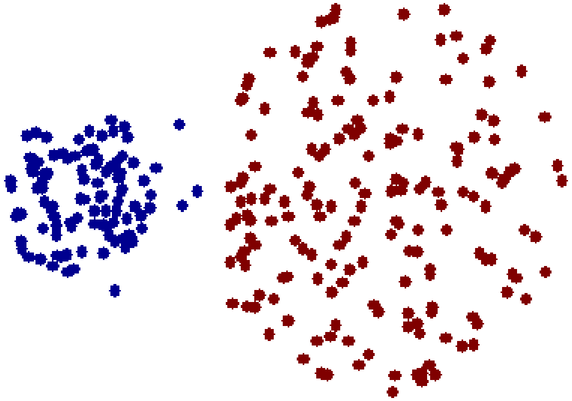
فصل پنجم
شناسایی الگو
انتخاب ویژگی
Feature Selection

محمدجواد فدائی‌اسلام

Feature selection or feature reduction

- از میان تعدادی ویژگی، چگونه می توان مهمترین آنها را انتخاب کرد تا تعداد آنها کاهش یابد و در عین حال اطلاعات در آنها برای جداکنندگی بین کلاس‌ها بیشتر حفظ شود.
- هدف ما باید انتخاب ویژگی هایی باشد که در فضای ویژگی فاصله بین کلاس‌ها را زیاد و پراکندگی درون کلاسی را کاهش دهد.



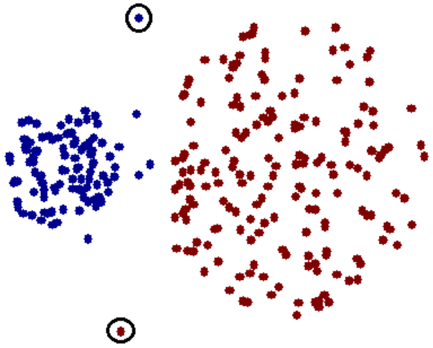
رهیافت‌های انتخاب ویژگی

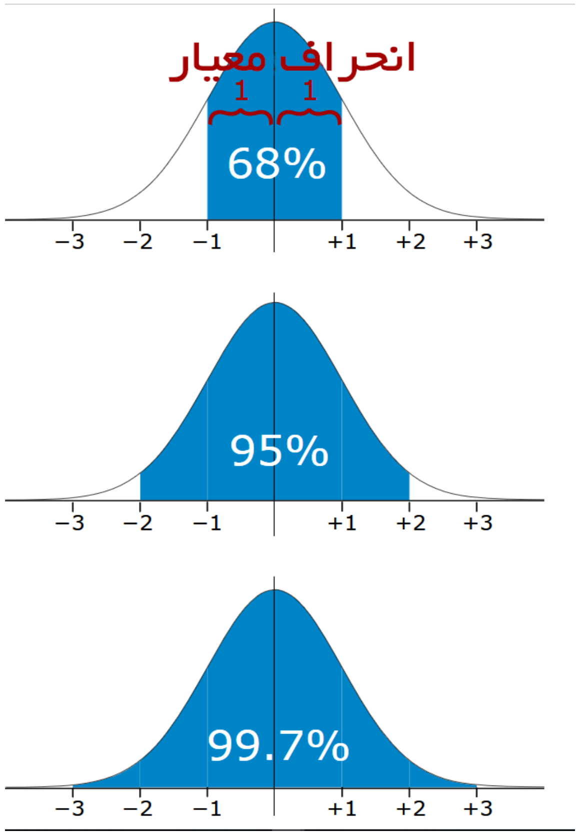
دو رهیافت برای انتخاب ویژگی وجود دارد:

- ویژگی ها را به صورت جداگانه بررسی کرده و ویژگی هایی را که توانایی جداکنندگی کمی دارند، کنار بگذارید.
- گزینه بهتر، بررسی ترکیب ویژگی‌ها است.

پیش پردازش داده‌ها: حذف داده‌های دور افتاده
OUTLIER REMOVAL

- داده دورافتاده به عنوان نقطه‌ای تعریف می‌شود که از میانگین داده‌ها بسیار خیلی دور است.
- در توزیع‌های نرمال، در فاصله‌ای به اندازه دو برابر انحراف استاندارد از میانگین ۹۵ درصد داده‌ها قرار می‌گیرند.





پیش‌پردازش داده‌ها: حذف داده‌های دورافتاده (ادامه)
OUTLIER REMOVAL

- اگر تعداد دورافتاده‌ها بسیارکم باشد، معمولاً دور ریخته می‌شوند.
- اگر اینگونه نباشد و آنها نتیجه توزیع با دم‌های طولانی باشند (پراکندگی داده‌ها زیاد باشد)، ممکن است طراح مجبور شود توابع هزینه‌ای را انتخاب کند که به نقاط دورافتاده حساس نیستند.
- به عنوان مثال، معیار کمترین مربعات نسبت به دورافتاده‌ها بسیار حساس است، زیرا خطاهای بزرگ به دلیل مجذورشدن در تابع هزینه، غالب عملکرد را تشکیل می‌دهند.

پیش‌پردازش داده‌ها: نرمال‌سازی داده‌ها
DATA NORMALIZATION

- ویژگی‌های با دامنه بزرگ ممکن است تأثیر بیشتری (غالبی) نسبت به ویژگی‌های با دامنه کم در تابع هزینه داشته باشند.
  - مثال: قد به متر در بازه [۰,۵ ،۲] قرار دارد و وزن در دامنه [۴۰ ،۱۵۰]
- نرمال‌سازی همه ویژگی‌ها باعث می‌شود تا مقادیر آنها در محدوده‌های مشابه قرار گیرند.

نرمال‌سازی داده‌ها- روش خطی
DATA NORMALIZATION – LINEAR METHOD

For N available data of the kth feature we have

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik}, \quad k = 1, 2, \ldots, l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_{ik} - \bar{x}_k)^2$$

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

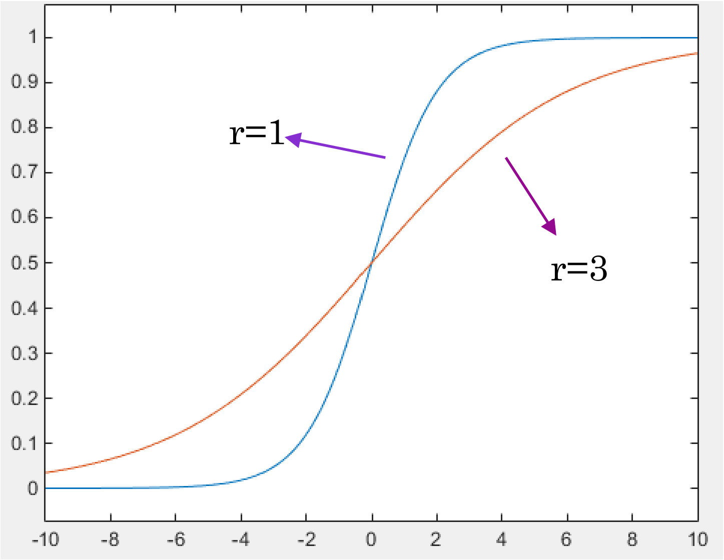
نرمال‌سازی داده‌ها- روش غیرخطی
DATA NORMALIZATION – NONLINEAR METHOD

The so-called softmax scaling is a popular candidate.
It consists of two steps

$$y = \frac{x_{ik} - \bar{x}_k}{r\sigma_k}, \quad \hat{x}_{ik} = \frac{1}{1 + \exp(-y)}$$

