فصل پنجم : انتخاب ویژگی

FEATURE SELECTION

در فصل های گذشته، فرض کردیم ویژگی‌ها از قبل انتخاب شده‌اند و تنها به طراحی طبقه بند پرداخته شد.

در این فصل به بررسی روشهای انتخاب این ویژگی‌ها می‌پردازیم.

هدف‌ها:

۱- انتخاب تعداد بهینهٔ ویژگی‌ها (l)

۲- انتخاب بهترین l ویژگی

l بزرگ ۳ عیب دارد: ۱- پیچیدگی محاسباتی بیشتر

۲- کارایی تعمیم پذیری پایین

تعداد نمونه‌های آموزشی

۳- تخمین خطای ضعیف (برای تخمین خطای خوب باید $\frac{N}{l}$ زیاد باشد) حداقل ۱۰ تا ۲۰ $= \frac{N}{l}$

تعداد ویژگی‌ها (ابعاد)

برای N داده شده: l باید به حدی بزرگ باشد که بتواند الف) تفاوت های کلاس ها را از یکدیگر آموزش دهد.

ب) شباهت های داخل هر کلاس را آموزش بدهد.

l باید به حدی کوچک باشد که تفاوت های نمونه های داخل یک کلاس را آموزش ندهد.

در عمل $l < \frac{N}{3}$ یک گزینهٔ قابل قبول برای برخی کاربردها است.

FEATURE SELECTION

- The goals:
  - Select the “optimum” number l of features
  - Select the “best” l features

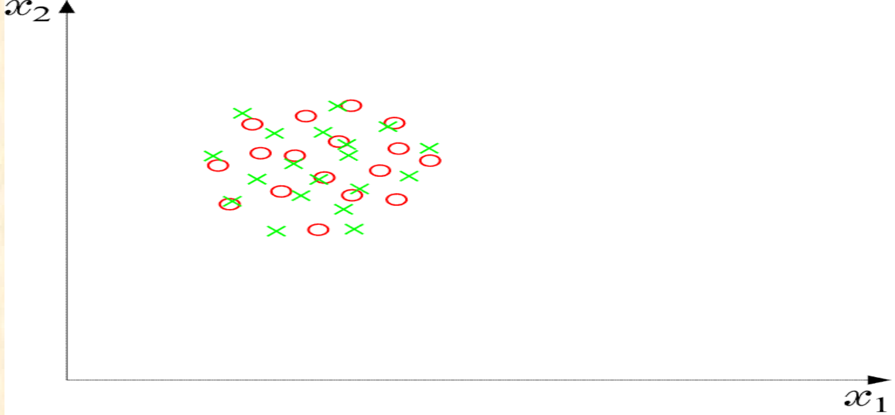
- Large l has a three-fold disadvantage:
  - High computational demands
  - Low generalization performance
  - Poor error estimates

- Given N
  - l must be large enough to learn
    - what makes classes different
    - what makes patterns in the same class similar
  - l must be small enough not to learn what makes patterns of the same class different
  - In practice, $l < N/3$ has been reported to be a sensible choice for a number of cases

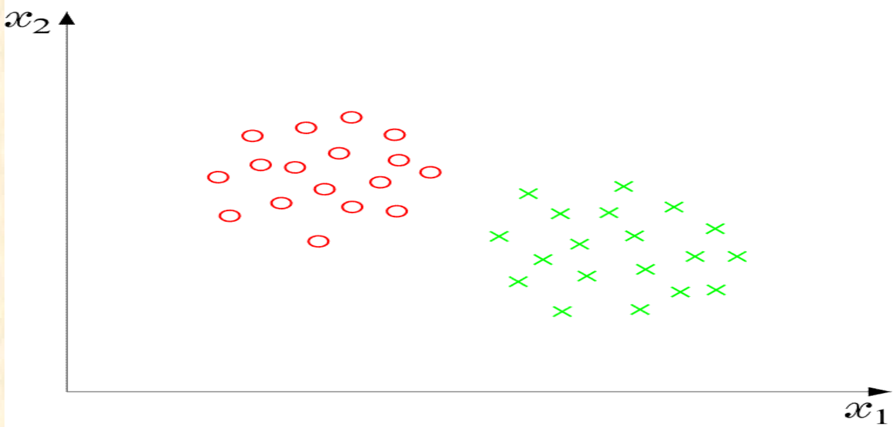
کدام l تا ویژگی؟

- Once l has been decided, choose the l most informative features
  - Best: **Large** between class distance,
    Small within class variance

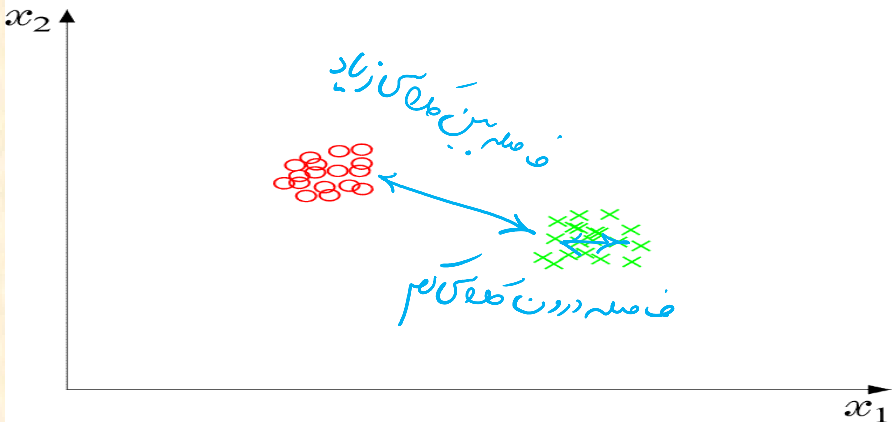
بهترین l ویژگی: بیشترین فاصله بین کلاسی
کمترین فاصله درون کلاسی (واریانس داخل کلاسی



Bad choice



Not bad choice



Good choice

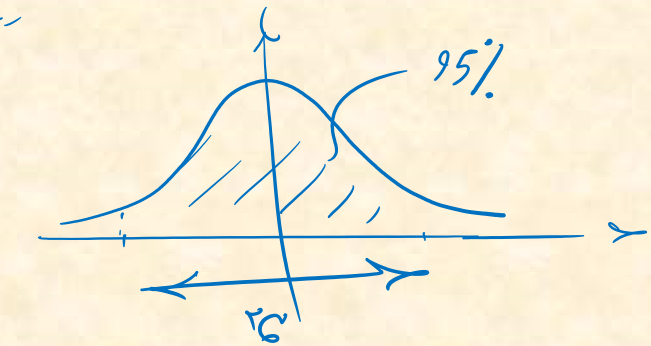
5.2 : پیش پردازش : Preprocessing

5.2.1 : حذف داده های پرت Outlier Removal

داده پرت نقطه ای است که در فاصله خیلی دور از میانگین متغیر تصادفی مورد نظر قرار دارد.

برای متغیر تصادفی با توزیع گاوسی : فاصله دو برابر انحراف معیار ، 95% داده ها را پوشش می دهد.

" " ۳ " " " ، 99% " " " "



این مورد در اندازه گیری های نویزی بیشتر مشاهده می شود.

ولی در مواردی که داده های دور ناشی از دنباله های یک توزیع دنباله دار باشد طراح طبقه بند باید تابع هزینه را طوری تنظیم کند که این نمونه ها را نیز شامل شود.

یکی از معیارهای حساس به داده های پرت، معیار کمینه مربعات خطا است.
Least Squares Error

زیرا خطای زیاد به دلیل توان دوم ایجاد می شود.

5.2.2 نرمالیزه کردن داده‌ها Data Normalization

وقتی مقادیر ویژگی‌ها در محدوده‌های مختلفی قرار دارند نیاز به نرمالیزه کردن داده‌ها است.

different dynamic range

زیرا در غیر این صورت ویژگی‌های با مقادیر بزرگتر اثر بیشتری روی تابع هزینه خواهند داشت که ممکن است انعکاس دهنده درستی از اهمیت آن ویژگی‌ها نباشد.

روش سرراست نرمالیزه کردن با استفاده از تخمین میانگین و واریانس:

$$\begin{cases} \bar{x}_k = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} x_{ik} \quad ; \quad k=1,\dots,l \\[2ex] \sigma_k^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N} (x_{ik}-\bar{x}_{ik})^2 \end{cases}$$

ویژگی نرمالیزه شده با میانگین صفر و واریانس یک

$$\hat{x}_{ik} = \frac{x_{ik}-\bar{x}_{ik}}{\sigma_k}$$

روش خطی

$[-1,1]$, $[0,1]$

در کنار روش های خطی، روش های غیرخطی نیز وجود دارند که برای حالت هایی که توزیع داده ها اطراف میانگین به صورت یکنواخت نیست کاربرد دارند. در روش های غیرخطی تبدیل هایی با توابع غیرخطی (تفسیر لگاریتمی، sigmoid و ...) برای گسترش داده ها در بازه های مورد نظر به کار می روند.

برای مثال روش مقیاس بندی (Scaling) Soft max :

$$y = \frac{x_{ik} - \bar{x}_{ik}}{r\sigma_k}$$

$$\hat{x}_{ik} = \frac{1}{1+e^{-y}}$$

- داده ها را در محدوده $[0,1]$ قرار می دهند.
- برای y های خیلی کوچک رابطه خطی است.

