

فرادرس

فراتر از یک کلاس درس
www.faradars.org

آموزش یادگیری ماشین (Machine Learning)

(تئوری - عملی) - بخش دوم

درس هفتم: کشف داده‌های پرت

مدرس:

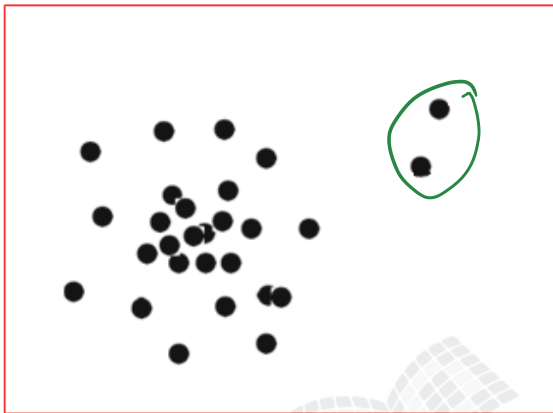
فرشید شیرافکن

دانشجوی دکترای بیو انفورماتیک

دانشگاه تهران

داده‌های پرت

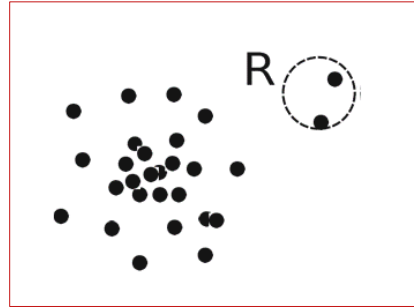
داده پرت: داده‌ای که به طور معنا داری از داده‌های طبیعی دور است.



نویز: یک خطای اتفاقی و تغییر کوچک است که به داده‌ها اعمال شده و آن‌ها را تا حدودی جابه‌جا می‌کند.

انواع داده‌های پرت

Types of Outliers

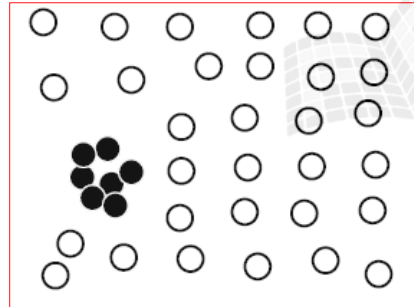


۱- سراسری (Global Outliers) ← پرت بودن نسبت به کل داده‌ها ←

۲- جمعی (Collective Outliers) ← پولسغوی

۳- زمینه‌ای (Contextual Outliers) ← پرت بودن در یک زمینه خاص

آیا دمای تهران پرت است؟
بله. زمستان جمعی



✓ یک مجموعه داده می‌تواند چند نوع داده پرت داشته باشد.

✓ یک داده می‌تواند متعلق به چند نوع داده پرت باشد.

روش‌های تشخیص داده‌های پرت

۱- با نظارت (Supervised)

کاربر تعدادی از نمونه‌ها را به عنوان داده‌های پرت معرفی می‌کند.

۲- بدون نظارت (Unsupervised)

اطلاع قبلی از داده‌های پرت نداریم و به کمک روش‌های زیر، آن‌ها را پیدا می‌کنیم.

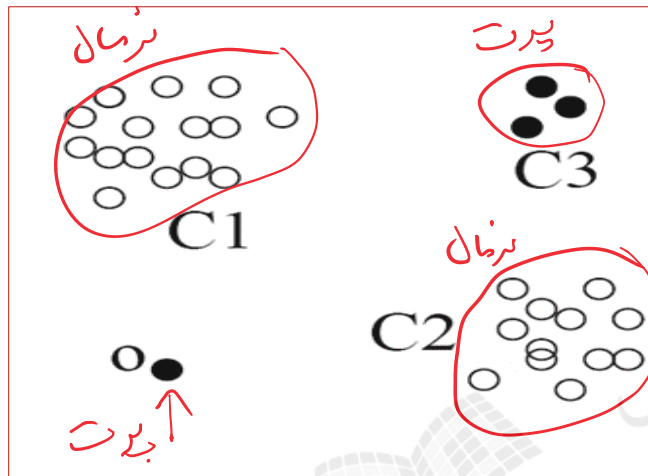
۱- مبتنی بر خوشه‌بندی (Clustering-Based)

۲- مبتنی بر نزدیکی (proximity-based)

۳- مبتنی بر دسته‌بندی (Classification-Based)

۴- آماری (Statistical)

روش‌های مبتنی بر خوشه‌بندی



به کمک خوشه‌بندی می‌توان داده‌های پرت را پیدا کرد.
داده‌های نرمال متعلق به خوشه‌های بزرگ و متراکم هستند.

