بازشناسی آماری الگو

جلسه ششم

بازنمایی نمونه ها

استخراج و انتخاب مشخصه ها

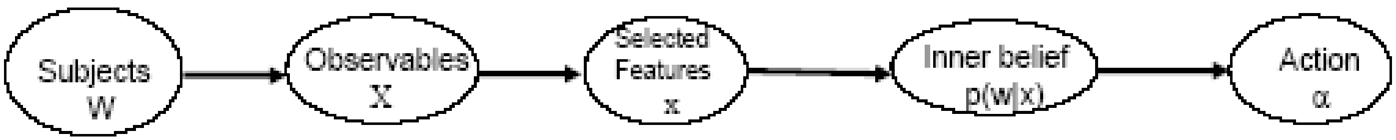
فهرست مطالب

- مقدمه و جایگاه بحث
- استخراج مشخصه ها
- رویکرد های اصلی با مربی و بدون مربی
- تقسیم بندی روش ها
  - تبدیلات (فضای فرکانسی)
  - روش های آماری
  - روش های مبتنی بر خصوصیات مسئله
- تحلیل مولفه های اساسی و جدایی پذیری خطی

مقدمه

- قبلا گفته شد که مراحل مختلف یک مسئله مهندسی بازشناسی (آماری) الگو شامل مراحل زیر است:
  - بررسی مورد یا موارد (نمونه ها) از دیدگاه جدا سازی
  - بررسی کمیت های قابل اندازه گیری نمونه ها (و با توجه به جدا سازی)
  - بررسی روشی که نمونه ها را باید با آن نشان داد (بردار مشخصه ها)
  - طراحی کلاسیفایر
    - عملیات دستکاری داده ها برای سادگس تفسیر
    - اخذ تصمیم و اقدام
  - تست کلاسیفایر

صورت بندی ریاضی مسئله در بازشناسی آماری



X--- all the observables using existing sensors and instruments
x --- is a set of features selected from components of X or linear/non-linear functions of X.
p(w|x) --- is our belief/perception about the subject class with uncertainty represented by probability.
α --- is the action that we take for x.