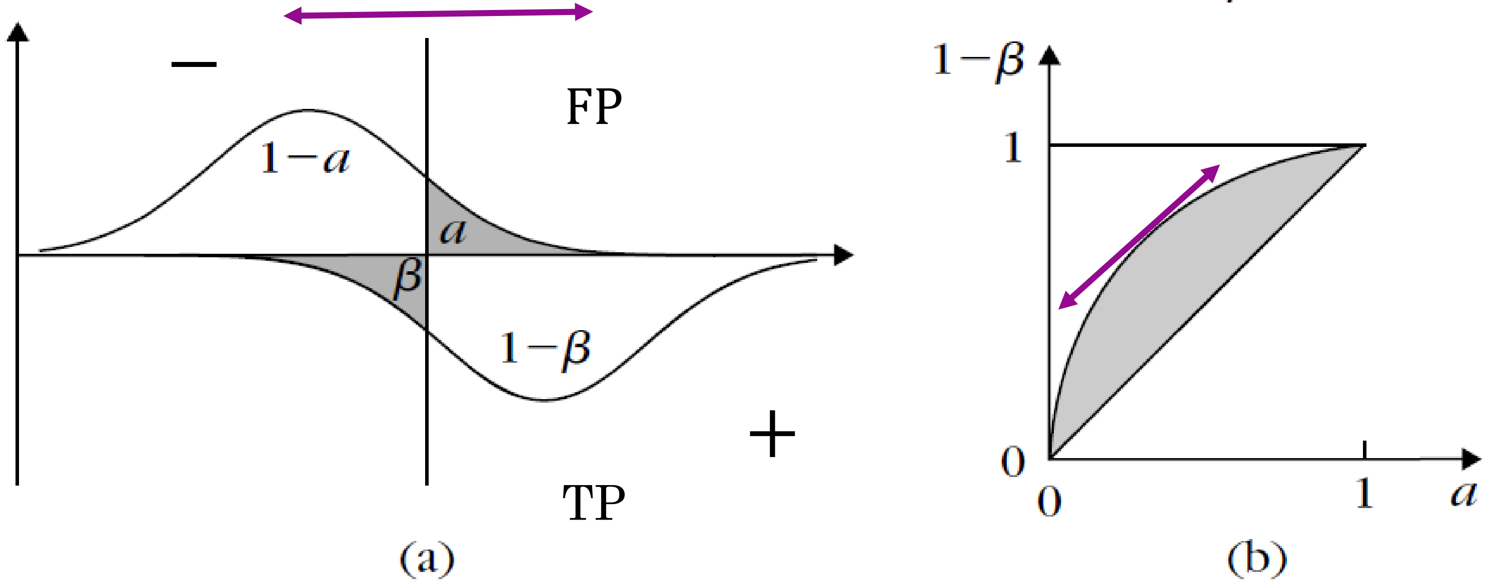
$$S(x) = \frac{1}{1 + e^{-x}}$$

$S(x)$ = sigmoid function

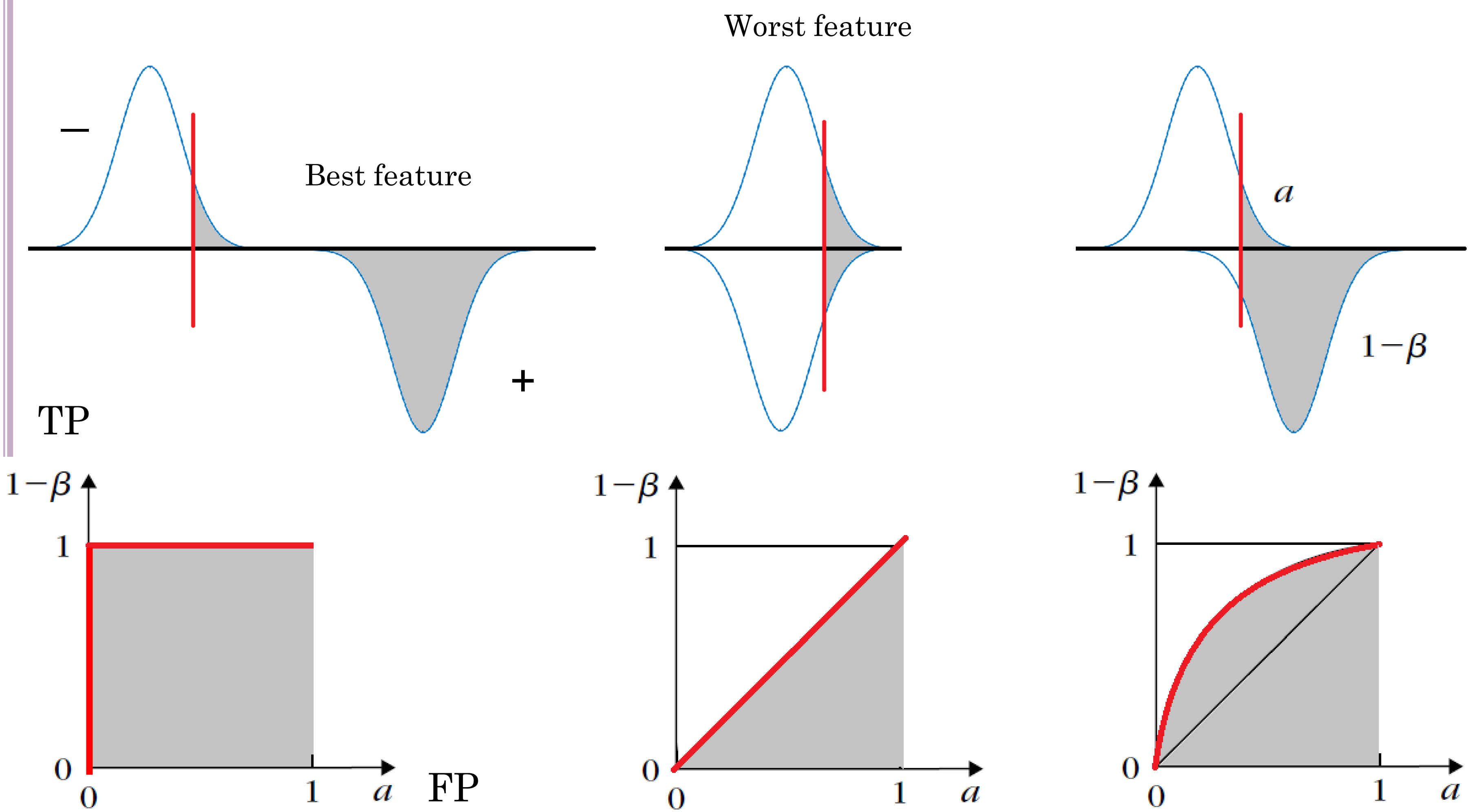


ROC CURVE

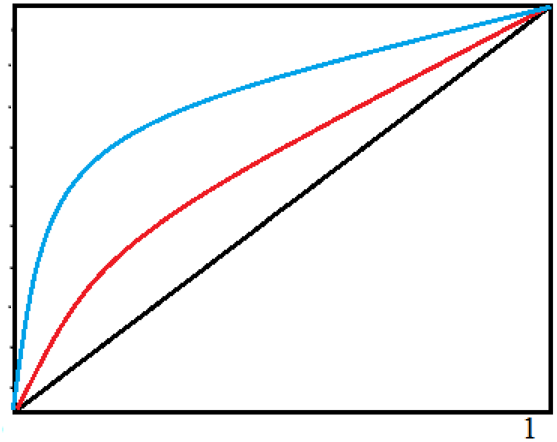
By moving the threshold over "all" possible positions, different values of a and β result.



