سوال ④ تعاریف:

dimition : وسعت قابل اندازه گیری از یک نوع خاص را بُعد می گویند؛ مانند طول و عرض و ارتفاع

outlier : یک متغیر از dataset را "داده ی پرت" می گوییم که مقادیرش خیلی با دیگر متغیر ها متفاوت باشد.

independent var : متغیر مستقل متغیری است که وابستگی ای به دیگر متغیر ها ندارد بعنوان مثال در رابطه ی زیر y یک متغیر وابسته است : $y = 2x$

Sampling startified : برای انتخاب نمونه ای که نماینده ی گروه های مختلف باشد، از این نمونه گیری استفاده می کنیم

پیش پردازش : در داده کاوی به دسته کاری اطلاق داده هایش از استفاده، به منظور افزایش عملکرد، پیش پردازش می گوییم.

Scanned with CamScanner

سوال ۵

PCA یکی از تکنیک ها یا الگوریتم‌های کاهش ابعاد خطی است که مجموعه‌ای از متغیر های همبسته (p) را به تعداد k (که k<p) تا از متغیرهای غیرهمبسته (که همان principal component هستند) تبدیل می‌کند ولی در عین حال تا جایی که ممکن است. تغییرات را در مجموعه داده‌ولیه حفظ می‌کند. در ML این الگوریتم را تحت عنوان یک روش unsupervised برای کاهش بعد تشخیص داده می‌کنند

Scanned with CamScanner

برای مثال PCA را کاهش ابعاد iris flowers استفاده می‌کنیم:

کد این مثال را بازان پایتون نوشتم و screen shot آن را در صفحه بعدی قرار خواهم داد و خروجی آن را نیز قرار خواهم داد.

ابتدا داده‌ها را به دو دسته data و target تقسیم می‌کنیم؛ سپس با استفاده از تابع PCA (که آن را اول import کردیم) عملیات کاهش ابعاد را انجام می‌دهیم و پارامتری که این تابع دریافت می‌کند بیانگر همان تعداد مورد نظر ما بعد از عملیات کاهش بعد است.

حال همانطور که در کد مشخص است، اگر خروجی بگیریم، بعدها از ۴ به ۲ کاهش داده شدند و خروجی نیز توسط کتابخانه‌ی math plot نشان داده شده است.

Scanned with CamScanner

pca-example.py > ...