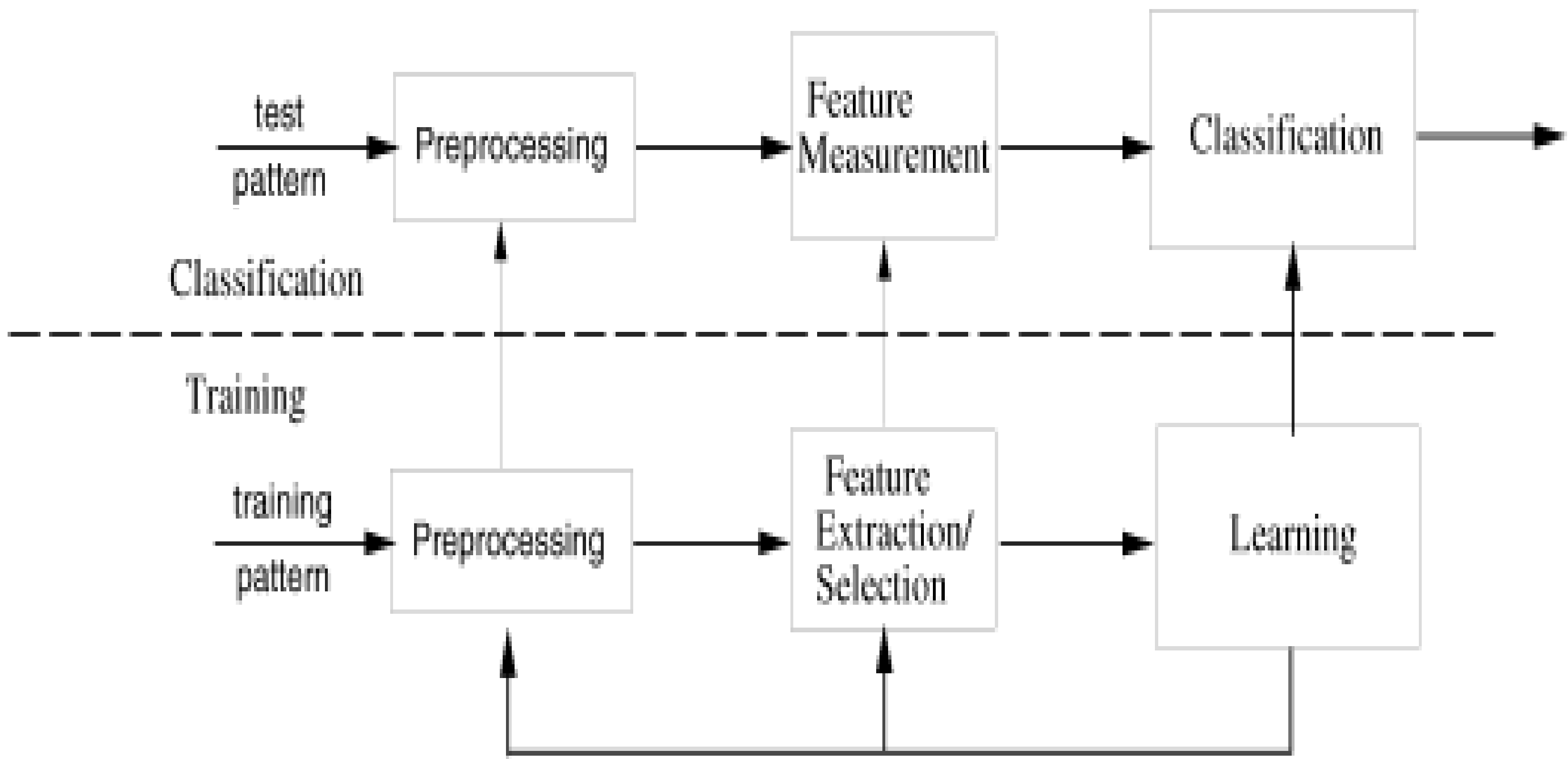
We denote the three spaces by

$$\mathbf{x} \in \Omega^{d}, \quad w \in \Omega^{C}, \quad \alpha \in \Omega^{\alpha}$$

$\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is a vector

w is the index of classes, $\Omega^{C} = \{w_1, w_2, \ldots, w_k\}$

مدل معمول عملیاتی درSPR



بازنمایی نمونه ها
تبیین صورت مسئله

- دیدیم که در نمونه های در دسترس، (مجموعه های یادگیری و آزمایش) ابتدا باید کمیت هایی را انتخاب کنیم که قابلیت اندازه گیری آسان داشته باشند.
- قرار است از این کمیت های سنجش شده، کمیت های دیگری به دست بیایند که قابیلت تفکیک بهتری ایجاد کنند (استخراج مشخصه ها)
- در بیشتر موارد، تعداد مشخصه های قابل محاسبه، آنقدر زیاد می شود که استفاده از همه آنها ممکن نبوده و باید از بین آنها چند تایی را انتخاب نمود(انتخاب مشخصه ها)

ملاحظات اصلی در استخراج مشخصه ها

- نحوست ابعاد dimension reduction
- راحتی عملیات(تبدیل ها)
- وابستگی و استقلال اطلاعاتی مشخصه ها
- جدائی پذیری زیر فضا ها
- تفکیک بین عملیات استنتاج و تصمیم

مسئله نحوست ابعاد
Curse of dimensionality

- در برخی مسائل تعداد المان های نتیجه مرحله اندازه گیری (سنجش) خیلی زیاد است (مثلا کاربرد های پردازش تصویر)
- مشکلات ناشی ابعاد بسیار زیاد:
  - افزایش حجم حافظه مورد نیاز
  - افزایش حجم پردازش (زمان و ...)
  - مشکلات عملی پردازش (تکین شدن بسیاری از ماتریس ها...)
  - افزایش احتمال گیر کردن در تله های محلی در جستجو های بهینه سازی
  - دشوار سازی پیاده سازی الگوریتم تفکیک