سه نمودار ROC برای سه ویژگی متفاوت



ROC - AUC



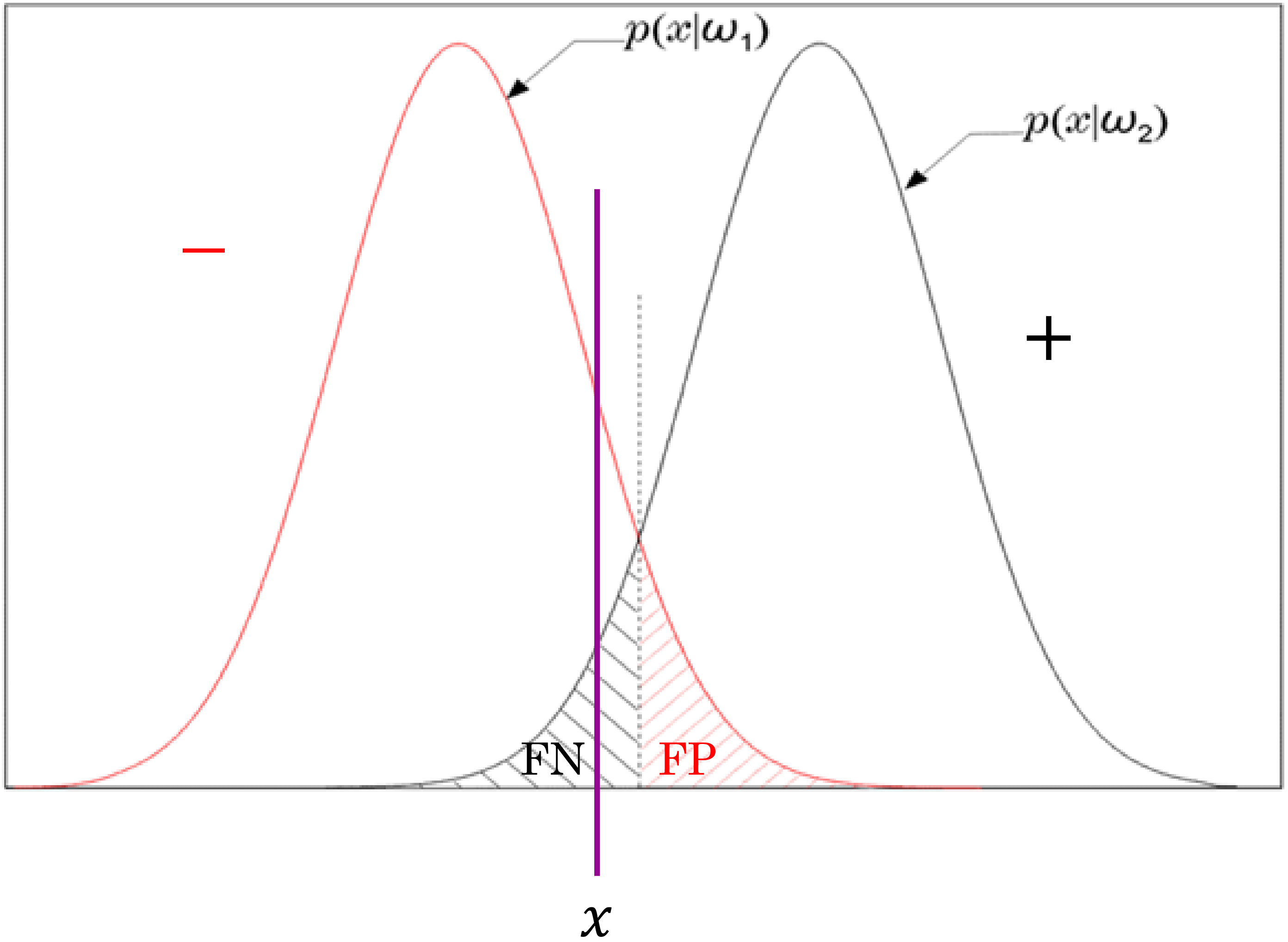
طبقه‌بند فیروزه‌ای از طبقه‌بند قرمز کارایی بهتری دارد. یا به عبارتی ویژگی که طبقه‌بند فیروزه‌ای با آن ساخته شده است از ویژگی قرمز بهتر است.

- Max AUC = 1

اندازه‌گیری تفکیک‌پذیری کلاس
Class Separability Measures

تاکید فصل قبل، بر روی جداکنندگی ویژگی‌ها به صورت انفرادی بود. اما این روش‌ها همبستگی که میان ویژگی‌ها وجود دارد و توانایی بردار ویژگی برای کلاس‌بندی را تحت تاثیر قرار می‌دهد را مورد توجه قرار نمی‌دهد. اندازه‌گیری جداکنندگی **یک بردار ویژگی** (ترکیب ویژگی‌ها) در اینجا مورد توجه ما است.

Classification Error - Review



DIVERGENCE

$$P(\omega_1|x) = \frac{p(x|\omega_1)P(\omega_1)}{p(x)}$$

- با توجه به قانون بیز، داده با بردار ویژگی x متعلق به کلاس ω_1 است اگر

$$P(\omega_1|x) > P(\omega_2|x)$$

- می‌دانیم هرچه اختلاف بین $P(\omega_1|x)$ و $P(\omega_2|x)$ بیشتر باشد کلاس‌بندی بهتر است. از این رو نسبت $\frac{P(\omega_1|x)}{P(\omega_2|x)}$ می‌تواند برای ارزیابی به کار رود. با توجه به قانون بیز داریم:

$$D_{12}(x) = ln\frac{p(x|\omega_1)}{p(x|\omega_2)}$$

- در حالت میانگین داریم:

$$D_{12} = \int_{-\infty}^{+\infty} p(x|\omega_1)\, ln\frac{p(x|\omega_1)}{p(x|\omega_2)}dx$$

- به طور مشابه برای کلاس ω_2 در حالت میانگین داریم:

$$D_{21} = \int_{-\infty}^{+\infty} p(x|\omega_2)\, ln\frac{p(x|\omega_2)}{p(x|\omega_1)}dx$$

DIVERGENCE

$$d_{12} = D_{12} + D_{21}$$

- مقدار بالا divergence نام دارد. اگر مساله چندکلاسه باشد داریم:

$$d_{ij} = D_{ij} + D_{ji}$$

- درحالت میانگین جداپذیری اینگونه محاسبه می‌شود:

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

- خصوصیات دیورجانس

- $d_{ij} \geq 0$
- $d_{ij} = 0, if\ i = j$
- $d_{ij} = d_{ji}$

Trace of matrix

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} -1 & 0 & 3 \\ 11 & 5 & 2 \\ 6 & 12 & -5 \end{pmatrix}$$

$$\text{trace}(\mathbf{A}) = \operatorname{tr}(\mathbf{A}) = \sum_{i=1}^{3} a_{ii} = a_{11} + a_{22} + a_{33} = -1 + 5 + (-5) = -1$$

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{\ell}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

DIVERGENCE-NORMAL DISTRIBUTION

Assuming now that the density functions are Gaussians $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i)$ and $\mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$, respectively, the computation of the divergence is simplified, and it is not difficult to show that

$$d_{ij} = \frac{1}{2}\text{trace}\{\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I\} + \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \qquad (5.22)$$

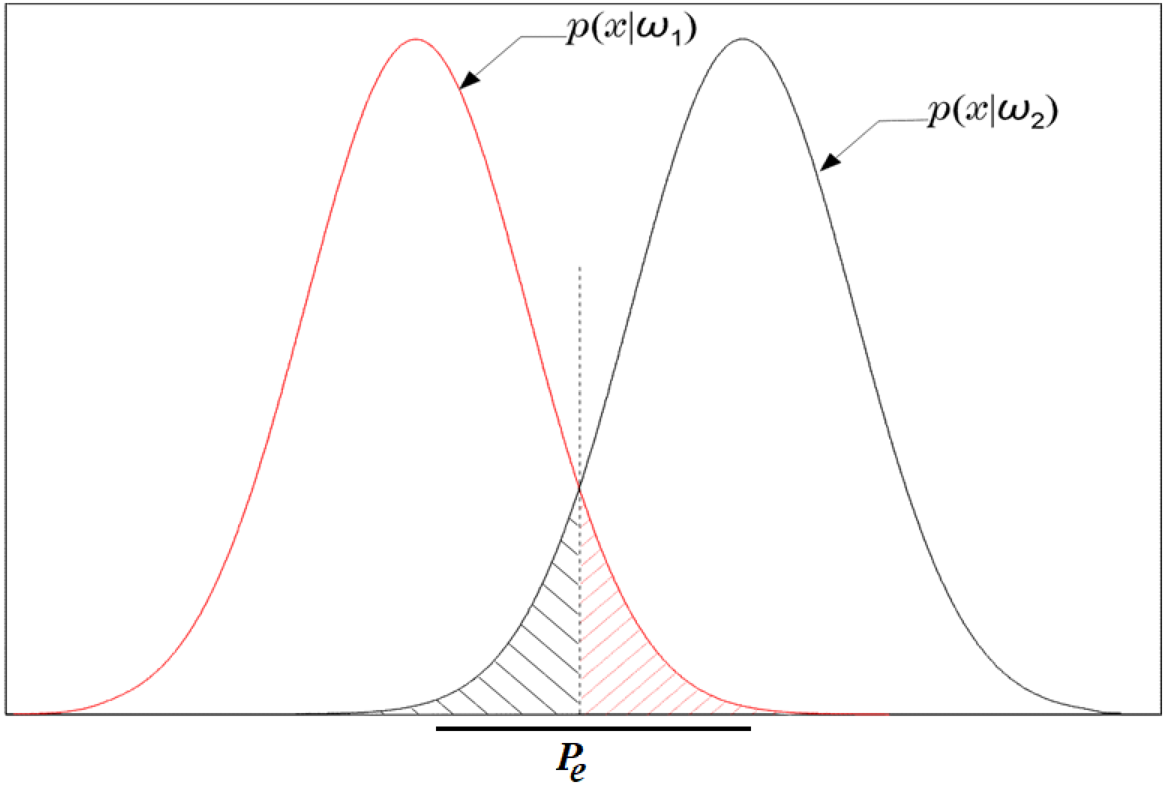
For the one-dimensional case this becomes

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

- اگر $\Sigma_i = \Sigma_j = \Sigma$؛ انگاه دیورجانس آن برابر با فاصله ماهالونوبیس می‌شود.