داده‌های پرت:

- فاصله آن‌ها تا نزدیک‌ترین خوشه زیاد است یا
- در خوشه‌های کوچک و اسپارس هستند یا
- به هیچ خوشه‌ای متعلق نیستند.

مزایا و معایب روش‌های مبتنی بر خوشه‌بندی

مزایا:

- نیاز به داشتن برچسب نمونه‌ها نیست.
- نیاز نیست حتما داده‌ها عددی باشد و برای بسیاری از نمونه داده‌ها کار می‌کند.

معایب:

- وابسته به روش خوشه‌بندی است. (ممکن است نتواند داده‌های پرت را پیدا کند).
- برای داده‌های بزرگ مناسب نیست. (چون در ابتدا باید خوشه‌بندی انجام شود و هزینه زیاد می‌شود).

روش‌های مبتنی بر نزدیکی

۱- بر اساس فاصله (Distance based)

نمونه‌ای پرت است که تعداد همسایه‌های آن از یک حدی کمتر باشد.
مشکل: اثر بخشی این روش، وابسته به معیار نزدیکی استفاده شده دارد.

۲- بر اساس چگالی (Density-based)

نمونه‌ای پرت است که چگالی اطراف آن از نمونه‌های دیگر کمتر باشد.
مشکل: تعیین مناسب شعاع همسایگی و حداقل تعداد همسایه‌ها، چالش برانگیز است.

روش‌های مبتنی بر دسته‌بندی

در این روش از ابتدا برچسب نرمال و پرت بودن نمونه‌های آموزشی را می‌دانیم، پس می‌توان مدلی را برای بررسی نمونه‌های جدید، مشابه روش‌های کلاسه‌بندی آموزش دهیم.

چالش‌های پیش رو:

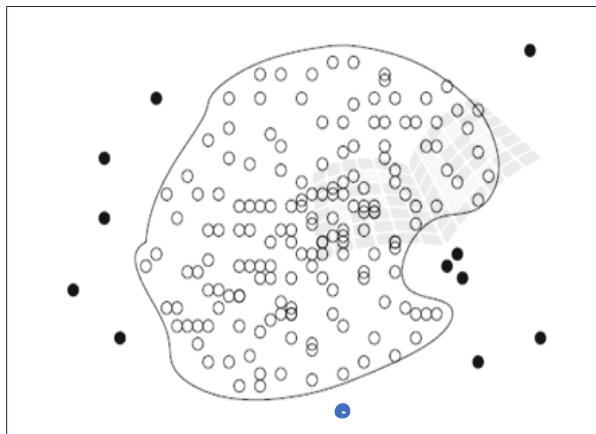
۱- تعداد داده‌های نرمال خیلی بیشتر از داده‌های پرت است و دسته‌بند، بد عمل می‌کند.

۲- ممکن است دسته‌بند نتواند داده‌های پرت جدید را خوب تشخیص دهد.

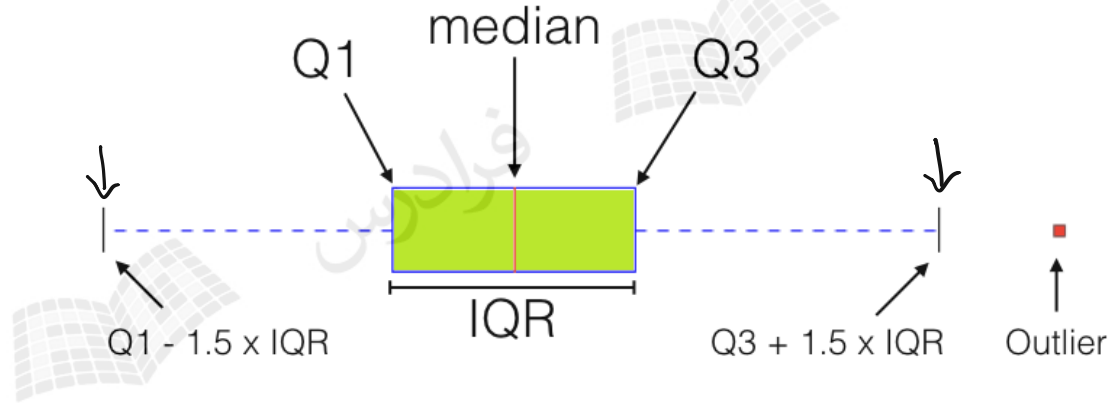
برای رفع این مشکل می‌توان از مدل تک کلاسی (One-Class Model)

استفاده کرد. مرز بین داده‌های نرمال را با روشی مانند SVM پیدا می‌کنیم.

