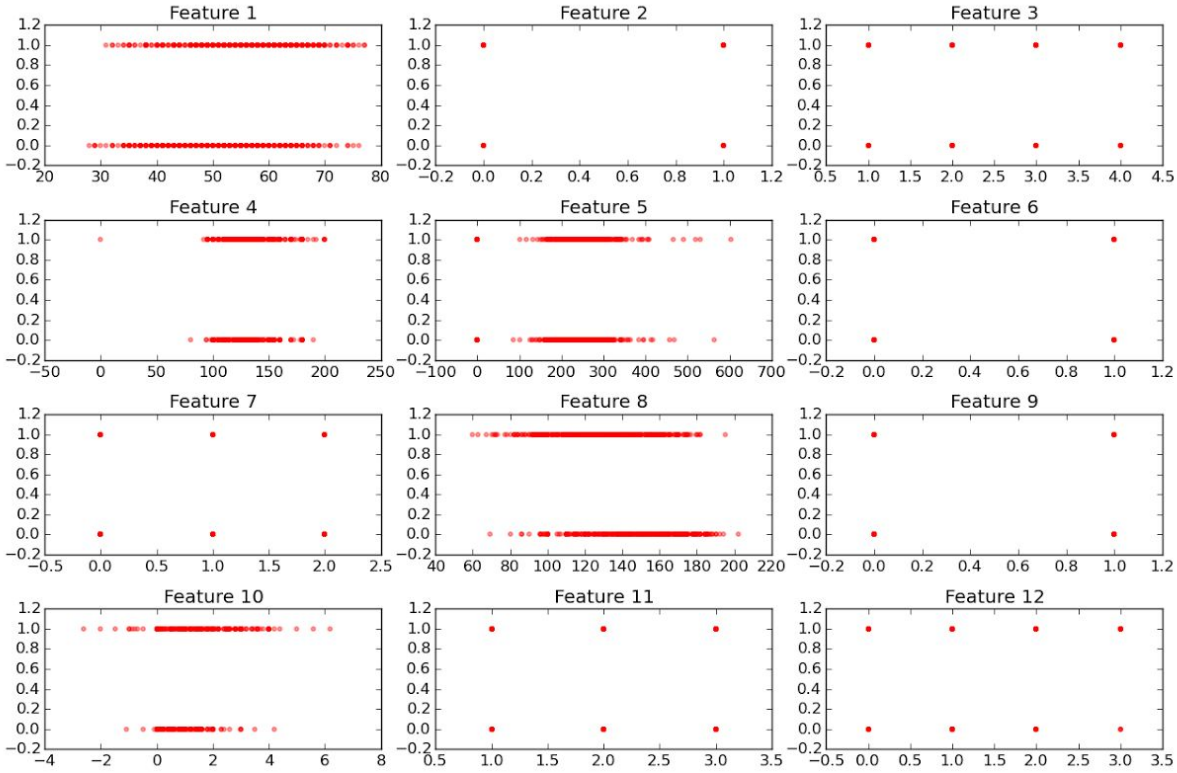
مهدی طاهراحمدی ۹۲۳۱۰۴۲

استاد: دکتر نیک آبادی

الف و ب)