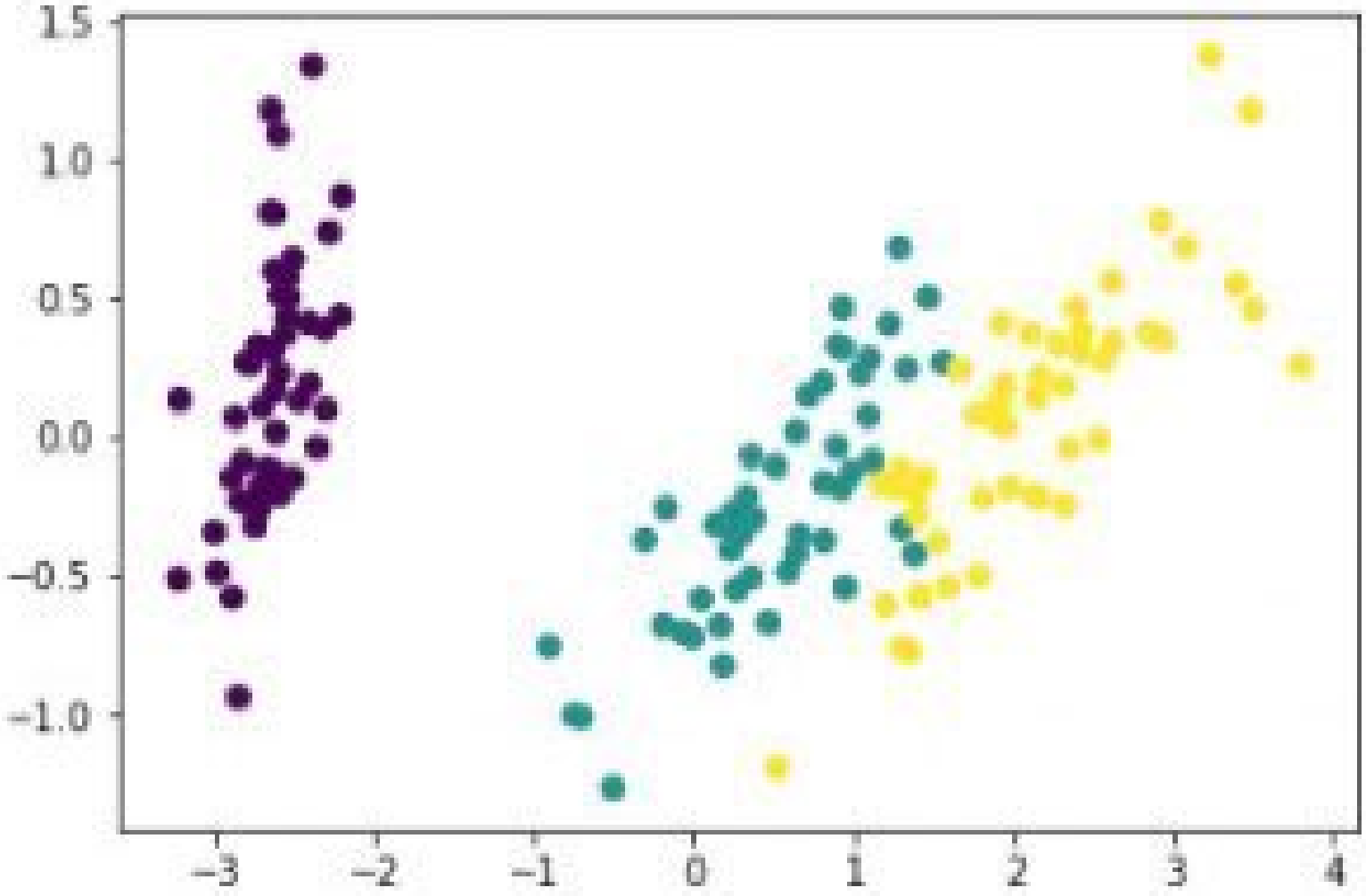
```python
from sklearn import datasets
from sklearn.decomposition import PCA
from matplotlib import pyplot as plt

iris = datasets.load_iris()
x = iris.data
y = iris.target

Dimr = PCA(n_components = 2)
Dimr.fit(X)
trans = Dimr.transform(X)

plt.scatter(trans[:,0],trans[:,1],c= y)
plt.show()
```

Scanned with CamScanner



Scanned with CamScanner

سوال ⑥ اینکه همبستگی بین دو متغیر صفر باشد به این معنات که بین متغیرها رابطه‌ی خطی وجود ندارد اما این به این معنایت که لزوما متغیرها مستقل باشند بلکه ممکن است وابسته باشند ولی رابطه‌ی بینشان خطی نباشد

Scanned with CamScanner

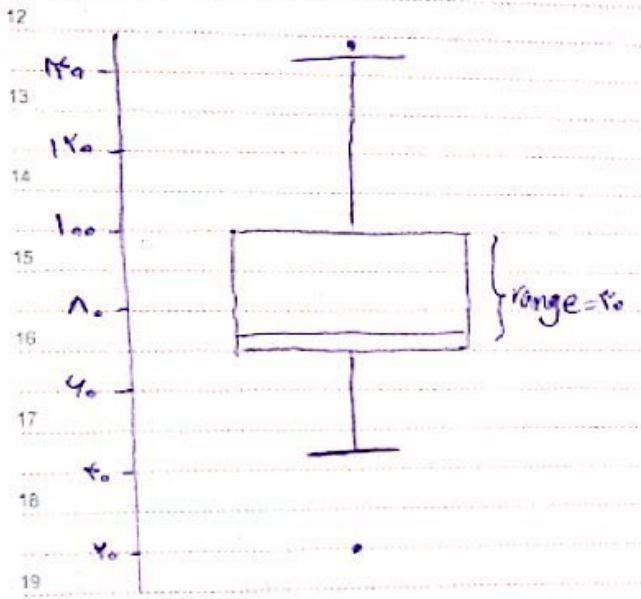
4 March 2013 • ۲۱ ربیع الثانی ۱۴۳۴

سوال ۷

نام روش	فرمول اصلی	عناصر برای بعد بالا	حساس به مقدیر بردار	بازه خروجی
Euclidean Distance	$\sqrt{\sum_{i=1}^{n}(A_i-B_i)^2}$	✗	✓	$[0,+\infty)$
Cosine Similarity	$\frac{A.B}{\|A\| \times \|B\|}$	✓	✗	$[-1,1]$
Hamming Distance	idk	✓	✗	$[0,+\infty)$
Manhattan Distance	$\sum_{i=1}^{n}\|A_i-B_i\|$	✓	✓	$[0,+\infty]$