داده‌های پرت در بیرون این مرز قرار دارند.



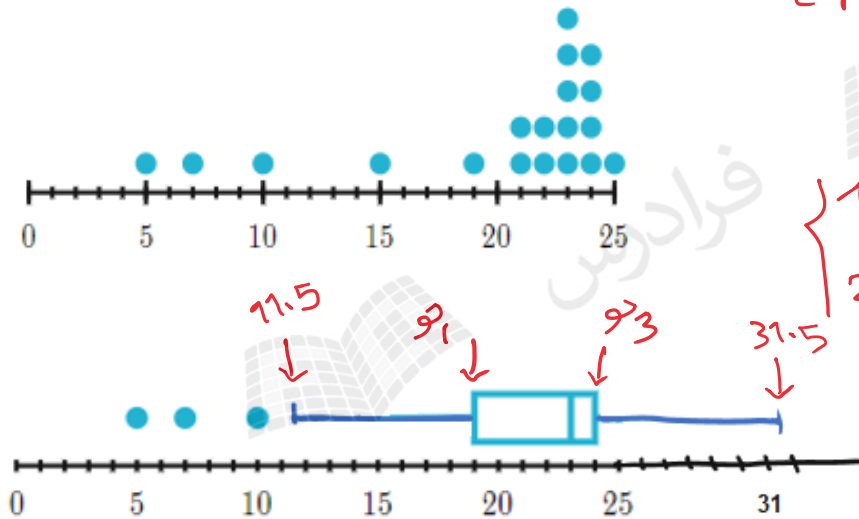
روش Boxplot



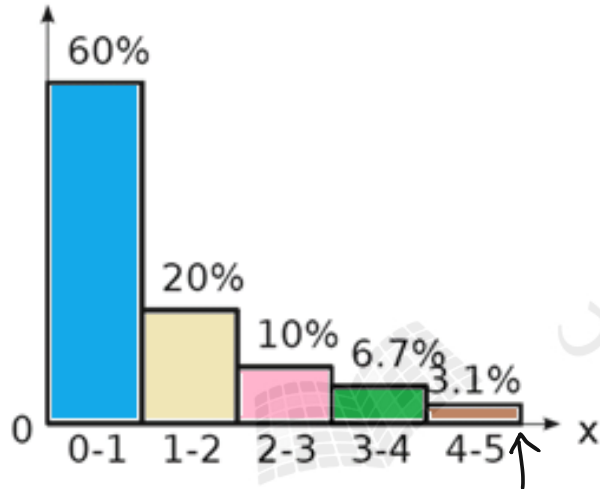
$$Q_3 - Q_1$$

مثال

5, 7, 10, 15, 19, 21, 21, 22, 22, 23, 23, 23, 23, 23, 24, 24, 24, 24, 25



هیستوگرام



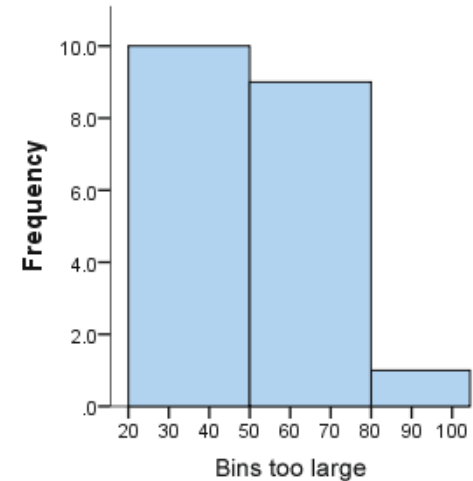
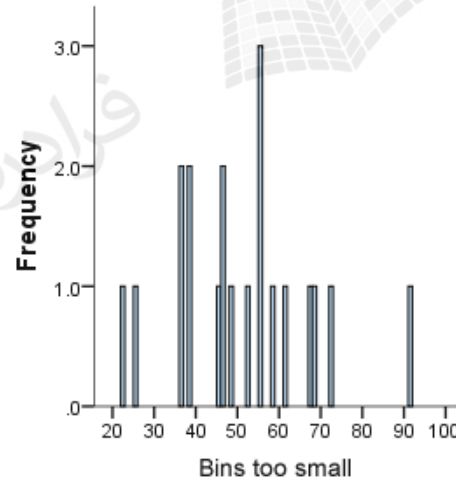
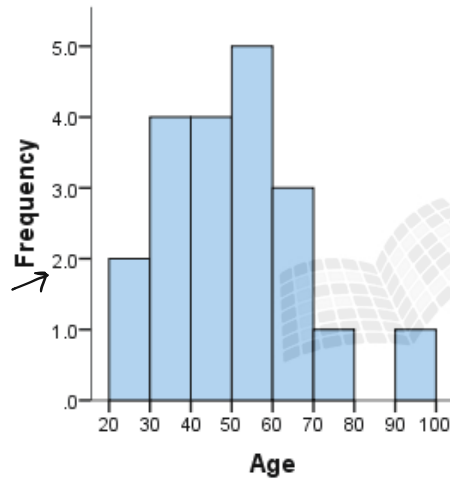
۹۹.۸

۰.۲ %

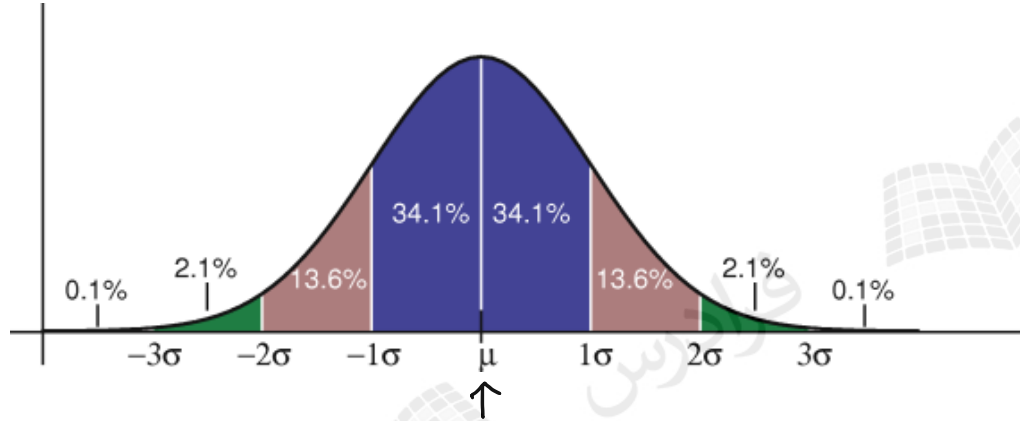
۶۰۰

مشکل روش هیستوگرام

36	25	38	46	55	68	72	55	36	38
67	45	22	48	91	46	52	61	58	55



توزیع نرمال



$$\mu \pm 3\sigma$$

$$99.7\%$$

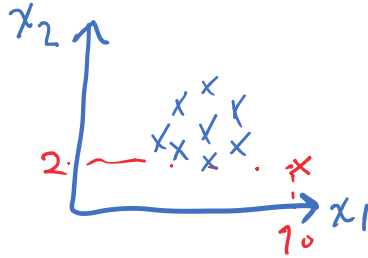
$n=10$ → $\begin{cases} \mu = 28.61 \\ \sigma = 1.51 \end{cases}$

نرمال → $\begin{cases} \mu = 0 \\ \sigma = 1 \end{cases} \pm 3$

$$28.61 + 3(1.51) = 24.1$$

$$28.61 - 3(1.51) = 33.1$$

الگوریتم کشف آنومالی (موارد غیرمتعارف)



۱- انتخاب ویژگی‌هایی که میزان وقوع غیرمتعارف آن‌ها نشان دهنده موارد غیرعادی است.