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

P_e



$$P_e = \int_{-\infty}^{\infty} \min\left[P(\omega_i)p(\boldsymbol{x}|\omega_i), P(\omega_j)p(\boldsymbol{x}|\omega_j)\right] d\boldsymbol{x}$$

Chernoff bound

The minimum attainable classification error of the Bayes classifier for two classes ω_1, ω_2 can be written as:

$$P_e = \int_{-\infty}^{\infty} \min\left[P(\omega_i)p(\boldsymbol{x}|\omega_i), P(\omega_j)p(\boldsymbol{x}|\omega_j)\right] d\boldsymbol{x} \tag{5.23}$$

Analytic computation of this integral in the general case is not possible. However, an upper bound can be derived. The derivation is based on the inequality

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

Combining (5.23) and (5.24), we get

$$P_e \leq P(\omega_i)^s P(\omega_j)^{1-s} \int_{-\infty}^{\infty} p(\boldsymbol{x}|\omega_i)^s p(\boldsymbol{x}|\omega_j)^{1-s} \, d\boldsymbol{x} \equiv \epsilon_{CB} \tag{5.25}$$

ϵ_{CB} is known as the *Chernoff bound*. The minimum bound can be computed by minimizing ϵ_{CB} with respect to s.

BHATTACHARYYA DISTANCE

A special form of the bound results for $s = 1/2$:

$$P_e \leq \epsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)} \int_{-\infty}^{\infty} \sqrt{p(\boldsymbol{x}|\omega_i)p(\boldsymbol{x}|\omega_j)}\, d\boldsymbol{x} \tag{5.26}$$

For Gaussian distributions $\mathcal{N}(\boldsymbol{\mu}_i, \Sigma_i), \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j)$ and after a bit of algebra, we obtain

$$\epsilon_{CB} = \sqrt{P(\omega_i)P(\omega_j)} \exp(-B)$$

where

$$B = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^T \left(\frac{\Sigma_i + \Sigma_j}{2}\right)^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2} \ln \frac{\left| \frac{\Sigma_i + \Sigma_j}{2} \right|}{\sqrt{|\Sigma_i||\Sigma_j|}} \tag{5.27}$$

- $|.|$: determinant
- $B: Bhattacharyya\ distance$
- اگر $\Sigma_i = \Sigma_j$ باشد، آنگاه فاصله باتاچاریا با فاصله ماهالانوبیس نسبت دارد.

EXAMPLE

Assume that $P(\omega_1) = P(\omega_2)$ and that the corresponding distributions are Gaussians $\mathcal{N}(\boldsymbol{\mu}, \sigma_1^2 I)$ and $\mathcal{N}(\boldsymbol{\mu}, \sigma_2^2 I)$. The Bhattacharyya distance becomes

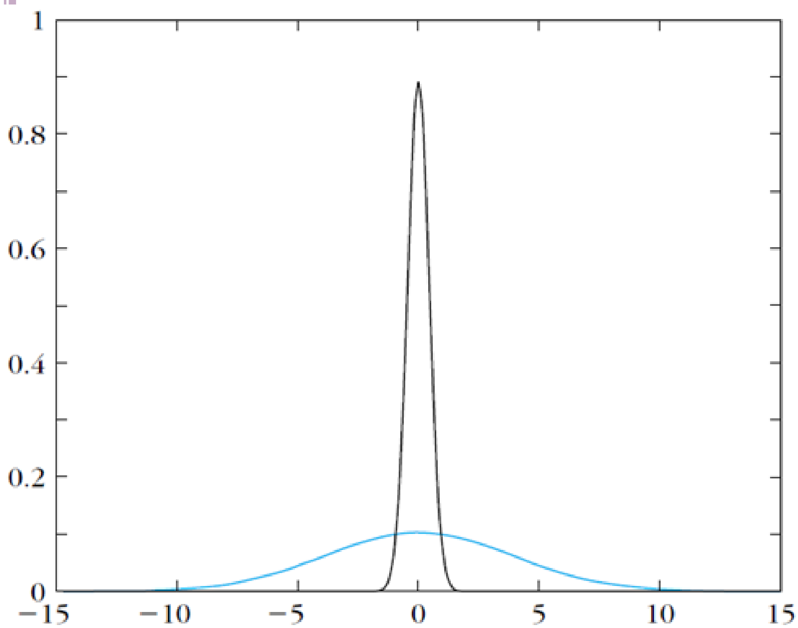
$$B = \frac{1}{2} \ln \frac{\left(\frac{\sigma_1^2+\sigma_2^2}{2}\right)^l}{\sqrt{\sigma_1^{2l}\sigma_2^{2l}}} = \frac{1}{2} \ln \left(\frac{\sigma_1^2+\sigma_2^2}{2\sigma_1\sigma_2}\right)^l$$

For the one-dimensional case $l = 1$ and for $\sigma_1 = 10\sigma_2$, $B = 0.8097$ and

$$P_e \leq 0.2225$$

If $\sigma_1 = 100\sigma_2$, $B = 1.9561$ and

$$P_e \leq 0.0707$$



cognition

انتخاب ویژگی- قسمت ۲
Feature Selection-part II

ماتریس‌های پراکندگی
SCATTER MATRICES

یک عیب عمده معیارهای تفکیک کلاس‌بندی که تاکنون در نظر گرفته شده این است که به راحتی محاسبه نمی‌شوند، مگر اینکه فرض گاوسی به کار گرفته شود. ما اکنون توجه خود را به مجموعه‌ای از معیارهای ساده‌تر، متکی بر اطلاعات مربوط به نحوه پراکندگی بردارهای ویژگی در فضای l بعدی خواهیم کرد.

WITHIN-CLASS SCATTER MATRIX

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

where Σ_i is the covariance matrix for class ω_i

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

and P_i the *a priori* probability of class ω_i.

BETWEEN-CLASS SCATTER MATRIX

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

where $\boldsymbol{\mu}_0$ is the global mean vector

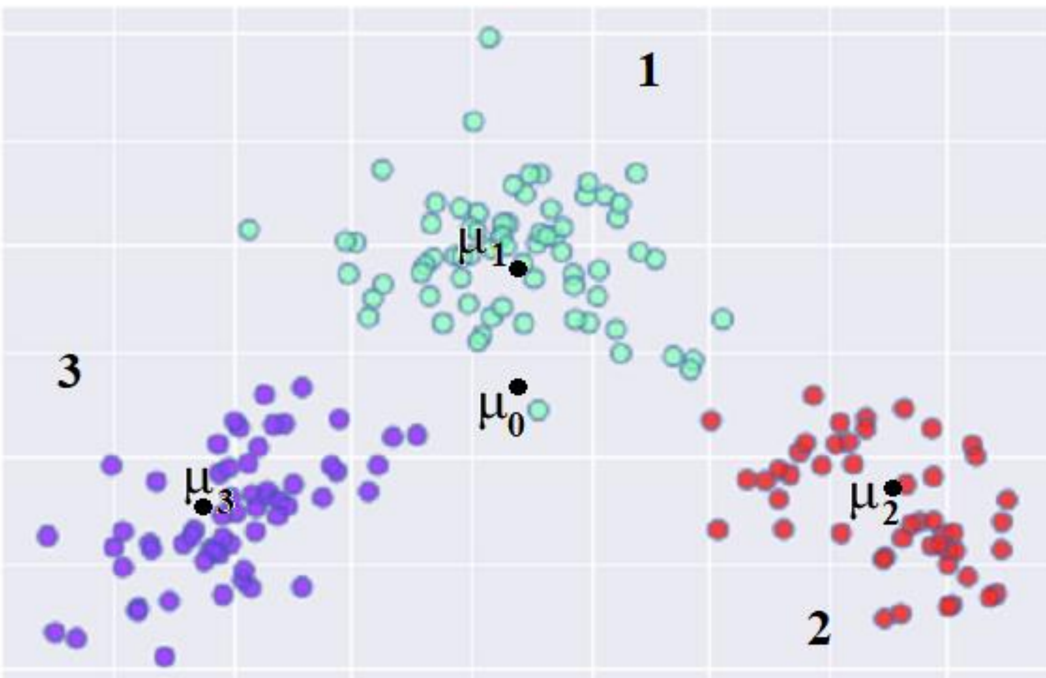
$$\boldsymbol{\mu}_0 = \sum_{i}^{M} P_i \boldsymbol{\mu}_i$$