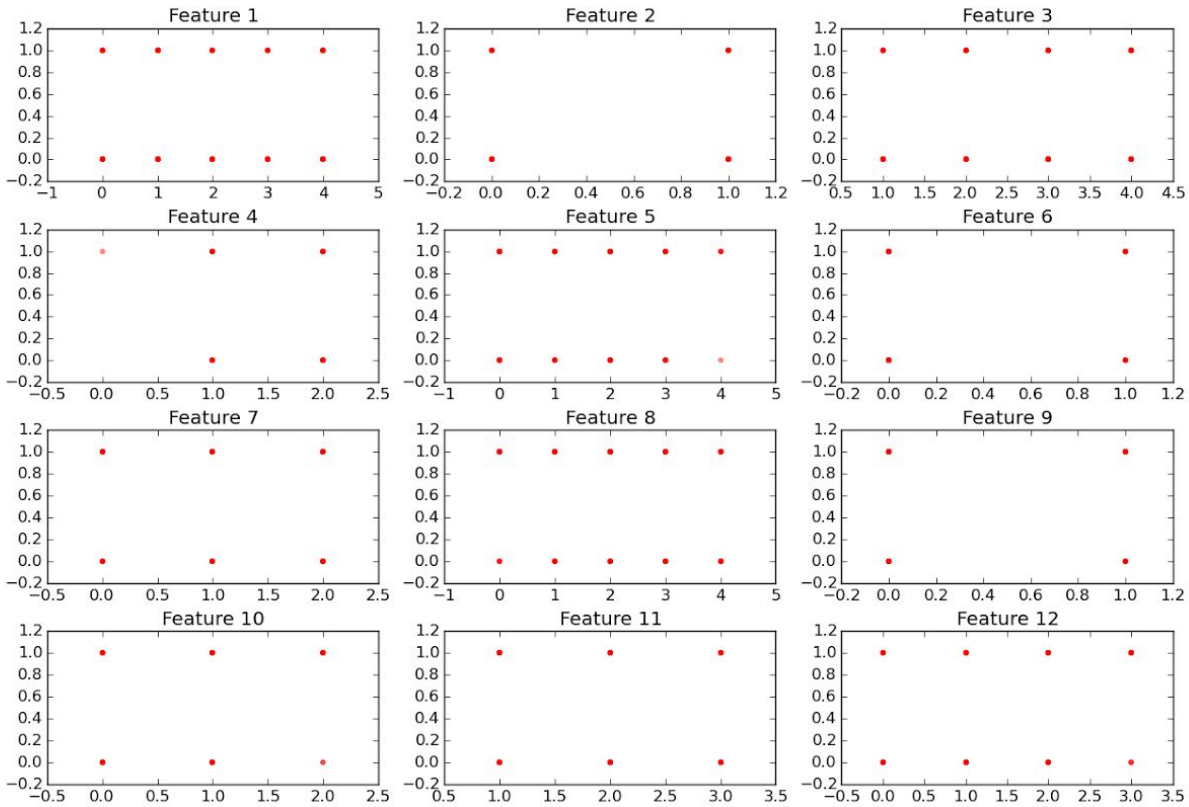
در ابتدا لازم است تا پیش پردازش هایی روی مجموعه داده ها انجام شود. ابتدا ۴ مجموعه داده مرتبط به شهر های مختلف را ادغام میکنیم. در مرحله اول، داده های نامشخص را با مقدار میانگین هر ستون پر کرده ایم. اینکار به سادگی با توابع موجود در کتابخانه pandas انجام می‌شود.

سپس نمودار scatter آن را رسم کردیم.