۲- محاسبه میانگین و واریانس هر ویژگی در مجموعه داده آموزشی.

$$\mu_j = \frac{1}{m} \sum_{i=1}^m x_j^{(i)}$$

$$\sigma_j^2 = \frac{1}{m} \sum_{i=1}^m (x_j^{(i)} - \mu_j)^2$$

$$\begin{matrix} \mu_1 & \mu_2 \\ \sigma_1^2 & \sigma_2^2 \end{matrix}$$

	x_1	x_2
1	✓	
2	✓	
...		
10	✓	

$m=10$
 $n=2$

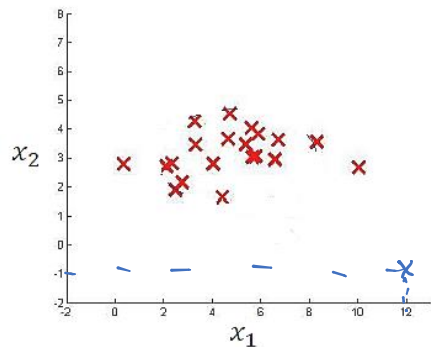
$$p(\mathbf{x}) = \prod_{j=1}^n p(x_j; \mu_j, \sigma_j^2) = \prod_{j=1}^n \frac{1}{\sigma_j \sqrt{2\pi}} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2}\right)$$

۳- بررسی غیرعادی بودن هر نمونه x با محاسبه $P(x)$

اگر $p(x) < \epsilon$ ، آنگاه x یک نمونه غیرعادی است.

$$p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

مثال



$$p(\mathbf{x}) = p(x_1; \mu_1, \sigma_1^2) p(x_2; \mu_2, \sigma_2^2)$$

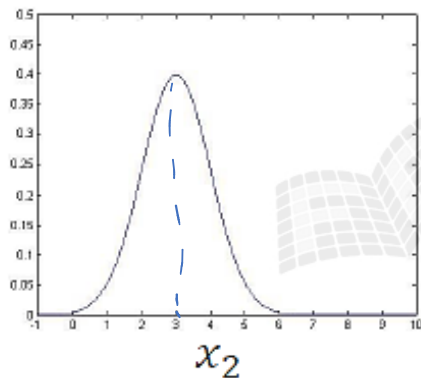
$$x_1 = 12$$

$$x_2 = -1$$

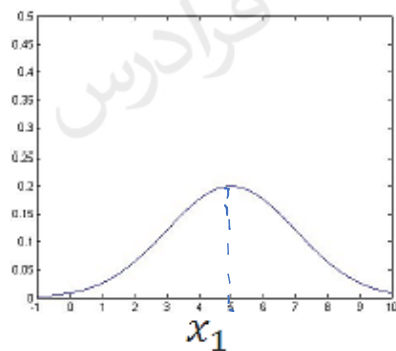
$$p(12, 5, 2^2) \times p(-1, 3, 1^2)$$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