MIXTURE SCATTER MATRIX

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

That is, S_m is the covariance matrix of the feature vector with respect to the global mean. It is not difficult to show (Problem 5.12) that

$$S_m = S_w + S_b$$



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

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

$$S_m = S_w + S_b$$

Criterion based on trace

$$J_1 = \frac{\mathrm{trace}\{S_m\}}{\mathrm{trace}\{S_w\}}$$

$$\underbrace{S_m = S_w}_{\text{Within}} + \underbrace{S_b}_{\text{Between}}$$

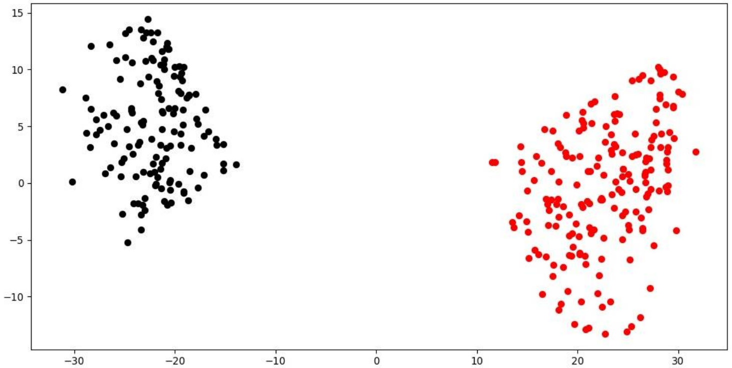
$$S_w = \sum_{i=1}^{M} P_i \Sigma_i$$

Criterion based on determinant

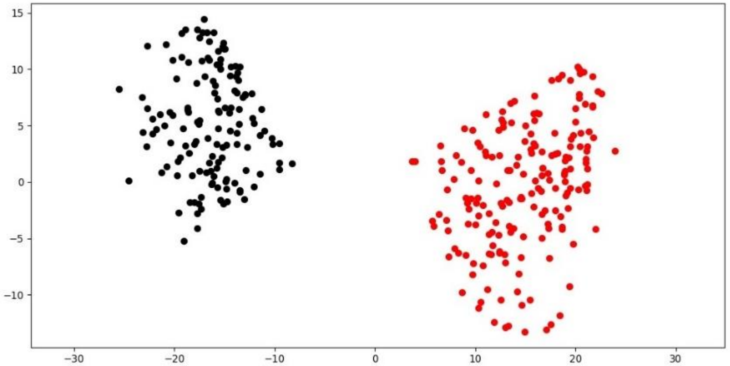
$$J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$$

$$J_3 = \mathrm{trace}\{S_w^{-1} S_m\}$$

SCATTER MATRIX



A



B

$$S_{WA} = S_{WB}$$

$$S_{bA} \geq S_{bB} \Longrightarrow S_{mA} \geq S_{mB} \Longrightarrow J_{1A} \geq J_{1B}$$

SCATTER MATRIX (SPECIAL FORM, 1DIM.)

These criteria take a special form in the one-dimensional, two-class problem. In this case, it is easy to see that for equiprobable classes $|S_w|$ is proportional to $\sigma_1^2 + \sigma_2^2$ and $|S_b|$ proportional to $(\mu_1 - \mu_2)^2$. Combining S_b and S_w, the so-called *Fisher's discriminant ratio (FDR)* results

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

For the multiclass case, averaging forms of *FDR* can be used. One possibility is

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

where the subscripts i, j refer to the mean and variance corresponding to the feature under investigation for the classes ω_i, ω_j, respectively.

SCATTER MATRIX (SPECIAL FORM, 1DIM.)

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

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

$$S_m = S_w + S_b$$

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

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

انتخاب مجموعه ویژگی
FEATURE SUBSET SELECTION

- تاکنون معیارهایی برای سنجش کارایی ویژگی‌های منفرد و یا بردارهای ویژگی در طبقه‌بندی ارایه شد. با توجه به این معیارها به انتخاب تعدادی ویژگی می‌پردازیم.
- انتخاب ویژگی به صورت اسکالر
  - ویژگی‌ها به صورت جداگانه بررسی می‌شوند. هر یک از معیارهای اندازه‌گیری می‌تواند برای ارزیابی تفکیک‌پذیری کلاس در نظر گرفته شود (مثل: ROC، FDR، و ....). برای هر ویژگی مقدار معیار محاسبه می‌شود این مقادیر مرتب شده و بر اساس آن ویژگی انتخاب می‌شود.
- انتخاب بردار ویژگی
  - جستجوی نیمه‌بهینه پیشرو/ پسرو

(Suboptimal Forward/Backward Search)

  - تولید ویژگی بهینه

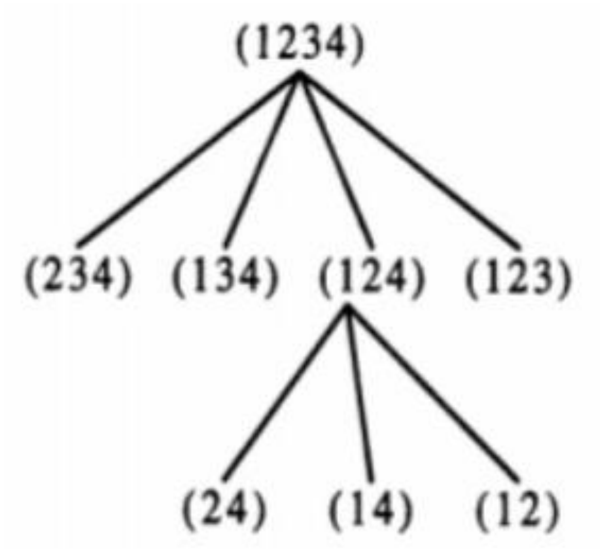
Suboptimal Searching Techniques
Sequential Backward Selection

We will demonstrate the method via an example. Let $m = 4$, and the originally available features are x_1, x_2, x_3, x_4. We wish to select two of them. The selection procedure consists of the following steps:

- Adopt a class separability criterion, C, and compute its value for the feature vector $[x_1, x_2, x_3, x_4]^T$.
- 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 corresponding criterion value. Select the combination with the best value, say $[x_1, x_2, x_3]^T$.
- From the selected three-dimensional feature vector eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T$, $[x_1, x_3]^T$, $[x_2, x_3]^T$, compute the criterion value and select the one with the best value.