5.2.3 : داده های از دست رفته Missing Data :

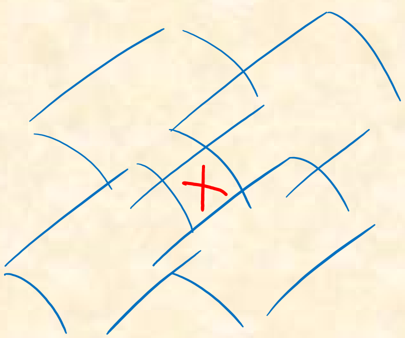
در عمل، برخی نمونه ها از داده ها ثبت نشده اند.

در علوم انسانی: فرم های نظرسنجی survey : پاسخ های ناقص ← داده های از دست رفته

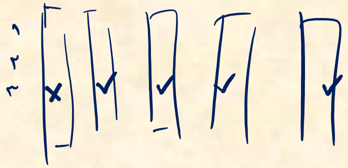
در سنجش از دور : نواحی مشخصی تنها با بخشی از حسگرها تحت پوشش قرار می گیرد.

در شبکه های حسگر : اطلاعات از حسگرهای توزیع شده جمع آوری، ادغام می شوند.

← داده های ناقص



روشهای تکمیل داده های ناقص: Imputation



۱- قرار دادن صفر به جای داده های از دست رفته

۲- قرار دادن میانگین بدون شرط که از روی مقادیر موجود داده محاسبه شده است.

۳- قرار دادن میانگین شرطی، درصورت موجود بودن تخمینی از pdf های داده های از دست رفته از روی داده های موجود.

فلسفه اصلی انتخاب ویژگی : - حذف ویژگی‌های تکی که اطلاعات ضعیفی دارند.

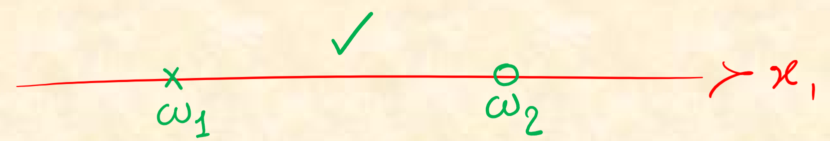
- بررسی توأم ویژگی‌های قوی باقیمانده.

- ❖ The basic philosophy
  - ➢ Discard individual features with poor information content
  - ➢ The remaining information rich features are examined jointly as vectors

5.4: انتخاب ویژگی بر اساس روش تست فرضیه آماری :

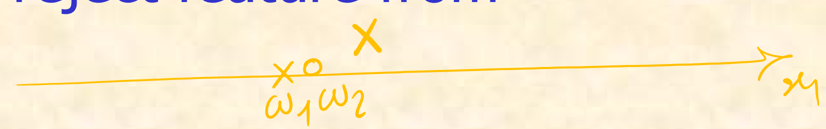
- ❖ Feature Selection based on statistical Hypothesis Testing
  - ➢ The Goal: For each individual feature, find whether the values, which the feature takes for the different classes, differ significantly.

    That is, answer

    - $H_1:\ \theta_1 \neq \theta_0$: The values differ significantly
    - $H_0:\ \theta_1 = \theta_0$: The values do not differ significantly

    If they do not differ significantly reject feature from subsequent stages.

- ❖ Hypothesis Testing Basics

گام اول در انتخاب ویژگی این است که به ویژگی ها به دید مستقل نگاه می کنیم و توانایی تمیز دادن کلاس ها را توسط آنها بررسی می کنیم.

- فرض کنید x متغیر تصادفی نمایش دهنده یک ویژگی مشخص باشد.

- هدف بررسی این نکته است که مقادیر این متغیر تصادفی برای کلاس های مختلف ω_1 و ω_2 تفاوت چشمگیری دارند یا خیر.

برای پاسخ به این سوال از فرمول بندی ریاضی تست فرضیه آماری استفاده می کنیم:

Alternative Hypothesis H_1: مقادیر ویژگی تفاوت چشمگیر دارند

Null Hypothesis H_0: 〃 〃 〃 〃 ندارند

برای حل این مسئله تفاوت ها از مقدار میانگین را برای یک ویژگی مشخص در کلاس های مختلف در نظر می گیریم و بررسی می کنیم که آیا این تفاوت ها از صفر فاصله قابل توجهی دارند یا خیر.

5.4.1 اصول فرضیه آماری:

فرض کنید x متغیر تصادفی با توزیع احتمال مشخص است که پارامتر θ آن نامعلوم است.
برای مثال اگر توزیع گاوسی باشد این پارامتر θ می‌تواند میانگین یا واریانس باشد.

$$\begin{cases} H_1: \theta \neq \theta_0 \\ H_0: \theta = \theta_0 \end{cases}$$

روال تصمیم: فرض کنید $x_i \,;\, i=1,2,\dots,N$ نمونه‌های آموزشی متغیر تصادفی x باشند. تابع $f(.,\dots,.)$
انتخاب می‌شود. بنابراین متغیر $q = f(x_1, x_2, \dots, x_N)$ ، f به صورتی انتخاب می‌شود که توزیع

test statistics آماره تست

احتمال q به راحتی برحسب پارامتر نامعلوم θ فرمول‌بندی شود: $P_q(q;\theta)$

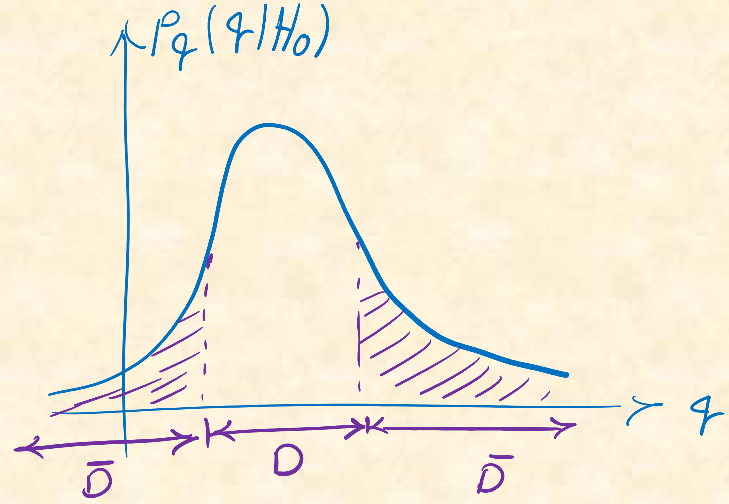
D: بازه‌ای که q در آن با احتمال زیادی فرض H_0 برقرار باشد (بازه قبول) acceptance interval

$\bar{D}$: مکمل D است. فرض H_1 برقرار است. (بازه بحرانی) critical interval

احتمال تصمیم غلط :

فرض کنید H_0 فرض درست باشد. بنابراین احتمال رخ دادن خطا:

$$p(q\in\bar{D}|H_0)=\rho$$



$\bar{D}$ روی $P_q(q|H_0)$ انتگرال

significance level : ρ

سطح اهمیت

حالت با واریانس معلوم :

فرض کنید که x متغیر تصادفی و x_i ; $i=1,2,\dots,N$ نمونه های حاصل از این فرآیند باشند. فرض کنید که نمونه ها به صورت متغیر مستقل از هم (iid) و $E[(x-\mu)^2]=\sigma^2$, $E[x]=\mu$

تخمین مستقل برای μ بر حسب نمونه ها :

$$\boxed{\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i}$$

از آنجایی که با تغییر نمونه ها تخمین متفاوتی حاصل می شود، $\bar{x}$ نیز متغیر تصادفی است که با تابع احتمال $f_{\bar{x}}(\bar{x})$ مشخص می شود که :

$$E[\bar{x}]=\frac{1}{N}\sum_{i=1}^{N}E[x_i]=\frac{1}{\cancel{N}}\cancel{N}\mu=\mu$$

$\bar{x}$ یک تخمین نااریب unbiased از میانگین x (μ) است.

$$E[(\bar{x}-\mu)^2]=E[(\frac{1}{N}\sum x_i-\mu)^2]=\frac{1}{N^2}\sum_{i=1}^{N}E[(x_i-\mu)^2]+\frac{2}{N^2}\sum_{i}\sum_{j\neq i}E[(x_i-\mu)(x_j-\mu)]$$

از فرض استقلال آماری

$E[(x_i-\mu)(x_j-\mu)]=0$

$$\rightarrow \sigma^2_{\bar{x}}=\frac{1}{N}\sigma^2$$

هرچه N بزرگتر باشد، واریانس تخمین حول μ کمتر می باشد ← تخمین دقیق تر

فرض کنید $\hat{\mu}$ داده شده، و باید تصمیم زیر را بگیریم :

$$H_1 : E[x] \neq \hat{\mu}$$
$$H_0 : E[x] = \hat{\mu}$$

برای این منظور آماره تست زیر را تعریف می‌کنیم :

$$q = \frac{\bar{x} - \hat{\mu}}{\sigma / \sqrt{N}}$$

با توجه به قضیه حد مرکزی تابع توزیع احتمال $\bar{x}$ تحت فرض H_0 (یعنی $\hat{\mu}$ داده شده) تقریباً گاوسی است :

Appendix A

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\,\sigma} e^{\frac{-N(\bar{x}-\mu)^2}{\sigma^2}}$$

$$\sim N\left(\hat{\mu}, \frac{\sigma^2}{N}\right)$$

بنابراین توزیع q تحت فرض H_0 به صورت $N(0,1)$

برای سطح اهمیت ρ : بازه اعتماد $D = [-x_\rho, x_\rho]$ معمولاً انتخاب می‌شود که در این بازه احتمال q ، $1-\rho$ باشد. (q احتمال حضور در بازه D)

Table 5.1

$1-\rho$	0.8	0.85	0.9	...
x_ρ	1.282	1.440	1.645	...