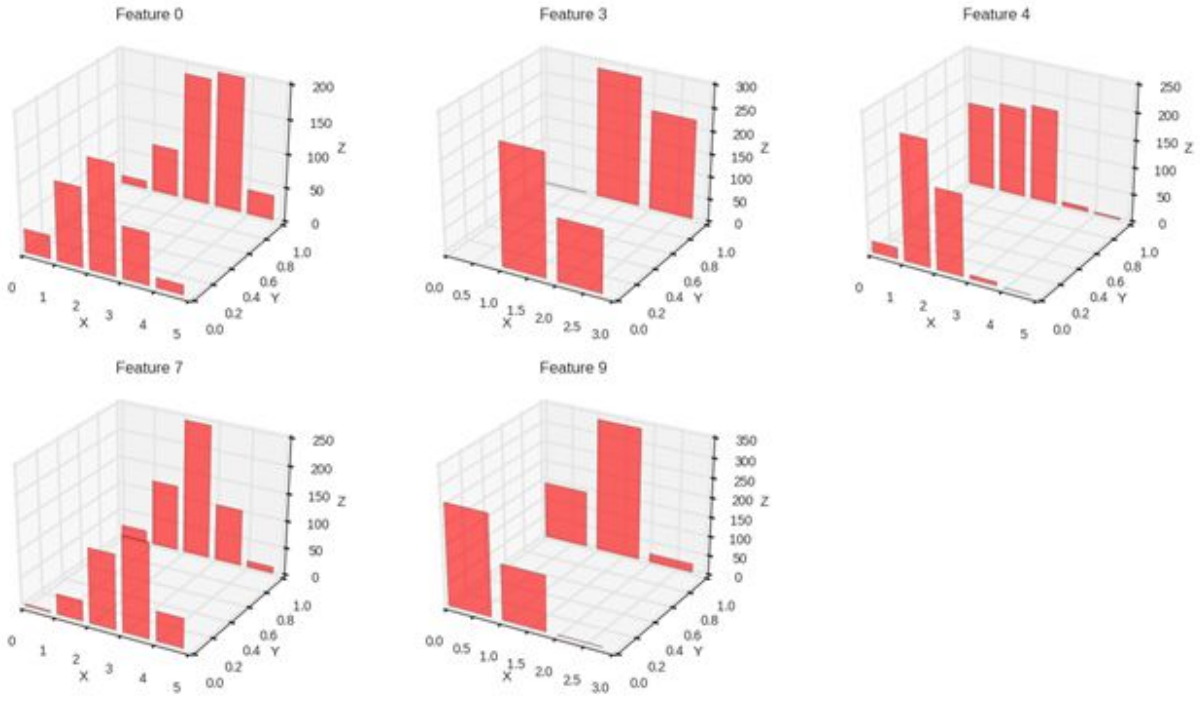


همانطور که مشاهده میشود، بعد از پر کردن مقادیر نامشخص با میانگین ستون ها، داده های بعضی از ستون ها، مقادیر گسسته دارند. پس گسسته سازی را روی داده های پیوسته را انجام می‌دهیم. این کار را برای ویژگی های ۰، ۳، ۴، ۷ و ۹ انجام می‌دهیم. این ویژگی ها به ترتیب به ۵، ۳، ۵، ۵ و ۳ کلاس تبدیل شده اند. این تعداد بر مبنای انحراف معیار مقادیر این ویژگی ها و همچنین اطلاع در مورد ویژگی و تعداد حالت های ممکن آن با استفاده از دانش پزشکی انتخاب شده است.

نمودار scatter داده ها پس از گسسته سازی به این شکل در می‌آید:



از روی این نمودار scatter اطلاعات مفیدی بدست نمی آید زیرا فراوانی هر مقدار گسسته مشاهده نمیشود. پس از نمودار میله ای سه بعدی استفاده میکنیم:



تحلیل سه آماره از داده ها به شرح زیر است:

	0	1	2	3	4	5	6	7	8	9	10	11	12
mean	53.51	0.78	3.25	132.1	199.1	0.16	0.6	137.5	0.38	0.88	1.7	0.67	5.08
min	28	0	1	0	0	0	0	60	0	2.6-	1	0	3
max	77	1	4	200	603	1	2	202	1	6.2	3	3	7

مشاهده می‌شود که داده های ستون ۱، ۲، ۵، ۶، ۸، ۱۰، ۱۱، ۱۲ دارای مقادیر گسسته بودند ولی بعد از اینکار یک مقدار به range داده های گسسته اضافه شد. پس برای پر کردن داده های نامشخص آن ها از مد استفاده می‌کنیم. برای بقیه ستون ها همچنان از میانگین استفاده می‌کنیم.

در مرحله بعد، روی داده برچسب کار میکنیم. برچسب های این مجموعه داده که همان ویژگی ۱۴ ام است، ۵ مقدار مختلف می‌تواند اختیار کند، ولی بنابر تعریف پروژه این مقادیر را به دو کلاس ۱ یعنی "وجود بیماری" و ۰ به معنی "عدم وجود بیماری" تقسیم می‌کنیم.