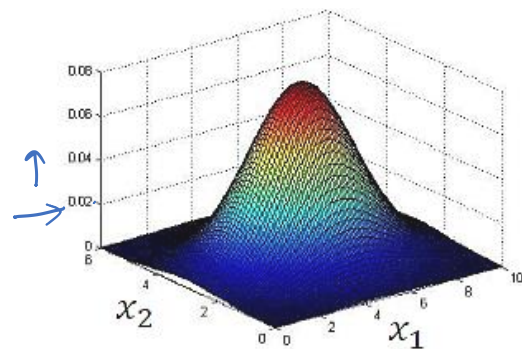
$$p(x) < 0.02$$



$$\mu_2 = 3, \sigma_2 = 1$$



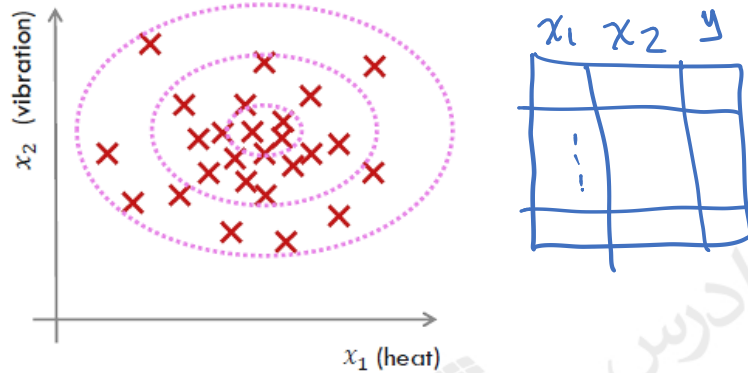
$$\mu_1 = 5, \sigma_1 = 2$$



منحنی گاسین مشترک

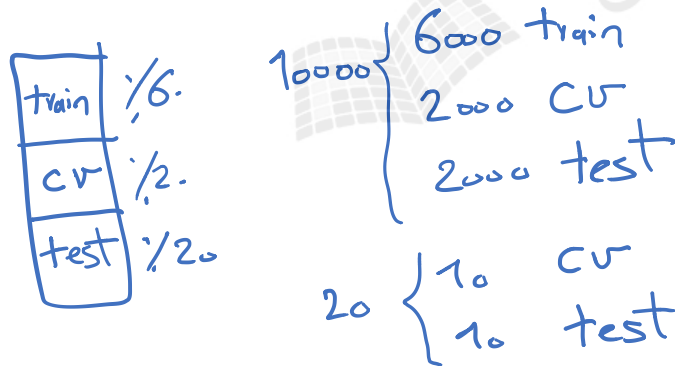
ساخت و ارزیابی سیستم‌های کشف موارد غیرمتعارف

یک تولید کننده موتور هواپیما در مرحله کنترل کیفیت دو ویژگی گرمای تولیدی و ارتعاشات موتور را اندازه گیری می کند.



تعداد موتورهای تولید شده سالم: $y=0 \leftarrow 10000$
 تعداد موتورهای تولید شده غیر سالم: $y=1 \leftarrow 20$
 داده‌های نمونه‌های متعارف:

- تخصیص یافته به مجموعه آموزشی: 6000
 - تخصیص یافته با علامت $y=0$ به مجموعه اعتبارسنجی: 2000
 - تخصیص یافته با علامت $y=1$ به مجموعه تست: 2000
- تقسیم داده‌های نمونه‌های غیرعادی:
- تخصیص یافته با علامت $y=1$ به مجموعه اعتبارسنجی: 10
 - تخصیص یافته با علامت $y=1$ به مجموعه تست: 10



بعد از تقسیم داده‌ها به سه دسته آموزشی، اعتبارسنجی و تست، مدل گوسین $p(x)$ را تنها با استفاده از داده‌های آموزشی می‌سازیم. سپس در حین آموزش، داده‌های اعتبارسنجی را مورد آزمون قرار می‌دهیم. بعد از پایان آموزش هم داده‌های تست را مورد آزمون قرار می‌دهیم.

$$y = \begin{cases} 1, & p(x) < \varepsilon \\ 0, & p(x) \geq \varepsilon \end{cases} \quad \text{محدادی}$$

به کمک شاخص‌های زیر می‌توان عملکرد الگوریتم را سنجید:


۱- میزان نسبت‌های TP و TN و FP و FN.

۲- میزان دقت و بازخوانی

۳- F1-score

در انتها با امتحان داده‌های اعتبارسنجی می‌توان اپسیلونی را انتخاب کرد که بهترین F1 score حاصل شود.

نکات

- برای کشف موارد غیر متعارف بهتر است از روش های بدون سرپرست استفاده کرد.
 $train : \{ x^{(1)} , x^{(2)} , \dots , x^{(n)} \}$
- ✓ اگر یک ویژگی از توزیع احتمالاتی گوسین پیروی نمی کرد، آن را به ویژگی دیگر تبدیل کنید.

- ✓ در انتخاب ویژگی ها برای ساخت مدل کشف موارد غیر متعارف، از ویژگی هایی استفاده کن که مقدار آنها در نمونه های غیر متعارف خیلی کم یا خیلی زیاد باشد.

ویژگی‌هایی با همبستگی زیاد

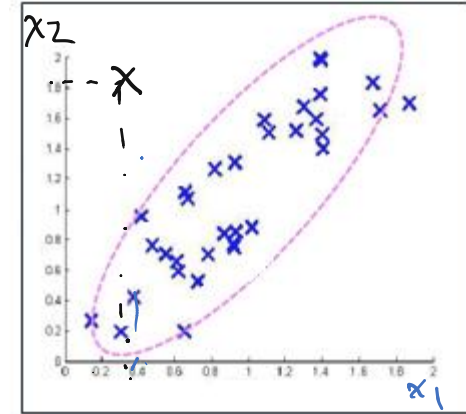
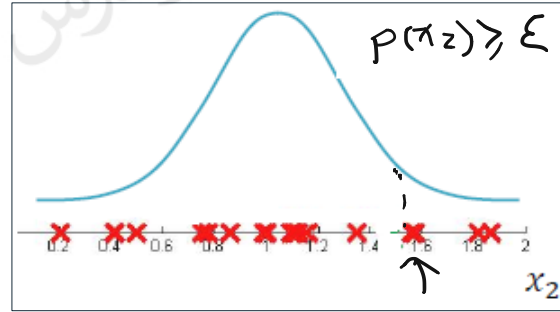
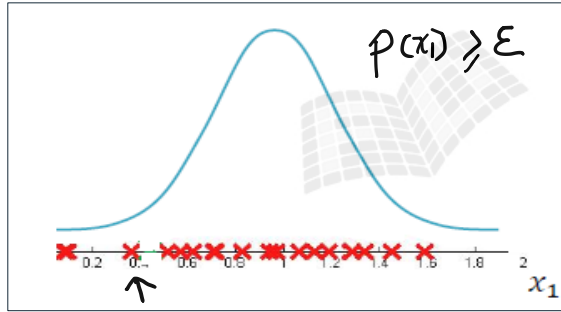
Loop