گام‌ها:

۱- با N نمونه آموزشی داده شده از متغیر $\underline{x}$، q را به دست می‌آوریم.

۲- سطح اهمیت ρ را انتخاب می‌کنیم.

۳- از روی جدول مقادیر $N(0,1)$. بازه قبول $D=[-x_\rho, x_\rho]$ را متناظر با $(1-\rho)$ به دست می‌آوریم.

۴- اگر $q \in D$ باشد، H_0 را قبول می‌کنیم در غیر این صورت H_1 را قبول می‌کنیم.

مثال: متغیر تصادفی x با $\sigma = ۰٫۲۳$ ، $N=۱۶$ ، $\underline{\bar{x} = ۱٫۳۵}$ ، $\rho = ۰٫۰۵$ را مفروض بگیرید، $\hat{\mu} = ۱٫۴$ درست است یا خیر؟

$$1-\rho = ۰٫۹۵ \rightarrow x_\rho = 1{,}967$$

Table 5.1

$$\longrightarrow \text{prob}\left\{-۱٫۹۷ < \frac{\bar{x}-\hat{\mu}}{\sigma/\sqrt{N}} < ۱٫۹۷\right\} = ۰٫۹۵$$

$$\frac{۱٫۳۵-۱٫۴}{۰٫۲۳/۴} = \frac{-۰٫۰۵}{۰٫۰۵۷۵} = -۰٫۸۷ \in D \checkmark$$

H_0 قابل قبول

$E[\underline{x}] = \hat{\mu}$

- The steps:
  - N measurements $x_i, i = 1,2,\ldots, N$ are known
  - Define a function of them

    $$q = f(x_1, x_2, \ldots, x_N):$$ test statistic

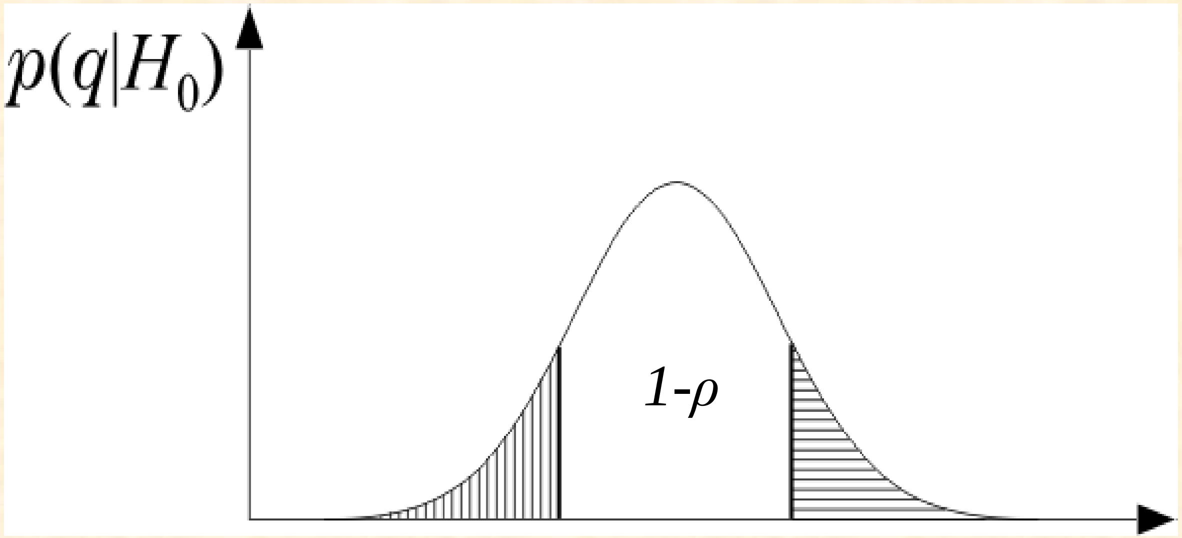
    so that $\boxed{p_q(q;\theta)}$ is easily parameterized in terms of θ.
  - Let D be an interval, where q has a high probability to lie under H_0, i.e., $p_q(q|\theta_0)$
  - Let $\bar{D}$ be the complement of D

    $D \longrightarrow$ Acceptance Interval

    $\bar{D} \longrightarrow$ Critical Interval
  - If q, resulting from $x_1, x_2, \ldots, x_N$, lies in D we accept H_0, otherwise we reject it.

➢ Probability of an error

$$p_q(q \in \overline{D} | H_0) = \rho$$



- ρ is preselected and it is known as the significance level.

❖ Application: The known variance case:

- Let x be a random variable and the experimental samples, $x_i = 1, 2, \ldots, N$, are assumed mutually independent. Also let

$$E[x] = \mu$$

$$E[(x-\mu)^2] = \sigma^2$$

- Compute the sample mean

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- This is also a random variable with mean value

$$E[\bar{x}] = \frac{1}{N}\sum_{i=1}^{N} E[x_i] = \mu$$

That is, it is an Unbiased Estimator

- The variance $\sigma_{\bar{x}}^2$

$$E[(\bar{x}-\mu)^2] = E[(\frac{1}{N}\sum_{i=1}^{N} x_i - \mu)^2]$$

$$= \frac{1}{N^2}\sum_{i=1}^{N} E[(x_i-\mu)^2] + \frac{1}{N^2}\sum_i \sum_j E[(x_i-\mu)(x_j-\mu)]$$

Due to independence

$$\sigma_{\bar{x}}^2 = \frac{1}{N}\sigma_x^2$$

That is, it is Asymptotically Efficient

- Hypothesis test

$$H_1 : E[x] \neq \hat{\mu}$$

$$H_0 : E[x] = \hat{\mu}$$

- Test Statistic: Define the variable

$$\boxed{q = \frac{\bar{x}-\hat{\mu}}{\sigma/\sqrt{N}}}$$

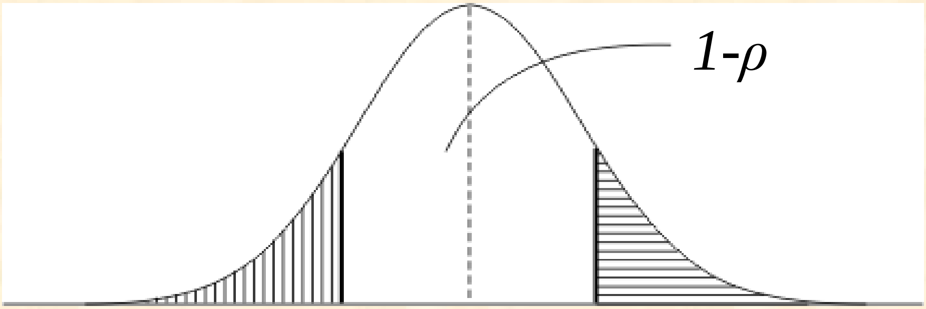
- Central limit theorem under H_0

$$p_{\bar{x}}(\bar{x}) = \frac{\sqrt{N}}{\sqrt{2\pi}\sigma} \exp\left(-\frac{N(\bar{x} - \hat{\mu})^2}{2\sigma^2} \right)$$

- Thus, under H_0

$$p_q(q) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{q^2}{2} \right) \quad q \approx N(0,1)$$

- The decision steps
  - Compute q from x_i, $i=1,2,\ldots,N$
  - Choose significance level ρ
  - Compute from $N(0,1)$ tables $D=[-x_\rho, x_\rho]$



  - if $q \in D$ accept H_0

    if $q \in \overline{D}$ reject H_0

- An example: A random variable x has variance $\sigma^2=(0.23)^2$. $N=16$ measurements are obtained giving $\bar{x}=1.35$. The significance level is $\rho=0.05$.

Test the hypothesis

$$H_0: \mu = \hat{\mu} = 1.4$$

$$H_1: \mu \neq \hat{\mu}$$

- Since σ² is known, $q = \dfrac{\bar{x} - \hat{\mu}}{\sigma / 4}$ is $N(0,1)$.

  From tables, we obtain the values with acceptance intervals $[-x_\rho, x_\rho]$ for normal $N(0,1)$

$1\text{-}\rho$	0.8	0.85	0.9	0.95	0.98	0.99	0.998	0.999
x_ρ	1.28	1.44	1.64	1.96	2.32	2.57	3.09	3.29

- Thus

$$\text{Prob}\left\{-1.967 < \frac{\bar{x} - \hat{\mu}}{0.23/4} < 1.967\right\} = 0.95$$

or

$$\text{Prob}\{-0.113 < \bar{x} - \hat{\mu} < 0.113\} = 0.95$$

or

$$\boxed{\text{Prob}\{1.237 < \hat{\mu} < 1.463\} = 0.95}$$

- Since $\hat{\mu} = 1.4$ <u>lies</u> within the above <u>acceptance</u> interval, we accept H_0, i.e.,

$$\mu = \hat{\mu} = 1.4$$

The interval [1.237, 1.463] is also known as confidence interval at the $1\text{-}\rho=0.95$ level.