راحتی عملیات(تبدیل ها)

- در بسیاری ازموارد نیازمند پردازش های پیچیده ای همچون کانولوشن هستیم.
- با تغییر فضا (استفاده از تبدیل ها) داده ها را می توان با عملیات ساده تری پردازش نمود.
- معمولترین تبدیل های مرسوم:
  - تبدیل فوریه
  - تبدیل فوریه پنجره دار
  - تبدیل موجک

وابستگی اطلاعاتی(آماری) مشخصه ها

- اطلاعات سنجش شده و یا مشخصه های انتخاب شده برای بازنمایی نمونه ها، در بسیاری از موارد دارای همپوشانی و وابستگی correlation هستند
- وجود وابستگی:
  - افزایش بیهوده حجم اطلاعات
  - ایجاد مشکلات محاسباتی در حل مسائل
- اگر بتوان داده ها را در فضایی از مشخصه ها که بردار های یکه متعامد دارد نگاشت داد کار راحت تر می شود.

جدائی پذیری زیر فضا ها

- در دیدگاه آماری هدف غایی طبقه بندی، پیدا کردن زیر فضا های متناظر با کلاس هاست.
- با توجه به خصوصیات مسئله، ممکن است بتوان فضایی از مشخصه ها پیدا کرد که در آن نواحی کلاس ها از هم بهتر قابل تفکیک باشد
  - بخصوص با استفاده از خطوط
  - استفاده از توابع خوش ساختار

تفکیک بین عملیات استنتاج (تفسیر) وانتساب به کلاس

- در بسیاری موارد ممکن است بتوان صورت مسئله طبقه بندی را به شکلی ترتیب داد که عملیات ریاضی مورد نیاز برای استنتاج (تفسیر) موثرتر باشد.
- سپس با روش های ریاضی- منطقی و با تر کیب نتایج به مسئله ساده تری رسید.
- در این دیدگاه، عملیات ریاضی مربوط به تفسیر (طراحی کلاسیفایر) عملا به حوزه استخراج مشخصه ها منتقل می شود.
- در گام بعدی، با استفاده از کلاسیفایر های بسیار ساده تر، مسئله حل می گردد.
- این امر معادل است با انتقال داده ها به فضایی دیگر

ارزش اطلاعات در کاربری های مختلف

- هنگام استفاده از اطلاعات برای کاربری های مختلف (مثلا فشرده سازی یا تفکیک)، یک سوال اصلی آن است که آیا می توان مراحل مختلف بازشناسی، مانند استخراج مشخصه ها و کلاسه بندی نمونه ها را جدا از هم در نظر گرفت یا خیر.
- درکاربری های مختلف و یا حتی در مسائل مختلف یک کاربری، ارزش انواع اطلاعات ممکن است بشدت دستخوش تغییر شود.
- به عبارت دیگر حتی برای کاربری تفکیک، ممکن است یک نمونه بخصوص، در مسائل مختلف، با اطلاعات مختلفی سازگاری بهتری نشان دهد.
- این دیدگاه به این ایده منجر می شودکه در صورت امکان بهتر است اطلاعات کلاس ها نیز در مرحله استخراج مشخصه ها مورد استفاده قرار گیرد.
- این مسئله به دو رویکرد با مربی و یا بدون مربی در استخراج مشخصه ها منجر می شود

تفاوت استخراج مشخصه ها و انتخاب مشخصه ها

- در استخراج مشخصه ها
  - معمولا یک نگاشت از فضای سنجش ها (یا مشخصه های قبلی) به فضای جدیدی مطرح است
- در انتخاب مشخصه ها:
  - از برخی از مشخصه هائیکه تاثیر کمتری در بازشناسی دارند ، یا با دقت کمتری تهیه شده اند صرفنظر می شود.
  - این امر با استفاده از آزمایش مشخصه های بدست آمده در تفکیک برخی نمونه های در دسترس (بهینه سازی) صورت می گیرد.