$$x_5 = \frac{x_2}{x_1}$$

$$x_6 = \frac{x_2^2}{x_1}$$

پایش عملکرد کامپیوترها در مرکز پایگاه داده

- x_1 ترافیک شبکه
- x_2 بار CPU
- x_3 میزان استفاده از حافظه کامپیوتر
- x_4 تعداد دسترسی به دیسک



مدل گوسین چند متغیره

۱- تخمین پارامترهای مدل

$$\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)} \quad \Sigma = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T \quad \Sigma = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$

ماتریس کواریانس

۲- محاسبه مقدار $p(x)$ برای داده جدید x

$$p(x; \mu, \Sigma) = \frac{1}{|\Sigma|^{1/2} (2\pi)^{n/2}} \exp \left(-\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)$$

داده جدید

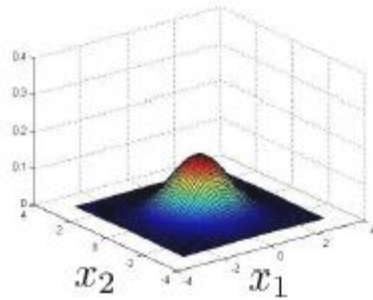
تجزیه

واریان

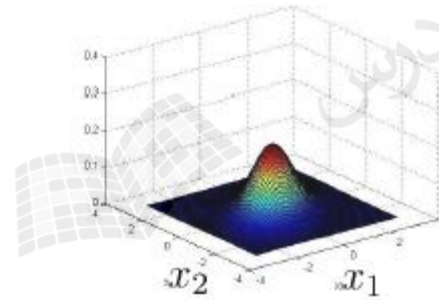
m تعداد نمونه ها
 n تعداد ویژگی

۳- اگر $p(x) < \epsilon$ آنگاه x یک نمونه غیرعادی است.

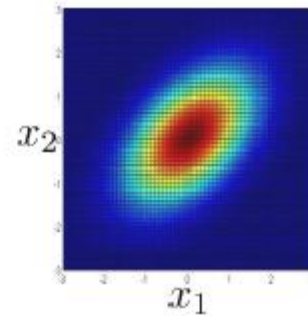
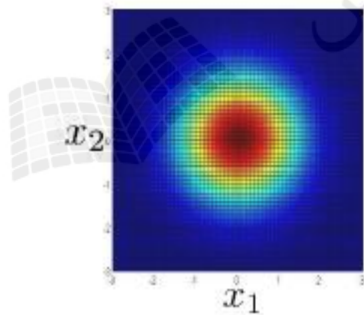
$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$



خطاتراز



مقایسه مدل اولیه و مدل گوسین چندمتغیره

مدل اولیه

- از نظر محاسباتی بسیار ارزان تر است.
- تعداد نمونه های آموزشی می تواند کم باشد.
- برای کشف موارد غیرمتعارفی که با ترکیب غیرعادی ویژگی های مستقل همراه باشد، باید ویژگی های جدیدی را به طور دستی ایجاد کرد.

مدل گوسی چند متغیره

- از نظر محاسباتی گران تر است.
- تعداد نمونه ها باید بیشتر از تعداد ویژگی ها باشد.
- مدل به طور خودکار همبستگی بین ویژگی ها را در بر می گیرد.

این اسلایدها بر مبنای نکات مطرح شده در فرادرس
«آموزش یادگیری ماشین (Machine Learning) (تئوری - عملی) - بخش دوم»
تهیه شده است.

برای کسب اطلاعات بیشتر در مورد این آموزش به لینک زیر مراجعه نمایید.

faradars.org/fvdm94062