We say that: There is no evidence at the 5% level that the mean value is not equal to $\hat{\mu}$

حالتِ واریانس نامشخص :

❖ The Unknown Variance Case

➢ Estimate the variance. The estimate

$$\boxed{\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

$$\rightarrow E[\hat{\sigma}^2] = \frac{1}{N-1}\sum_{i=1}^{N} E[(x_i-\bar{x})^2]$$

$$= \frac{1}{N-1}\sum_{i=1}^{N} E[((x_i-\mu)-(\bar{x}-\mu))^2]$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(\sigma^2 + \frac{\sigma^2}{N} - 2\cancel{E[(x_i-\mu)(\bar{x}-\mu)]}^{\sigma^2/N}\right)$$

is unbiased, i.e.,

$$E[\hat{\sigma}^2] = \sigma^2$$

نمونه‌های آزمایش مستقل از هم‌اند :

$$E[(x_i-\mu)(\bar{x}-\mu)] =$$

$$= \frac{1}{N} E[(x_i-\mu)((x_1-\mu)+\dots+(x_N-\mu))]$$

$$= \frac{\sigma^2}{N}$$

➢ Define the test statistic

$$\boxed{q = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{N}}}$$

$$E[\hat{\sigma}^2] = \frac{N}{N-1}\,\frac{N-1}{N}\,\sigma^2 = \sigma^2$$

بدون بایاس

توزیع q توزیع t با درجه آزادی $N-1$ است (پیوست A)

- This is no longer Gaussian. If x is Gaussian, then q follows a t-distribution, with N-1 degrees of freedom

مثال 5.2 : مثال قبلی را با $\hat{\sigma} = ۰٫۲۳$ در نظر بگیرید.

$N = ۱۶ \longrightarrow N - ۱ = ۱۵$

$\rho = ۰٫۰۲۵ \rightarrow ۱ - \rho = ۰٫۹۷۵$

- An example:

x is Gaussian, $N = 16$, obtained from measurements,

$\bar{x} = 1.35$ and $\hat{\sigma}^2 = (0.23)^2$. Test the hypothesis

$H_0:\ \mu = \hat{\mu} = 1.4$

at the significance level $\rho = 0.025$.

جدول بازهٔ قبول برای توزیع t

➢ Table of acceptance intervals for t-distribution

Degrees of Freedom (N-۱)	1-ρ	0.9	0.95	0.975	0.99
12		1.78	2.18	2.56	3.05
13		1.77	2.16	2.53	3.01
14		1.76	2.15	2.51	2.98
15		1.75	2.13	2.49	2.95
16		1.75	2.12	2.47	2.92
17		1.74	2.11	2.46	2.90
18		1.73	2.10	2.44	2.88

➢ $$\text{Prob}\left\{-2.49 < \frac{\bar{x}-\hat{\mu}}{\hat{\sigma}/4} < 2.49\right\}$$

$$1.207 < \hat{\mu} < 1.493$$

Thus, $\hat{\mu} = 1.4$ is accepted ✓

$$\frac{\bar{x}-\hat{\mu}}{\hat{\sigma}/\sqrt{N}} = \frac{۱,۳۵-۱,۴}{۰,۲۳/۴} = -۰,۸۷$$

$$-۲,۴۹ < -۰,۸۷ < ۲,۴۹$$

✓ $\hat{\mu} = ۱,۴$? ✓

❖ Application in Feature Selection

کاربرد تست t در انتخاب ویژگی:

آیا ویژگی داده شده برای مسأله طبقه بندی مناسب است یا خیر؟

- The goal here is to test against zero the difference μ_1-μ_2 of the respective means in ω_1, ω_2 of a single feature.

$E[x] = \mu_1$

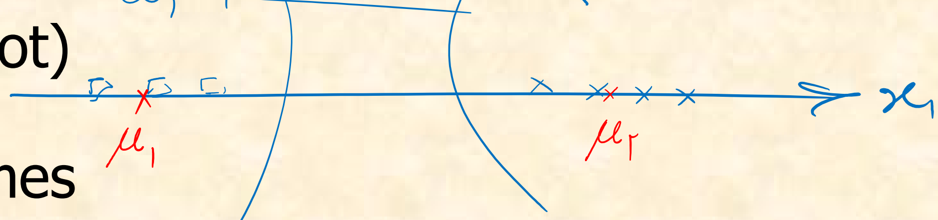
- Let x_i $i=1,\ldots,N$, the values of a feature in ω_1

$E[y] = \mu_2$

- Let y_i $i=1,\ldots,N$, the values of the same feature in ω_2

- Assume in both classes $\boxed{\sigma_1^2 = \sigma_2^2 = \sigma^2}$ (unknown or not)

واریانس های برابر



- The test becomes

$$H_0 : \Delta\mu = \mu_1 - \mu_2 = 0$$

$$H_1 : \Delta\mu \neq 0$$

$z = x - y$

$E[z] = \mu_1 - \mu_2$

- Define

  $z = x - y$

- Obviously

  $E[z] = \mu_1 - \mu_2$

- Define the average

$$\bar{z} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i) = \bar{x} - \bar{y}$$

- Known Variance Case: Define

حالتِ واریانس معلوم :

آماره تست :

test statistics

$$q = \frac{(\bar{x} - \bar{y}) - (\hat{\mu}_1 - \hat{\mu}_2)}{\sigma\sqrt{\frac{2}{N}}}$$

- This is $N(0,1)$ and one follows the procedure as before.

5.1

- **Unknown Variance Case:** حالتِ واریانسِ نامعلوم:
  Define the test statistic

$$q = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{S_z \sqrt{\frac{2}{N}}}$$

$$S_z^2 = \frac{1}{2N-2}\left(\sum_{i=1}^{N}(x_i - \bar{x})^2 + \sum_{i=1}^{N}(y_i - \bar{y})^2\right) = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)$$

- q is t-distribution with $2N$-2 degrees of freedom, 5.2
- Then apply appropriate tables as before.

مثال: مقدار یک ویژگی در دو کلاس به صورت زیر داده شده است:

- **Example:** The values of a feature in two classes are:

ω_1: 3.5, 3.7, 3.9, 4.1, 3.4, 3.5, 4.1, 3.8, 3.6, 3.7

ω_2: 3.2, 3.6, 3.1, 3.4, 3.0, 3.4, 2.8, 3.1, 3.3, 3.6

آیا این ویژگی با سطح اعتماد ۰٫۰۵ = ρ ویژگی مناسبی برای طبقه‌بندی دو کلاس است؟

Test if the mean values in the two classes differ significantly, at the significance level ρ=0.05

- We have

$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad , \hat{\sigma}_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$

$$\rightarrow \omega_1 : \quad \bar{x} = 3.73, \ \hat{\sigma}_1^2 = 0.0601$$

$$\rightarrow \omega_2 : \quad \bar{y} = 3.25, \ \hat{\sigma}_2^2 = 0.0672$$

For $N=10$

$$S_z^2 = \frac{1}{2}(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) = \frac{۱}{۲}(۰٫۰۶۰۱ + ۰٫۰۶۷۲) = ۰٫۰۶۳۶۵$$

$$q = \frac{(\bar{x} - \bar{y}) - 0}{S_z\sqrt{\frac{2}{10}}}$$

$$\boxed{q = 4.25}$$

- From the table of the t-distribution with $2N$-2=18 degrees of freedom and ρ=0.05, we obtain D=[-2.10,2.10] and since q=4.25 is outside D, H_1 is accepted and the feature is selected.

ناحیه قبول

❖ Class Separability Measures

۵.۶ : معیارهای جدایی‌پذیری کلاس‌ها

The emphasis so far was on individually considered features. However, such an approach cannot take into account existing correlations among the features. That is, two features may be rich in information, but if they are highly correlated we need not consider both of them. To this end, in order to search for possible correlations, we consider features jointly as elements of vectors. To this end:

تاکنون ویژگی‌ها را بصورت جداگانه در نظر گرفتیم. در این قسمت ویژگی‌ها را به صورت توأم بررسی می‌کنیم.

گام‌ها :

➢ Discard poor in information features, by means of a statistical test.

۱- حذف ویژگی‌های ضعیف با ارزیابی انفرادی آنها توسط تست آماری

۲- مشخص کردن بیشینه تعداد ویژگی‌ها (ℓ)

$A_1 = \pi r_1^2$ $\quad$ $A_2 = \pi r_2^2$

➢ Choose the maximum number, ℓ, of features to be used. This is dictated by the specific problem (e.g., the number, N, of available training patterns and the type of the classifier to be adopted).

- Combine remaining features to search for the "best" combination. To this end:

۳- ترکیب ویژگی‌های باقی مانده به صورت بردارهای l تایی و جستجوی بهترین l ویژگی

  - Use different feature combinations to form the feature vector. Train the classifier, and choose the combination resulting in the best classifier performance.