سه رویکرد اساسی در بازنمایی نمونه ها

- رویکرد تحلیلی
  - عموما مبتنی بر مسئله و دانش ما از آن
  - ریخت شناسی موضوع morphology
- رویکرد آماری
  - تبدیلات مختلف (فوریه، pca و ...)
  - تحلیل جداسازی خطی
- تفکیک مسئله
  - ECOC

روش های معمول برای استخراج مشخصه ها

Feature Extraction and Projection Methods

Method	Property	Comments
Principal Component Analysis (PCA)	Linear map; fast; eigenvector-based.	Traditional, eigenvector based method, also known as Karhunen-Loève expansion; good for Gaussian data.
Linear Discriminant Analysis	Supervised linear map; fast; eigenvector-based.	Better than PCA for classification; limited to $(c-1)$ components with non-zero eigenvalues.
Projection Pursuit	Linear map; iterative; non-Gaussian.	Mainly used for interactive exploratory data-analysis.
Independent Component Analysis (ICA)	Linear map, iterative, non-Gaussian.	Blind source separation, used for de-mixing non-Gaussian distributed sources (features).
Kernel PCA	Nonlinear map; eigenvector-based.	PCA-based method, using a kernel to replace inner products of pattern vectors.
PCA Network	Linear map; iterative.	Auto-associative neural network with linear transfer functions and just one hidden layer.
Nonlinear PCA	Linear map; non-Gaussian criterion; usually iterative	Neural network approach, possibly used for ICA.
Nonlinear auto-associative network	Nonlinear map; non-Gaussian criterion; iterative.	Bottleneck network with several hidden layers; the nonlinear map is optimized by a nonlinear reconstruction; input is used as target.
Multidimensional scaling (MDS), and Sammon's projection	Nonlinear map; iterative.	Often poor generalization; sample size limited; noise sensitive; mainly used for 2-dimensional visualization.
Self-Organizing Map (SOM)	Nonlinear; iterative.	Based on a grid of neurons in the feature space; suitable for extracting spaces of low dimensionality.

روش های معمول برای انتخاب مشخصه ها

Feature Selection Methods

Method	Property	Comments
Exhaustive Search	Evaluate all $\binom{d}{m}$ possible subsets.	Guaranteed to find the optimal subset; not feasible for even moderately large values of m and d.
Branch-and-Bound Search	Uses the well-known branch-and-bound search method; only a fraction of all possible feature subsets need to be enumerated to find the optimal subset.	Guaranteed to find the optimal subset provided the criterion function satisfies the monotonicity property; the worst-case complexity of this algorithm is exponential.
Best Individual Features	Evaluate all the m features individually; select the best m individual features.	Computationally simple; not likely to lead to an optimal subset.
Sequential Forward Selection (SFS)	Select the best single feature and then add one feature at a time which in combination with the selected features maximizes the criterion function.	Once a feature is retained, it cannot be discarded; computationally attractive since to select a subset of size 2, it examines only $(d-1)$ possible subsets.
Sequential Backward Selection (SBS)	Start with all the d features and successively delete one feature at a time.	Once a feature is deleted, it cannot be brought back into the optimal subset; requires more computation than sequential forward selection.
"Plus l-take away r" Selection	First enlarge the feature subset by l features using forward selection and then delete r features using backward selection.	Avoids the problem of feature subset "nesting" encountered in SFS and SBS methods; need to select values of l and $r(l > r)$.
Sequential Forward Floating Search (SFFS) and Sequential Backward Floating Search (SBFS)	A generalization of "plus-l take away-r" method; the values of l and r are determined automatically and updated dynamically.	Provides close to optimal solution at an affordable computational cost.