Scanned with CamScanner

سوال ⑧



مقدار min = ۲۰

مقدار max = ۱۴۶

مقدار median = ۷۴

Scanned with CamScanner

سوال ۹

(a) در یک مدل regression هنگامی که دو متغیر مستقل بسیار با یکدیگر ارتباط داشته باشند multicolinearity اتفاق می‌افتد؟ در واقع نمی‌توانیم در یک مدل regression مقدار یک متغیر مستقل را از متغیر مستقل دیگر [redacted] پیش بینی کرد.

Scanned with CamScanner

امتیاز

2 March 2013 • ۱۹ ربیع الثانی ۱۴۳۴

$$Z = \frac{x - \mu}{\sigma}$$

μ ← mean جمعیت

σ ← sd جمعیت

b) فرمول استفاده از روش z-score به صورت مقابل است:

یک z-score موقعیت یک امتیاز را برحسب فاصله آن از میانگین زمانی که در واحدهای انحراف استاندارد اندازه‌گیری می‌شود، توصیف می‌کند به این صورت که اگر مقدار بالاتر از میانگین باشد، امتیاز z مثبت و اگر کمتر از میانگین باشد منفی است.

روش IQR نیز، ۵۰٪ مقادیر متوسط را هنگامی که از کمترین به بالاترین مرتب می‌شود توصیف می‌کند.

برای یافتن IQR، ابتدا میانه‌ی نیمه‌ی پایین و نیمه‌ی بالایی داده‌ها را پیدا می‌کنیم و:

$$IQR = Q_3 - Q_1$$

Q_3 ← چارک ۳ و Q_1 ← چارک ۱

c) روش اول) استفاده از mean/median مقادیر:

این روش به این صورت کار می‌کند که مقدار mean/median برای مقادیر non-missing محاسبه می‌کند و با هر مقدار missing values جایگزین می‌شود به صورت مستقل و هرا از هم در هر ستون

مزایا: + استفاده از آن آسان است و بسیار سریع

+ برای داده‌های عددی کوچک بسیار خوب کار می‌کند

معایب: - این روش دقیق نیست (و فقط برای داده‌های عددی مناسب است)

- نتایج ضعیفی برای داده‌های encode شده به دست می‌آورد

- همبستگی بین ویژگی‌ها را در نظر نمی‌گیرد و تنها در سطح ستون کار می‌کند

Scanned with CamScanner

روش دوم) استفاده از مقادیر پرتکرار یا zero/constant :

این نیز روش دیگری برای راهبردهای آماریت برای انتساب مقادیر miss شده. و این روش برای ویژگی های طبقه بندی نیز کاربرد دارد (یعنی هم داده های عددی یا string) و به این صورت کار می کند که داده های miss شده را با داده‌ی پرتکرار در همان column جایگزین می کند. مزایا: + با categorical features خوب کار می کند.

معایب: - این روش نیز ارتباط بین ویژگی ها را در نظر نمی گیرد

- ممکن است اشکال bias را در data ها ایجاد کند

(یعنی برخی element ها وزن بسیار زیادی در dataset یا داشته باشند)

روش سوم) استفاده از K-NN :

از این الگوریتم برای طبقه بندی های ساده استفاده می شود. این الگوریتم از شباهت ویژگی ها برای پیش بینی دیتاهای جدید در هر نقطه استفاده می کند. این روش برای پیش بینی کردن مقادیر miss شده می تواند بسیار پرکاربرد باشد.

مزایا: + با توجه به نوع dataset ها می تواند بسیار دقیق تر از دو روش قبلی عمل کند.

معایب: - این روش با استفاده از ذخیره کل traning data ها، هزینه محاسباتی زیادی را به‌بار میآورد.

- این روش به نقاط پرت بسیار حساس است.

Scanned with CamScanner