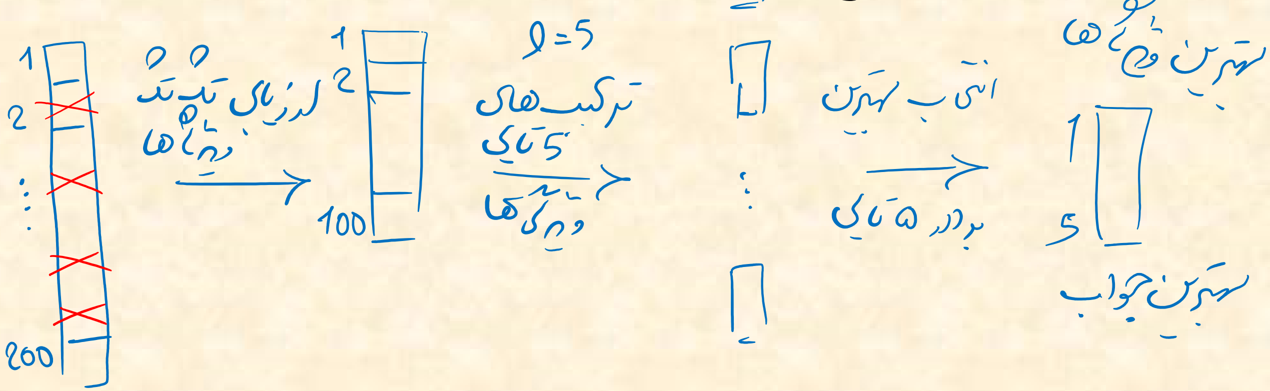
    A major disadvantage of this approach is the high complexity. Also, local minima, may give misleading results.

عیب این روش پیچیدگی محاسباتی زیاد آن است.

  - Adopt a class separability measure and choose the best feature combination against this cost.

ویژگی‌های اولیه

تعریف معیارهای جدایی پذیری کلاسها که به عنوان تابع هزینه انتخاب ویژگی استفاده می شود.



➢ Class separability measures: Let $\underline{x}$ be the current feature combination vector.

5.6.1
دیورژانس

- Divergence. To see the rationale behind this cost, consider the two – class case. Obviously, if on the average the value of $\ln \frac{p(\underline{x} \mid \omega_1)}{p(\underline{x} \mid \omega_2)}$ is close to zero, then $\underline{x}$ should be a poor feature combination. Define:

$p(\omega_1 \mid \underline{x}) > p(\omega_2 \mid \underline{x})$

$\frac{p(\omega_1 \mid \underline{x})}{p(\omega_2 \mid \underline{x})} > 1$

$\ln \frac{p(\omega_1 \mid \underline{x})}{p(\omega_2 \mid \underline{x})} > 0$

– $$D_{12} = \int_{-\infty}^{+\infty} p(\underline{x} \mid \omega_1) \ln \frac{p(\underline{x} \mid \omega_1)}{p(\underline{x} \mid \omega_2)} d\underline{x}$$

– $$D_{21} = \int_{-\infty}^{+\infty} p(\underline{x} \mid \omega_2) \ln \frac{p(\underline{x} \mid \omega_2)}{p(\underline{x} \mid \omega_1)} d\underline{x}$$

– $$d_{12} = D_{12} + D_{21}$$

چندکلاسه
معیار دیورژانس
در حالت دوکلاسه

d_{12} is known as the divergence and can be used as a class separability measure.

– For the multi-class case, define d_{ij} for every pair of classes ω_i, ω_j and the average divergence is defined as

1 2 3 — d_{12}, d_{13}, d_{23}

دیورژانس چندکلاسه

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} P(\omega_i)P(\omega_j)d_{ij}$$

– Some properties: kullback-leibler فاصله یا دیورژانس

$$\begin{cases} d_{ij} \geq 0 \\ d_{ij} = 0, \text{if } i = j \\ d_{ij} = d_{ji} \end{cases}$$

– Large values of d are indicative of good feature combination.

هرچه d بزرگتر باشد نشاندهنده بردار ویژگی مناسب است.

محاسبهٔ دیورژانس برای توزیع گوسی:

$\mathcal{N}(\underline{\mu}_i, \Sigma_i)$, $\mathcal{N}(\underline{\mu}_j, \Sigma_j)$

$d_{ij} =$ ______ (5.22)

برای حالت یک بعدی:

ساده شده $d_{ij} = \frac{1}{2}\left(\frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} - 2\right) + \frac{1}{2}(\mu_i - \mu_j)^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}\right)$

نتیجه: معیار جدایی پذیری تنها به میانگین در دو کلاس وابسته نیست و به واریانس نیز باید بستگی داشته باشد.

$\Sigma_i = \Sigma_j = \Sigma$ در حالت خاص (5.22)

$d_{ij} = (\underline{\mu}_i - \underline{\mu}_j)^T \Sigma^{-1} (\underline{\mu}_i - \underline{\mu}_j)$

فاصلهٔ ماهالانوبیس بین بردارهای میانگین مشخصه های زوج

نکته: رابطهٔ بین دیورژانس و خطای طبقه بندی رابطهٔ مستقیم نیست.

دیورژانس تبدیل یافته

transformed divergence

$\hat{d}_{ij} = 2\left(1 - e^{-\frac{d_{ij}}{8}}\right)$

5.6.2 فاصله بتاچاریا و حد چرنوف — Chernoff Bound & Battacharyya Distance

کمینه احتمال خطای قابل دستیابی توسط طبقه‌بند بیز برای دو کلاس ω_i و ω_j:

$$P_e = \int_{-\infty}^{+\infty} \min[\underbrace{P(\omega_i)\,p(\underline{x}|\omega_i)}_{a}, \underbrace{P(\omega_j)\,p(\underline{x}|\omega_j)}_{b}]\, d\underline{x} \qquad (5.23)$$

محاسبه تحلیلی انتگرال فوق در حالت کلی ممکن نیست. ولی می‌توان حد بالای آن را به دست آورد.

$$\min[a,b] \leq a^s b^{1-s} \quad \text{for } a,b \geq 0 \,,\ 0 \leq s \leq 1 \qquad (5.24)$$

$$(5.23),(5.24) \longrightarrow P_e \leq P(\omega_i)^s P(\omega_j)^{1-s} \int_{-\infty}^{+\infty} p(\underline{x}|\omega_i)^s\, p(\underline{x}|\omega_j)^{1-s}\, d\underline{x} \equiv \varepsilon_{CB}$$

حد chernoff

با کمینه کردن حد چرنوف نسبت به s می‌توان به کمینه خطا دست پیدا کرد.

یک حالت خاص این حد به ازای $s = 1/2$ به صورت زیر به دست آمده:

$$P_e \leq \varepsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)} \int_{-\infty}^{+\infty} \sqrt{p(\underline{x}|\omega_i)\, p(\underline{x}|\omega_j)}\, dx$$

در حالت گوسی $N(\underline{\mu}_i, \Sigma_i), N(\underline{\mu}_j, \Sigma_j)$

$$\downarrow \varepsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)}\; e^{-B} \uparrow$$

فاصله Battacharyya

$$B = \frac{1}{8}(\underline{\mu}_i - \underline{\mu}_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\underline{\mu}_i - \underline{\mu}_j) + \frac{1}{2} \ln \frac{\left|\frac{\Sigma_i + \Sigma_j}{2}\right|}{\sqrt{|\Sigma_i||\Sigma_j|}}$$

$|.|$ دترمینان

این فاصله را می‌توان به عنوان معیار جدایی پذیری کلاسها استفاده کرد.

در حالت $\Sigma_i = \Sigma_j$ این فاصله به حد بهینه شرنوف دست پیدا می‌کند و فاصله باتاچاریا به بخشی از فاصله ماهالانوبیس تبدیل می‌شود.

مثال: فرض کنید $P(\omega_1) = P(\omega_2)$ و توزیع های گاوسی $\mathcal{N}(\underline{\mu}, \sigma_1^2 I)$, $\mathcal{N}(\underline{\mu}, \sigma_2^2 I)$

فاصله باتاچاریا:

$$B = \frac{1}{2} \ln \frac{\left(\frac{\sigma_1^2 + \sigma_2^2}{2}\right)^l}{\sqrt{\sigma_1^{2l} \sigma_2^{2l}}} = \frac{1}{2} \ln \left(\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\right)^l$$

حالت خاص: $l = 1$

if $\sigma_1 = 10\sigma_2 \rightarrow B = 0.8097$

$P_e \leq 0.2225$

if $\sigma_1 = 100\sigma_2 \rightarrow B = 1.9561 \rightarrow P_e \leq 0.0707$

هرچه اختلاف واریانس‌ها بزرگتر باشد، حد خطا کاهش می‌یابد.