Sequential Backward Selection(Example)

- انتخاب دو ویژگی از چهار ویژگی

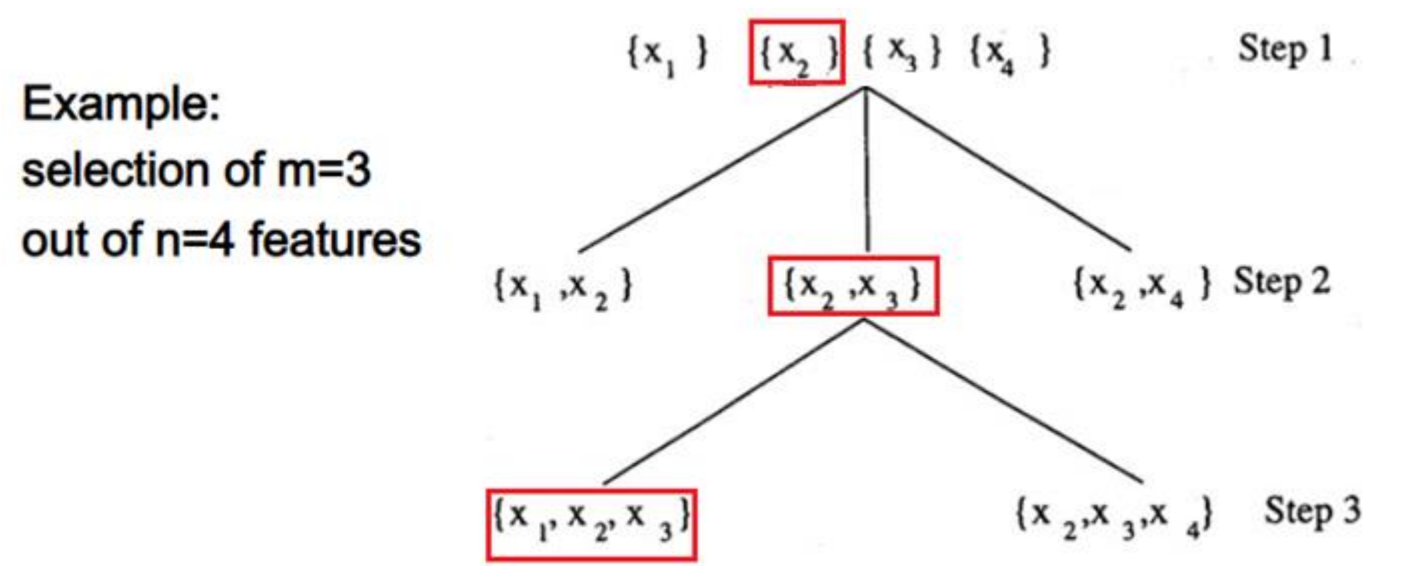


Suboptimal Searching Techniques Sequential Forward Selection

Here, the reverse to the preceding procedure is followed:

- Compute the criterion value for each of the features. Select the feature with the best value, say x_1.
- Form all possible two-dimensional vectors that contain the winner from the previous step, that is, $[x_1, x_2]^T$, $[x_1, x_3]^T$, $[x_1, x_4]^T$. Compute the criterion value for each of them and select the best one, say $[x_1, x_3]^T$.

Sequential Forward Selection (Example)

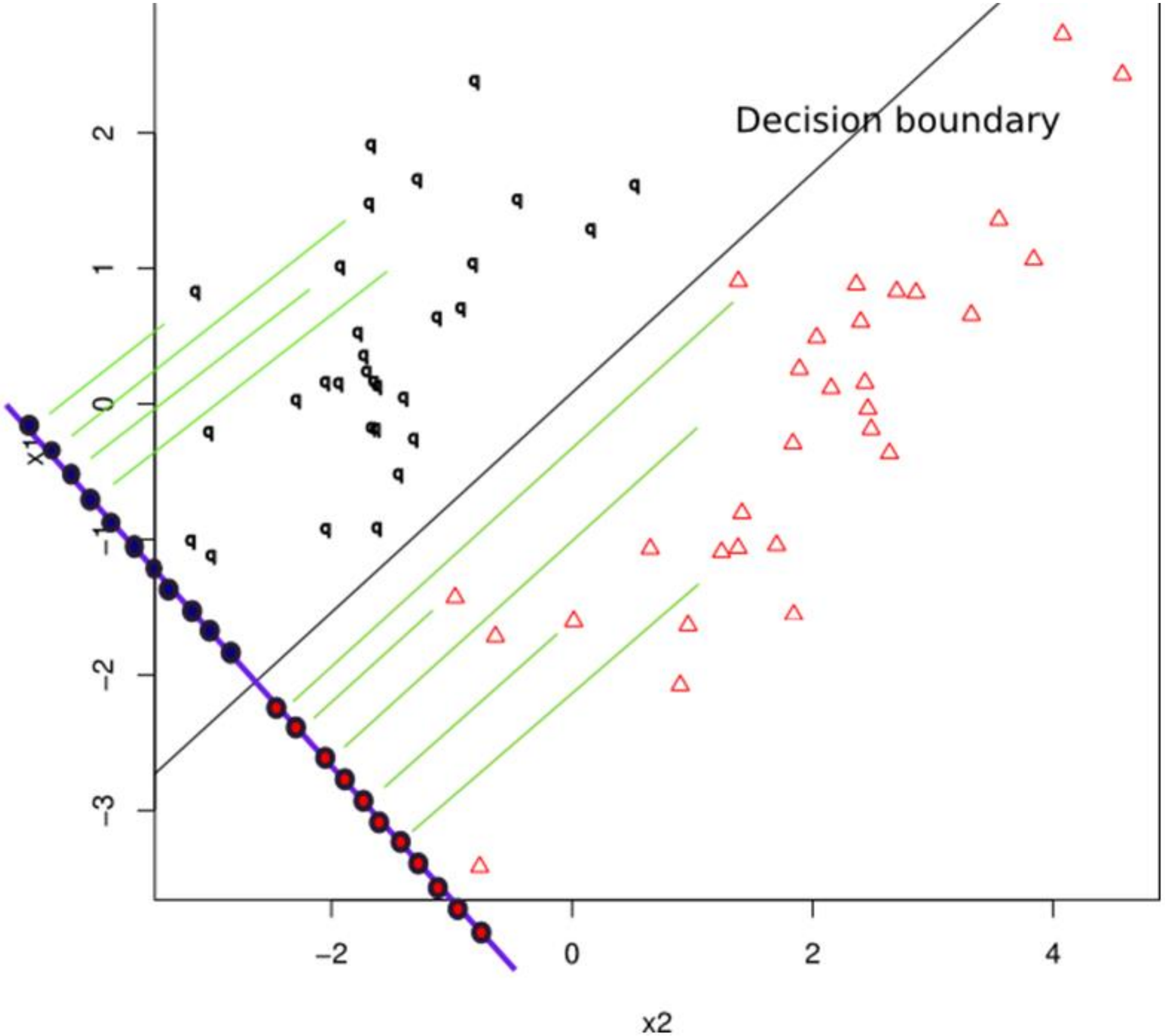


- آیا در مثال بالا کارایی زیر مجموعه $\{x_1, x_2, x_4\}$ ارزیابی شده است؟
- ایا تضمینی وجود دارد که این زیر مجموعه کارایی کمتری دارد؟

چرا روش‌های پیش‌رو و پس‌رو به بهینه سراسری نمی‌رسند

به طور مثال اگر m ویژگی داشته باشیم. برای انتخاب دو ویژگی تعداد $\binom{m}{2}$ محاسبه باید انجام پذیرد. تمام این حالات در روش پیش‌رو یا پس‌رو بررسی نمی‌شوند. بنابراین نمی‌توان ثابت کرد که به جواب بهینه می‌رسند.

LDA, Linear Discriminant Analysis



در این روش، تولید ویژگی و در عین حال طراحی یک طبقه‌بندی کننده خطی باهم انجام می‌شود.

Linear Discriminant Analysis

- در این روش مشخصات خطی (ابرصفحه) یافت می‌شود که تصویر داده‌ها بر روی آن دارای بیشترین جداکنندگی است(روش فیشر).

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

Between Class Scatter Matrix

where $\boldsymbol{\mu}_0$ is the global mean vector

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

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

Within Class Scatter Matrix

where Σ_i is the covariance matrix for class ω_i

and P_i the *a priori* probability of class ω_i

بردارهای ویژه متناظر با بیشترین مقدار ویژه $S_W^{-1}S_b$، ماتریس W را تشکیل می‌دهد.