تحلیلی که می‌توان ارائه کرد، به طور کلی این است که تعداد تکرار متغیر هدف(۰ یا ۱) با تغییر مقدار ویژگی مورد بررسی رابطه دارد، به طور مثال در ویژگی شماره ۲ مشاهده می‌شود که با افزایش مقدار

ویژگی مورد بحث، مقدار متغیر هدف با احتمال بیشتری ۱ است که رابطه ای منطقی است، چرا که ویژگی ۲ مربوط به مقدار کلسترول خون است و می‌دانیم با افزایش کلسترول احتمال رخ دادن بیماری قلبی افزایش می‌یابد.
یک نتیجه گیری دیگر مقدار همبستگی کلی است. به طور مثال در ویژگی ۵ با تغییر مقدار ویژگی، تغییری در احتمال ۰ یا ۱ بودن متغیر هدف به وجود نیامده است. یعنی ویژگی ۵ همبستگی زیادی با متغیر هدف ندارد.
برای ویژگی ۸ می‌توان نتیجه گرفت افزایش مقدار ویژگی، احتمال ۰ شدن متغیر هدف را به شدت کاهش می‌دهد.

ج)
برای پیاده سازی مدل بیز ساده، که روابط آن به شرح زیر است:

$$f(\theta|x^n) = \frac{f(x^n|\theta)f(\theta)}{\int f(x^n|\theta)f(\theta)d\theta} = \frac{\mathcal{L}_n(\theta)f(\theta)}{\int \mathcal{L}_n(\theta)f(\theta)d\theta} \propto \mathcal{L}_n(\theta)f(\theta). \tag{12.2}$$

The Naive Bayes Classifier

1. For each group k, compute an estimate $\widehat{f}_{kj}$ of the density f_{kj} for X_j, using the data for which $Y_i = k$.
2. Let
$$\widehat{f}_k(x) = \widehat{f}_k(x_1, \ldots, x_d) = \prod_{j=1}^{d} \widehat{f}_{kj}(x_j).$$
3. Let
$$\widehat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = k)$$
where $I(Y_i = k) = 1$ if $Y_i = k$ and $I(Y_i = k) = 0$ if $Y_i \neq k$.
4. Let
$$h(x) = \text{argmax}_k \, \widehat{\pi}_k \, \widehat{f}_k(x).$$

از تابع موجود در کتابخانه Sci-kit استفاده شده است. از آنجایی که متغیر ها گسسته اند، پس مدل Multinomial Naive Bayes برای هدف مورد نظر مناسب است. دقت مدل با استفاده از روش Leave One Out Cross Validation اندازه گیری شده است. این مقدار برابر 0.15 است.

د) مدل بیز ساده با دو زیر مجموعه مختلف از ویژگی ها:
ابتدا مدل را با مجموعه ویژگی های 0 1 2 آموزش می‌دهیم. دلیل انتخاب این ۳ ویژگی این است که اولا مدل ساده تر باشد، دوما از روی Bar Plot های رسم شده می‌توان نتیجه گرفت این ۳ ویژگی "آموزنده"[2] هستند. نتیجه حاصل به شکل زیر است:

Number of mislabeled points out of a total 920 points: 411
Feature set: [0, 1, 2]
LOOCV: 0.33

در مرحله بعد، به این مجموعه ویژگی ها، ویژگی 4 را اضافه می‌کنیم. این ویژگی هم یک ویژگی آموزنده است. انتظار می‌رود مدل بهبود بیابد. نتیجه همین طور است:
Number of mislabeled points out of a total 920 points: 391
Feature set: [0, 1, 2, 4]
LOOCV: 0.25
حال چند ویژگی غیر آموزنده[3] یا به زبان ساده "بد" را به مدل اضافه می‌کنیم. انتظار می‌رود خطا افزایش یابد. خروجی به شکل زیر است:
Number of mislabeled points out of a total 920 points: 350
Feature set: [0, 1, 2, 3, 4, 5]
LOOCV: 0.5

منابع:

[1].Larry Wasserman, All of Statistics: A Concise Course in Statistical Inference, Springer, 2010.
[2].http://www.uow.edu.au/student/qualities/statlit/module3/5.4interpret/index.html