میانگین‌های برابر

$\sigma_1 = 1$, $\sigma_2 = 0.01$

هرچه l بزرگتر باشد کاهش حد خطا بیشتر خواهد شد.

$$\frac{\sigma_2}{\sigma_1} \to 0 \longrightarrow P_e \to 0$$

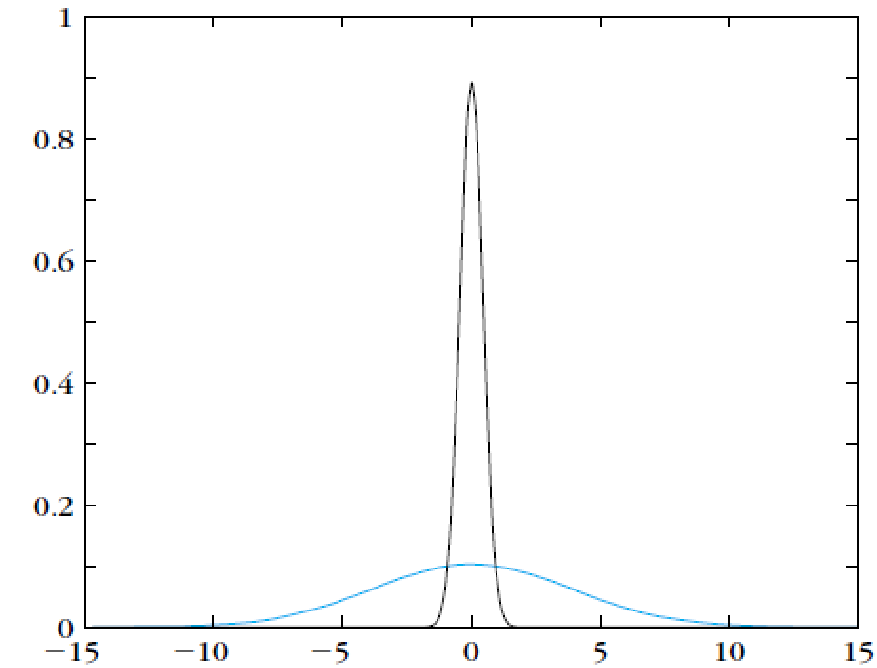


FIGURE 5.4

Gaussian pdfs with the same mean and different variances.

5.6.3 ماتریس‌های پراکندگی

- **Scatter Matrices.** These are used as a measure of the way data are scattered in the respective feature space.

چگونگی پراکندگی داده‌ها در فضای ویژگی را اندازه‌گیری می‌کند.

  - **Within-class** scatter matrix

ماتریس پراکندگی درون کلاس

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i$$

Σ_i : ماتریس کوواریانس کلاس ω_i

where

$$\Sigma_i = E\left[\left(\underline{x} - \underline{\mu}_i \right) \left(\underline{x} - \underline{\mu}_i \right)^T \right]$$

and

$$P_i \equiv P(\omega_i) \approx \frac{n_i}{N}$$

n_i the number of training samples in ω_i.

ردِ ماتریس S_w

Trace $\{S_w\}$ is a measure of the average variance of the features, over all classes.

$A_{n\times n}$ $\quad tr(A) = \sum_{i=1}^{n} a_{ii}$ مجموع عناصر قطر اصلی.

- Between-class scatter matrix

ماتریس پراکندگی بین کلاس

$$S_b = \sum_{i=1}^{M} P_i \left(\underline{\mu}_i - \underline{\mu}_0\right)\left(\underline{\mu}_i - \underline{\mu}_0\right)^T$$

میانگینِ میانگینِ کلاس‌ها

$$\underline{\mu}_0 = \sum_{i=1}^{M} P_i \underline{\mu}_i$$

Trace $\{S_b\}$ is a measure of the average distance of the mean of each class from the respective global one.

رد S_b معیاری از فاصله میانگین ، میانگین هر کلاس از میانگین سراسری است.

- Mixture scatter matrix

ماتریس پراکندگی مخلوط

$$S_m = E\left[\left(\underline{x} - \underline{\mu}_0\right)\left(\underline{x} - \underline{\mu}_0\right)^T\right]$$

It turns out that:

$$\boxed{S_m = S_w + S_b}$$

➢ Measures based on Scatter Matrices.

- $$J_1 = \frac{\mathrm{Trace}\{S_m\}}{\mathrm{Trace}\{S_w\}}$$

- $$J_2 = \frac{|S_m|}{|S_w|} = \left|S_w^{-1} S_m\right|$$

- $$J_3 = \mathrm{Trace}\left\{S_w^{-1} S_m\right\}$$

- Other criteria are also possible, by using various combinations of S_m, S_b, S_w.

The above J_1, J_2, J_3 criteria take high values for the cases where:

- Data are clustered together within each class.
- The means of the various classes are far.

حالت خاص: مسئله تک بعدی و دو کلاسه هم‌احتمال

- Fisher's discriminant ratio. In one dimension and for two equiprobable classes the determinants become:

$$|S_w| \propto \sigma_1^2 + \sigma_2^2$$

$$|S_b| \propto (\mu_1 - \mu_2)^2$$

and

$$\frac{|S_b|}{|S_w|} = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} = FDR$$

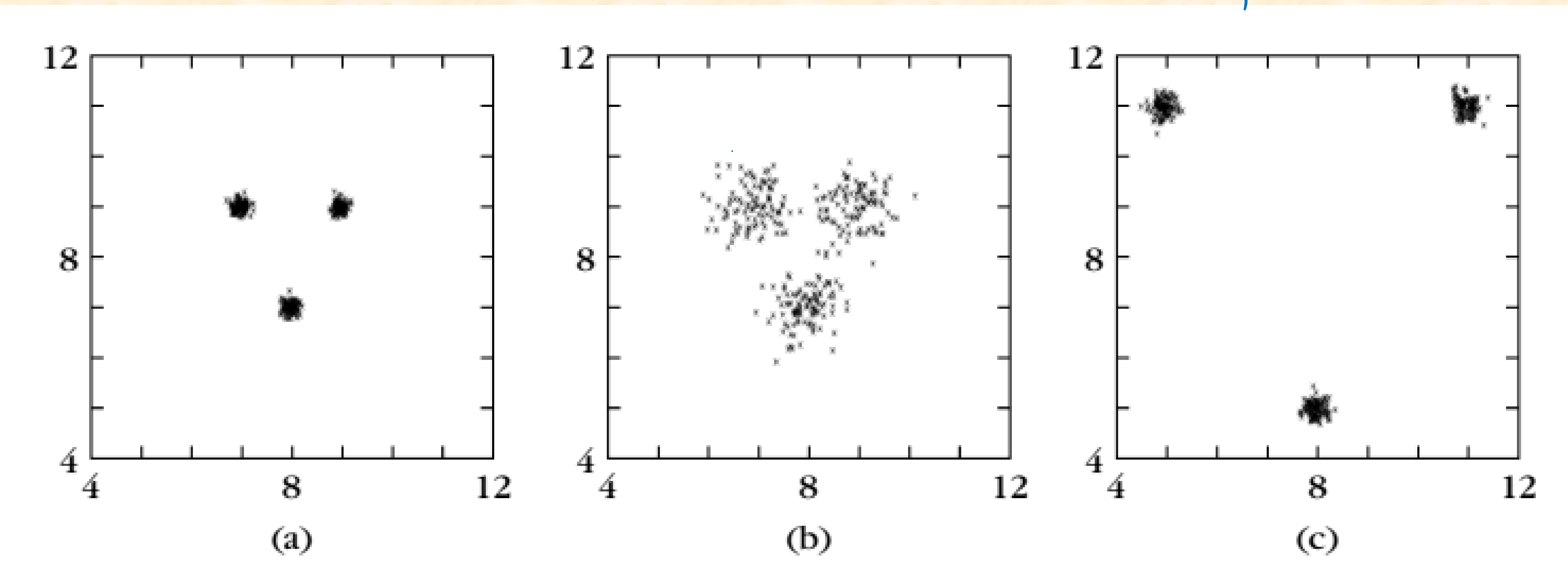
known as Fischer's ratio.

برای حالت چندکلاسه

$$FDR_1 = \sum_{i=1}^{M} \sum_{j \neq i}^{M} \frac{(\mu_i - \mu_j)^2}{\sigma_i^2 + \sigma_j^2}$$

مثال:

$J_{۳}$ = ۱۶۴٫۷ $J_{۳}$ = ۱۲٫۵ $J_{۳}$ = ۶۲۰٫۹



5.7 انتخاب زیرمجموعه ویژگی‌ها :

5.7.1 : انتخاب ویژگی اسکالر : مقادیر معیار $c(k)$ را حساب کرده، ویژگی‌ها را به صورت نزولی مرتب می‌کنیم.

$k = 1, 2, \dots, m$

l ویژگی متناظر با بهترین l مقدار $c(k)$ به عنوان بردار ویژگی انتخاب می‌شوند.

چند کلاسه :

$$c(k) = \min_{i,j} d_{ij}$$

حداقل دیورژانس روی همه جفت کلاس‌ها worst case

عیب : در روش انتخاب ویژگی اسکالر، ویژگی‌ها به صورت جداگانه مورد بررسی قرار می‌گیرند.

مزیت : پیچیدگی محاسباتی کم

برای در نظر گرفتن همبستگی correlation بین ویژگی‌ها می‌توان از روش زیر استفاده کرد :

x_{nk} ; $n = 1, 2, \dots, N$
$k = 1, 2, \dots, m$

k امین ویژگی از الگوی n ام

cross correlation coefficient ضریب همبستگی متقابل

$$\rho_{ij} = \frac{\sum_{n=1}^{N} x_{ni} x_{nj}}{\sqrt{\sum_{n=1}^{N} x_{ni}^2 \sum_{n=1}^{N} x_{nj}^2}}$$

رابطه 5.13

$|\rho_{ij}| \leq 1$

الگوریتم انتخاب ویژگی با در نظر گرفتن همبستگی متقابل:

- یک معیار جداپذیری کلاسها (C) را در نظر می‌گیریم و مقادیر آن را به ازای تمام ویژگی‌های موجود x_k، $k=1,2,\dots,m$ به دست می‌آوریم. این مقادیر به ترتیب نزولی مرتب می‌کنیم و ویژگی با بهترین مقدار C را انتخاب می‌کنیم.