Linear Discriminant Analysis

- روابط قبل، مساله را از یک فضای L به یک فضای با بُعد پایین می‌برد که تعداد بُعد جدید وابسته به تعداد بردار ویژه انتخاب شده است.
- برای تصویر بر روی یک خط و در یک مساله دو کلاسه می‌توان از رابطه زیر استفاده کرد.

$$v = S_w^{-1}(\mu_1 - \mu_2)$$

LDA

خط جداکننده و کوواریانس دو کلاس

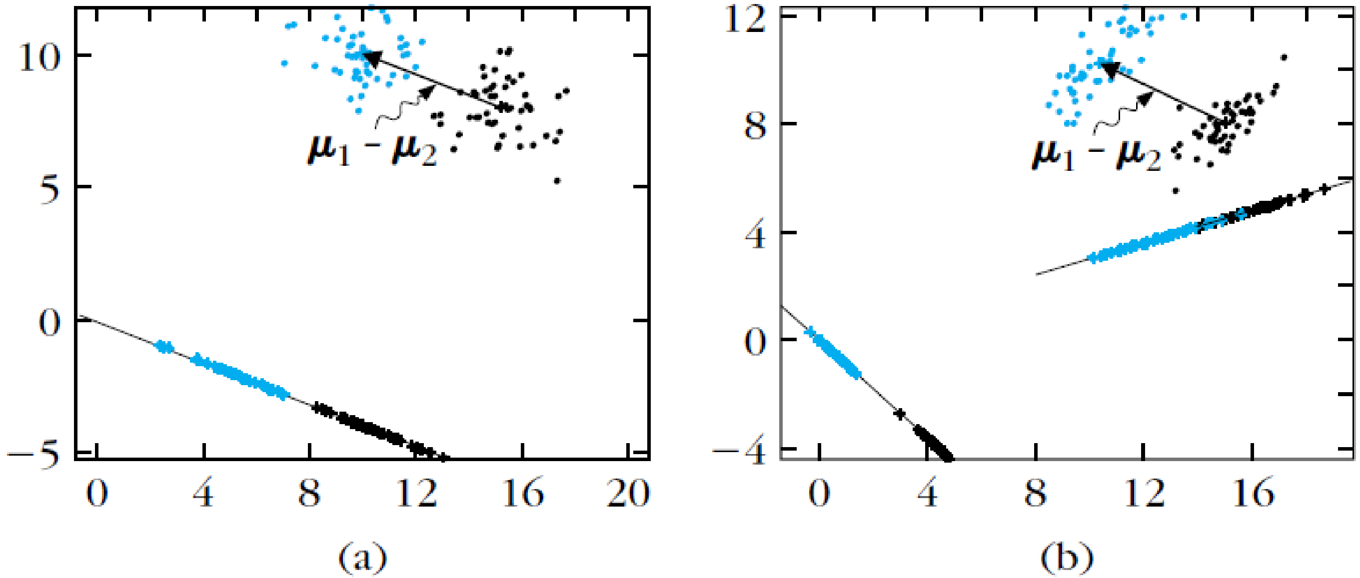
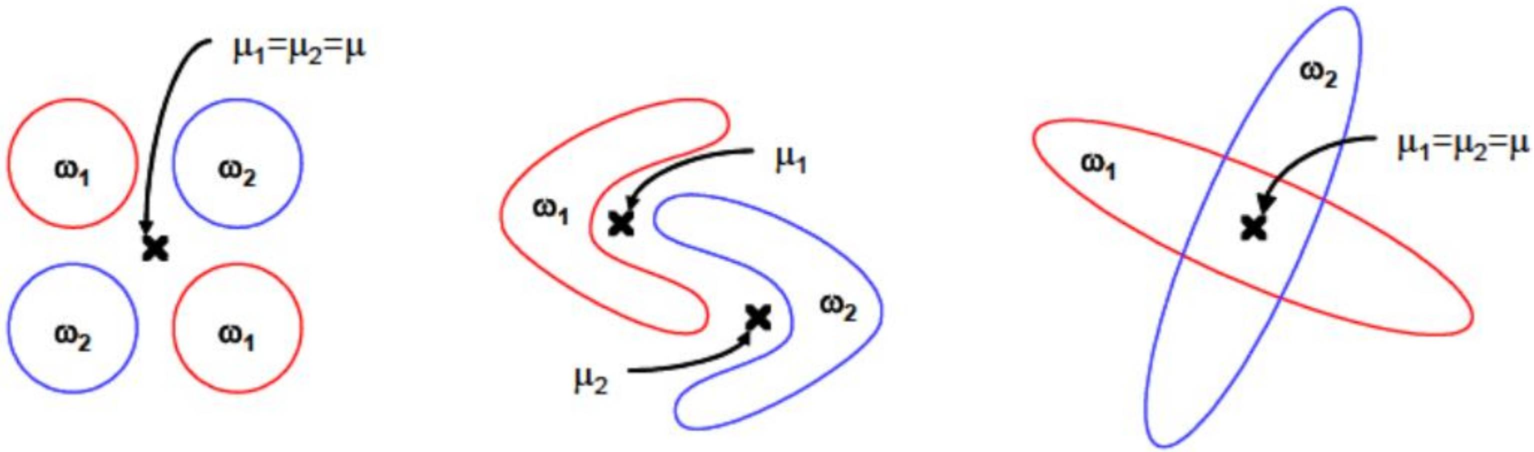


FIGURE 5.6

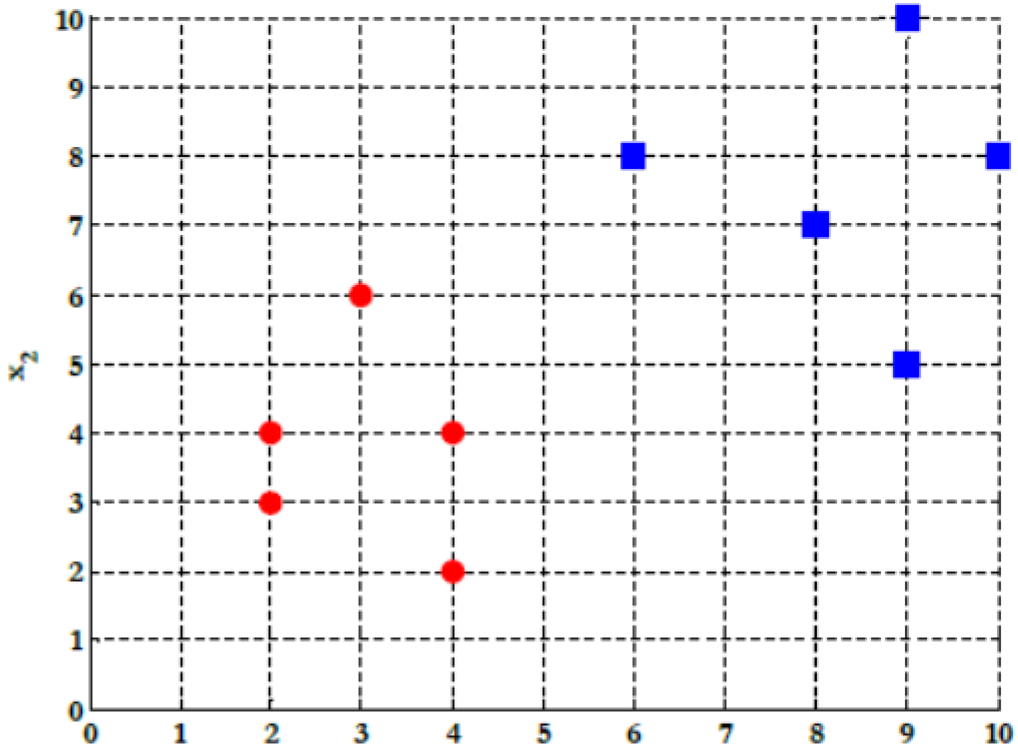
(a) The optimal line resulting from Fisher's criterion, for two Gaussian classes. Both classes share the same diagonal covariance matrix, with equal elements on the diagonal. The line is parallel to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. (b) The covariance matrix for both classes is nondiagonal. The optimal line is on the left. Observe that it is no more parallel to $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$. The line on the right is not optimal and the classes, after the projection, overlap.