فرض کنید این ویژگی x_{i_1} باشد.

- برای انتخاب ویژگی دوم، ضریب همبستگی متقابل که با رابطهٔ ρ_{ij} تعریف شد را بین x_{i_1} و $(m-1)$ ویژگی باقیمانده محاسبه می‌کنیم: $\rho_{i_1 j}$ برای $j \neq i_1$

- ویژگی x_{i_2} را طوری انتخاب می‌کنیم که:

$$i_2 = \arg\max_{j} \left\{ \alpha_1 C(j) - \alpha_2 |\rho_{i_1 j}| \right\}, \quad \text{for all } j \neq i_1$$

α_1 و α_2 ضرایب وزن‌دهی

به این ترتیب ویژگی را به عنوان ویژگی دوم انتخاب می‌کنیم که علاوه بر C زیاد، همبستگی متقابل کمتری با ویژگی اول داشته باشد.

- برای x_{i_k}، $k=3,\dots,l$:

میانگین همبستگی با ویژگی‌های قبلی

$$i_k = \arg\max_{j} \left\{ \alpha_1 C(j) - \frac{\alpha_2}{k-1} \sum_{r=1}^{k-1} |\rho_{i_r j}| \right\}$$

for $j \neq i_r,\ r=1,2,\dots,k-1$

می توان توابع هزینه های دیگری نیز در نظر گرفت :

مقاله Fine83 $\longrightarrow \left\{ \alpha_1 C_1(j) + \underline{\alpha_2 C_2(j)} - \dfrac{\alpha_3}{K-1} \sum_{r=1}^{K-1} |\rho_{irj}| \right\}$

دو معیار جدایی پذیری

5.7.2 انتخاب بردار ویژگی :

- راهکار فیلتر کردن : جستجو از بین همه حالت های ممکن :

$$\binom{m}{l} = \frac{m!}{l!(m-l)!}$$

معیار جدایی پذیری

مثال : $m=20$, $l=5$ $\longrightarrow \binom{20}{5} = 15{,}504$

مشکل پیچیدگی محاسباتی زیاد .

- راهکار Wrapper : در این راهکار به جای معیار جدایی پذیری کلاس ها، خطای طبقه بندی را در نظر می گیریم ، و ترکیبی که کمترین خطای طبقه بندی را داشته باشد را انتخاب می کنیم.

هر دو این روش‌ها، شکل پیچیده‌تر بستگی‌ها را دارند که برای حل آن می‌توان از روش‌های جستجوی زیر بهینه استفاده کرد:

Suboptimal Searching Techniques:

زیر بهینه

- انتخاب ترتیبی رو به عقب: (معکوس)

Sequential Backward Selection

فرض کنید $m = 4$ ویژگی در اختیار داریم، هدف انتخاب دو ویژگی از آن‌هاست.

x_1, x_2, x_3, x_4

- معیار جداپذیری کلاس C را در نظر می‌گیریم و مقدار آن را برای بردار $[x_1, x_2, x_3, x_4]^T$ به دست می‌آوریم.

- یک ویژگی را حذف می‌کنیم، چهار حالت ایجاد می‌شود: $[x_1, x_2, x_3]^T, [x_1, x_2, x_4]^T, [x_1, x_3, x_4]^T, [x_2, x_3, x_4]^T$

معیار را برای هر کدام از این چهار بردار به دست آوریم و ترکیب با بهترین مقدار معیار C را انتخاب می‌کنیم.

فرض کنید $[x_1, x_2, x_3]^T$ انتخاب شود.

- یک ویژگی را حذف کرده، سه حالت ایجاد می‌شود. $[x_1, x_2]^T, [x_1, x_3]^T, [x_2, x_3]^T$

معیار را برای هر کدام، بهترین مقدار را انتخاب می‌کنیم. مثلاً $[x_1, x_2]^T$ ✓ بردار با دو ویژگی

در مقایسه با حالت جستجوی کامل $\binom{m}{l}$ در این روش تعداد ترکیب ها از رابطه زیر به دست می آید:

$$1 + \frac{l}{2}\left((m+1)m - l(l+1)\right)$$ است که 5.15

- انتخاب ترتیبی رو به جلو (مستقیم) Sequential Forward Selection

- معیار را برای همه ویژگی ها محاسبه می کنیم، و بهترین ویژگی را انتخاب می کنیم مثلاً x_1

- برای همه ترکیب های دوتایی که شامل x_1 می شوند نیز $[x_1,x_2]^T, [x_1,x_3]^T, [x_1,x_4]^T$

معیار را محاسبه کرده و ترکیب با بهترین مقدار معیار را انتخاب می کنیم مثلاً $[x_1,x_2]^T$

اگر $l = 3$ یا بیشتر باشد روال را ادامه می دهیم.

در این حالت تعداد ترکیب ها $lm - l(l-1)/2$ می باشد.

سوال: در چه حالت هایی از چه روشی استفاده کنیم؟

Forward: اگر l کوچک و به 1 نزدیک باشد از روش

Backward: اگر l بزرگ و به m نزدیک باشد از روش

❖ Ways to combine features:

Trying to form all possible combinations of ℓ features from an original set of m selected features is a computationally hard task. Thus, a number of suboptimal searching techniques have been derived.

- **Sequential backward selection**. Let x_1, x_2, x_3, x_4 the available features (m=4). The procedure consists of the following steps:
  - Adopt a class separability criterion (could also be the error rate of the respective classifier). Compute its value for ALL features considered jointly $[x_1, x_2, x_3, x_4]^T$.
  - Eliminate one feature and for each of the possible resulting combinations, that is $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, compute the class reparability criterion value C. Select the best combination, say $[x_1, x_2, x_3]^T$.

- From the above selected feature vector eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T$, , $[x_2, x_3]^T$ compute $[x_1, x_3]^T$ and C select the best combination.

The above selection procedure shows how one can start from m features and end up with the "best" ℓ ones. Obviously, the choice is suboptimal. The number of required calculations is:

$$1+\frac{1}{2}\left((m+1)m-\ell(\ell+1)\right)$$

In contrast, a full search requires:

$$\binom{m}{\ell}=\frac{m!}{\ell!(m-\ell)!}$$

operations.

- Sequential forward selection. Here the reverse procedure is followed.
  - Compute C for each feature. Select the "best" one, say x_1
  - For all possible $2D$ combinations of x_1, i.e., $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$ compute C and choose the best, say $[x_1, x_3]$.
  - For all possible $3D$ combinations of $[x_1, x_3]$, e.g., $[x_1, x_3, x_2]$, etc., compute C and choose the best one.

The above procedure is repeated till the "best" vector with ℓ features has been formed. This is also a suboptimal technique, requiring:

$$\ell m - \frac{\ell(\ell-1)}{2}$$

operations.

- روشهای جستجو در Floating Search Methods:

در روشهای قبل در صورت حذف یک ویژگی یا اضافه شدن یک ویژگی فرصتی برای اضافه شدن یا حذف مجدد آنها وجود ندارد.

به این مشکل اثر آشیانه‌ای گفته می‌شود Nesting Effect

برای جبران آن از روش زیر استفاده می‌کنیم. Floating forward selection:

گام اول: اضافه کردن Inclusion: ویژگی که به همراه مجموعه قبلی بهترین مقدار C را نتیجه می‌دهد انتخاب می‌کنیم

گام دوم: تست Test: ویژگی را که با حذف شدن کمترین اثر روی مجموعه ویژگی‌ها را دارد پیدا می‌کنیم

گام سوم: حذف Exclusion: ویژگی گام قبل را حذف و به مرحله اول می‌رویم.

جزئیات بیشتر در درسنامه کتاب P. 287

مقداردهی اولیه: Sequential forward برای به دست آوردن X_2

جواب بهتر ولی با پیچیدگی بیشتر

➢ Floating Search Methods

The above two procedures suffer from the nesting effect. Once a bad choice has been done, there is no way to reconsider it in the following steps.

In the floating search methods one is given the opportunity in reconsidering a previously discarded feature or to discard a feature that was previously chosen.

The method is still suboptimal, however it leads to improved performance, at the expense of complexity.

➢ Remarks:

- Besides suboptimal techniques, some optimal searching techniques can also be used, provided that the optimizing cost has certain properties, e.g., monotonic.

- Instead of using a class separability measure (filter techniques) or using directly the classifier (wrapper techniques), one can modify the cost function of the classifier appropriately, so that to perform feature selection and classifier design in a single step (embedded) method.

- For the choice of the separability measure a multiplicity of costs have been proposed, including information theoretic costs.

5.8 تولید ویژگی‌های بهینه Optimal Feature Generation

تاکنون از معیارهای جدایی‌پذیری به شیوهٔ غیرفعال استفاده شده به این معنی برای اندازه‌گیری میزان کارایی ویژگی‌های انتخاب شده در طبقه‌بندی از این معیارها بهره بردیم.