مواردی که LDA جداکننده خوبی نیست



LDA
مثال – داده‌ها

- Samples for class ω_1 : $X_1=(x_1,x_2)=\{(4,2),(2,4),(2,3),(3,6),(4,4)\}$
- Sample for class ω_2 : $X_2=(x_1,x_2)=\{(9,10),(6,8),(9,5),(8,7),(10,8)\}$




```
% samples for class 1
X1 = [4,2;
      2,4;
      2,3;
      3,6;
      4,4];

% samples for class 2
X2 = [9,10;
      6,8;
      9,5;
      8,7;
      10,8];
```

گام اول: محاسبه میانگین

The classes mean are :

$$\mu_1 = \frac{1}{N_1}\sum_{x\in\omega_1} x = \frac{1}{5}\left[\begin{pmatrix}4\\2\end{pmatrix}+\begin{pmatrix}2\\4\end{pmatrix}+\begin{pmatrix}2\\3\end{pmatrix}+\begin{pmatrix}3\\6\end{pmatrix}+\begin{pmatrix}4\\4\end{pmatrix}\right]=\begin{pmatrix}3\\3.8\end{pmatrix}$$

$$\mu_2 = \frac{1}{N_2}\sum_{x\in\omega_2} x = \frac{1}{5}\left[\begin{pmatrix}9\\10\end{pmatrix}+\begin{pmatrix}6\\8\end{pmatrix}+\begin{pmatrix}9\\5\end{pmatrix}+\begin{pmatrix}8\\7\end{pmatrix}+\begin{pmatrix}10\\8\end{pmatrix}\right]=\begin{pmatrix}8.4\\7.6\end{pmatrix}$$

```
% class means
Mu1 = mean(X1)';
Mu2 = mean(X2)';
```

گام دوم: محاسبه کواریانس دو کلاس

- Covariance matrix of the first class:

$$S_1 = \sum_{x\in\omega_1}(x-\mu_1)(x-\mu_1)^T = \left[\begin{pmatrix}4\\2\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2 + \left[\begin{pmatrix}2\\4\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2$$

$$+\left[\begin{pmatrix}2\\3\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2 + \left[\begin{pmatrix}3\\6\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2 + \left[\begin{pmatrix}4\\4\end{pmatrix}-\begin{pmatrix}3\\3.8\end{pmatrix}\right]^2$$

$$=\begin{pmatrix}1 & -0.25\\ -0.25 & 2.2\end{pmatrix}$$

```
% covariance matrix of the first class
S1 = cov(X1);
```

گام دوم: محاسبه کواریانس دو کلاس ...

- Covariance matrix of the second class:

$$S_2 = \sum_{x\in\omega_2}(x-\mu_2)(x-\mu_2)^T = \left[\begin{pmatrix}9\\10\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2 + \left[\begin{pmatrix}6\\8\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2$$

$$+\left[\begin{pmatrix}9\\5\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2 + \left[\begin{pmatrix}8\\7\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2 + \left[\begin{pmatrix}10\\8\end{pmatrix}-\begin{pmatrix}8.4\\7.6\end{pmatrix}\right]^2$$

$$= \begin{pmatrix}2.3 & -0.05\\ -0.05 & 3.3\end{pmatrix}$$

```
% covariance matrix of the first class
S2 = cov(X2);
```

گام سوم: محاسبه ماتریس پراکندگی بین کلاسی

- Within-class scatter matrix:

$$S_w = S_1 + S_2 = \begin{pmatrix} 1 & -0.25 \\ -0.25 & 2.2 \end{pmatrix} + \begin{pmatrix} 2.3 & -0.05 \\ -0.05 & 3.3 \end{pmatrix}$$

$$= \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}$$

```
% within-class scatter matrix
Sw = S1 + S2 ;
```

گام چهارم: محاسبه V

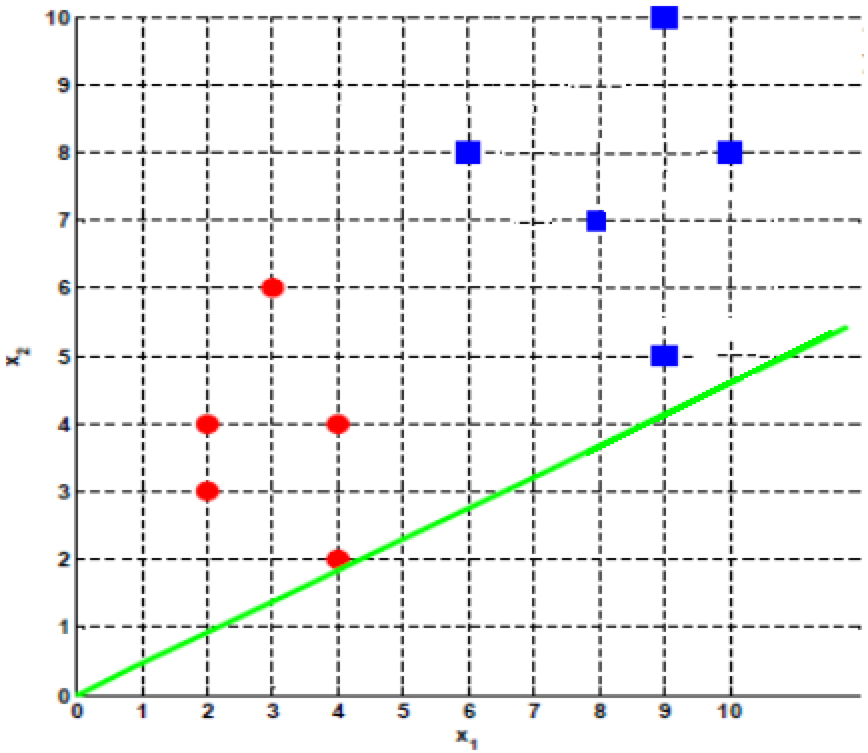
$$v = S_W^{-1}(\mu_1 - \mu_2) = \begin{pmatrix} 3.3 & -0.3 \\ -0.3 & 5.5 \end{pmatrix}^{-1} \left[\begin{pmatrix} 3 \\ 3.8 \end{pmatrix} - \begin{pmatrix} 8.4 \\ 7.6 \end{pmatrix} \right]$$

$$= \begin{pmatrix} 0.3045 & 0.0166 \\ 0.0166 & 0.1827 \end{pmatrix} \begin{pmatrix} -5.4 \\ -3.8 \end{pmatrix}$$

$$= \begin{pmatrix} 0.9088 \\ 0.4173 \end{pmatrix}$$

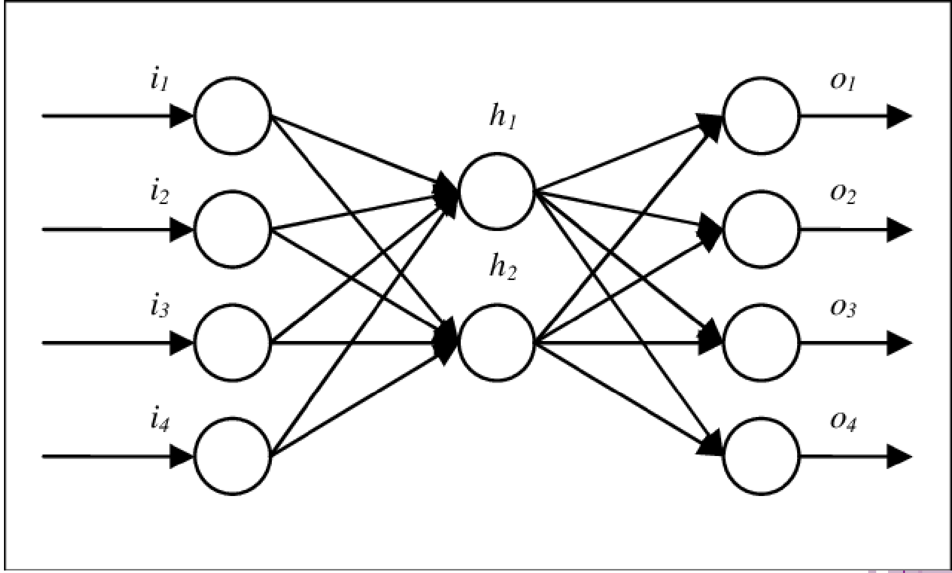
- نگاشت نقطه‌ها بر روی خط با رابطه $y = v^T x$

خط که داده‌ها بر آن تصویر می‌شوند



Using this vector leads to **good separability** between the two classes

NEURAL NETWORK AND FEATURE SELECTION



Recently, efforts have been made to use neural networks for feature generation and selection. A possible solution is via the so-called *auto-associative networks*. A network is employed having m input and m output nodes and a single hidden layer with l nodes with linear activations. During training, the desired outputs are the same as the inputs. That is,

$$\mathcal{E}(i) = \sum_{k=1}^{m} (\hat{y}_k(i) - x_k(i))^2$$

where the notation of the previous chapter has been adopted. Such a network has a unique minimum, and the outputs of the hidden layer constitute the projection of the input m-dimensional space onto an l-dimensional subspace.