در این بخش از معیارهای جدایی‌پذیری به شیوهٔ فعال و در فرآیند تولید ویژگی‌ها استفاده می‌کنیم.

روش LDA کار Fisher

Linear discriminant analysis آنالیز تمایز خطی

ـ حالت دوکلاسه: $\underline{x}$ بردار ویژگی‌ها در فضای m بعدی از دو کلاس

هدف: تولید ویژگی y به صورت ترکیب خطی از ویژگی‌های $\underline{x}$ به طوری که مسأله طبقه‌بندی را فشرده کرده، در فضای با ابعاد کمتر حل کند.

هدف به دست آوردن جهت $\underline{w}$ در فضای m بعدی که در امتداد آن دو کلاس به بهترین نحو ممکن از هم جدا شوند.

$$y = \frac{\underline{w}^T \underline{x}}{\|\underline{w}\|}$$ تصویر $\underline{x}$ در امتداد $\underline{w}$

$$y = w_1 x_1 + w_2 x_2 + \cdots$$

از $\|\underline{w}\|$ به دلیل اینکه در جهت تأثیر ندارد صرف‌نظر کنیم.

Fisher Discriminant Ratio

$$\underline{FDR} = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$$

μ_1, μ_2 ، σ_1^2, σ_2^2 به ترتیب میانگین‌ها و واریانس‌های y در کلاس‌های ω_1 و ω_2 می‌باشد.

$$y = \underline{w}^T \underline{x} \longrightarrow \mu_i = \underline{w}^T \underline{\mu}_i \quad , \quad i=1,2$$

$$\boxed{(\mu_1-\mu_2)^2} = \underline{w}^T (\underline{\mu}_1-\underline{\mu}_2)(\underline{\mu}_1-\underline{\mu}_2)^T \underline{w} \propto \boxed{\underline{w}^T S_b \underline{w}}$$

$$\sigma_i^2 = E[(y-\mu_i)^2] = E[\underline{w}^T(\underline{x}-\underline{\mu}_i)(\underline{x}-\underline{\mu}_i)^T\underline{w}] = \underline{w}^T \Sigma_i \underline{w} \qquad i=1,2$$

$$\boxed{\sigma_1^2+\sigma_2^2 \propto \underline{w}^T S_w \underline{w}}$$

$$FDR = \frac{\underline{w}^T S_b \underline{w}}{\underline{w}^T S_w \underline{w}}$$

با بیشینه سازی FDR نسبت به $\underline{w}$ به جواب می‌رسیم. این رابطه وقتی بیشینه می‌شود که:

$$S_b \underline{w} = \lambda S_w \underline{w}$$

بزرگترین مقدار ویژه ماتریس $S_w^{-1} S_b$

در اینجا نیازی به مسئله ما نیست:

$$\lambda S_w \underline{w} = S_b \underline{w} = (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{w} = \alpha (\underline{\mu}_1 - \underline{\mu}_2)$$

$$\cancel{\alpha / \lambda} \longrightarrow \boxed{\underline{w} = S_w^{-1}(\underline{\mu}_1 - \underline{\mu}_2)}$$ جواب LDA:

$$y = \underline{w}^T \underline{x}$$

مثال: شکل 5.6

حالت خاص: توزیع گاوسی برای کلاس‌ها با کوواریانس برابر

در این حالت حل روش LDA به طبقه‌بند بهینه بیز منجر شود.

$$g(\underline{x}) = \underbrace{(\underline{\mu}_1 - \mu_2)^T S_w^{-1}}_{\underline{w}^T} \underline{x} + w_0$$

3.14

طراحی همزمان طبقه‌بند با ترکیب ویژگی‌ها

$$g(\underline{x}) = (\underline{\mu}_1 - \mu_2)^T S_w^{-1} \left(\underline{x} - \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)\right) - \ln \frac{P(\omega_2)}{P(\omega_1)}$$

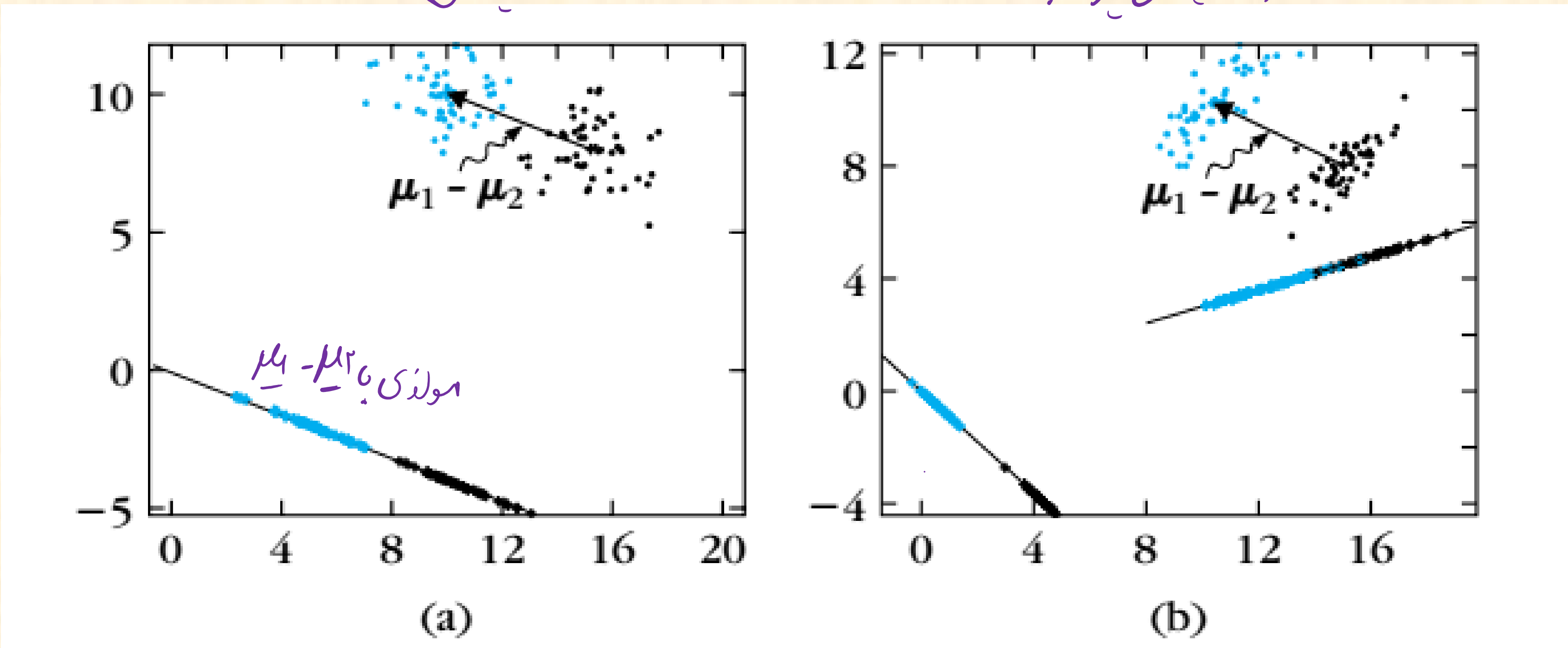
برای داده‌های غیر گاوسی نیز قابل استفاده است.

5.6

LDA: Linear Discriminant Analysis
2 Classes
2 Feature to 1 Feature

حالت کوواریانس قطری

حالت کوواریانس غیر قطری



حالت چند کلاسه:

$M > 2$

$$\underset{l\times 1}{\underline{y}} = \underset{l\times m}{A^T}\,\underline{x}_{m\times 1}$$

l تایی

m تایی

از این نیز میتوان $J_3 = \text{trace}\{S_w^{-1} S_b\}$ استفاده کرد.

$$S_{yw} = A^T S_{xw} A \qquad S_{yb} = A^T S_{xb} A$$

$$J_3(A) = \text{trace}\{(A^T S_{xw} A)^{-1} (A^T S_{xb} A)\}$$

problem 5.17

$$\frac{\partial J_3(A)}{\partial A} = 0 \longrightarrow \boxed{(S_{xw}^{-1} S_{xb})A = A(S_{yw}^{-1} S_{yb})}$$

$$\boxed{B^T S_{yw} B = I \quad ; \quad B^T S_{yb} B = D_{l\times l}}$$

قطری

$$\hat{y} = \underset{l\times l}{B^T} y = B^T A^T x$$

$$J_3(\hat{y}) = J_3(y)$$

5.44 $\quad (S_{xw}^{-1} S_{xb})\, \underset{m\times l}{C} = CD$

AB

- حالت $l = M-1$:

$$\hat{y} = C^T \underline{x} \longrightarrow J_{3_{max}}$$

$$J_{3,\underline{\hat{y}}} = J_{3,\underline{x}}$$

برای $\underline{M=2}$:

$$\hat{y} = (\underline{\mu}_1 - \underline{\mu}_2)^T S_{xw}^{-1} \underline{x}$$

رابطه‌ی LDA

- حالت $l < M-1$:

$$J_{3,\hat{y}} < J_{3,\underline{x}}$$

بیان هندسی:

5.7:

v_i ها بردارهای ویژه $S_w^{-1} S_b$