مثال: عملکرد چند نوع تبدیل در جدا سازی در فضای مشخصه ها

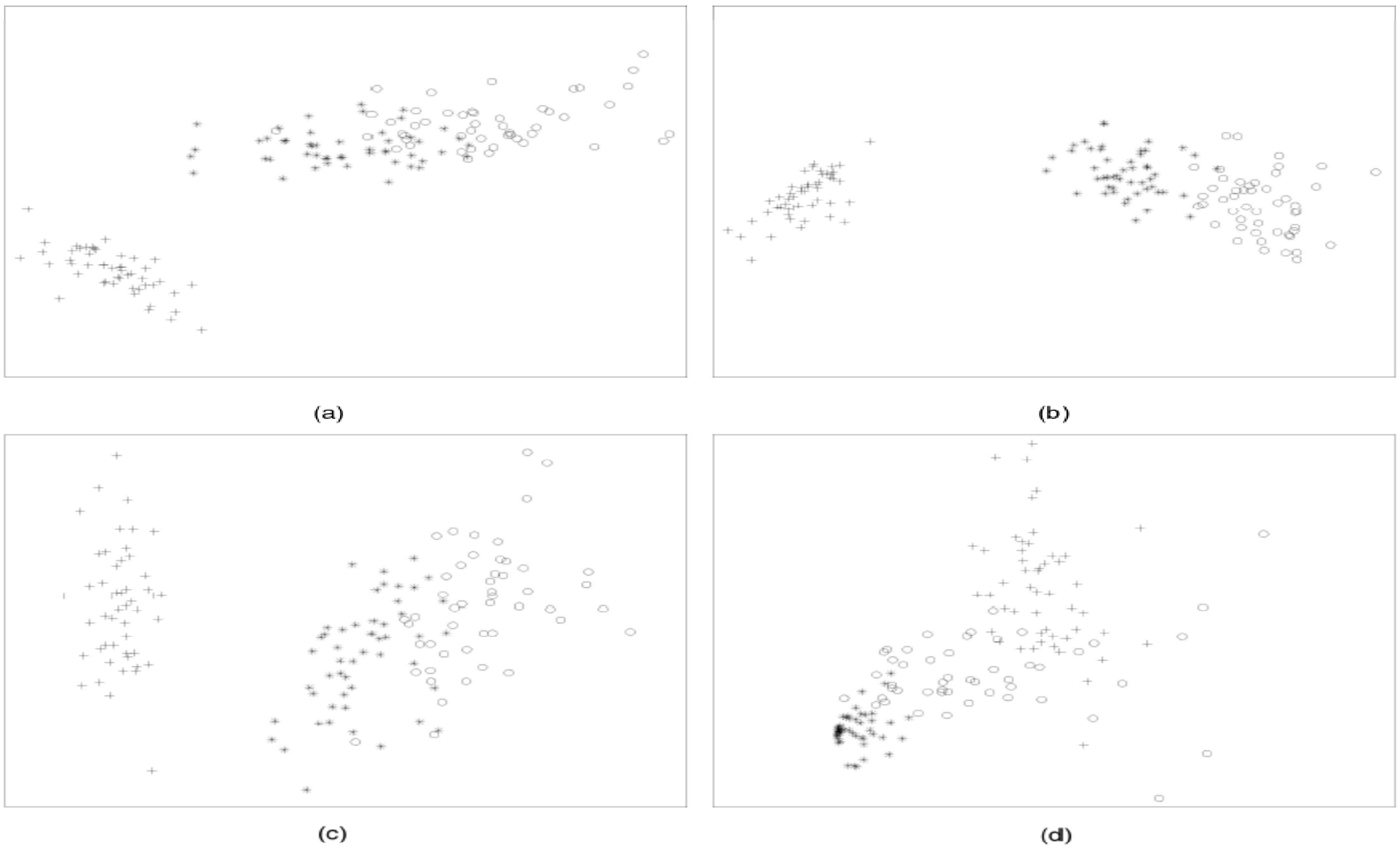
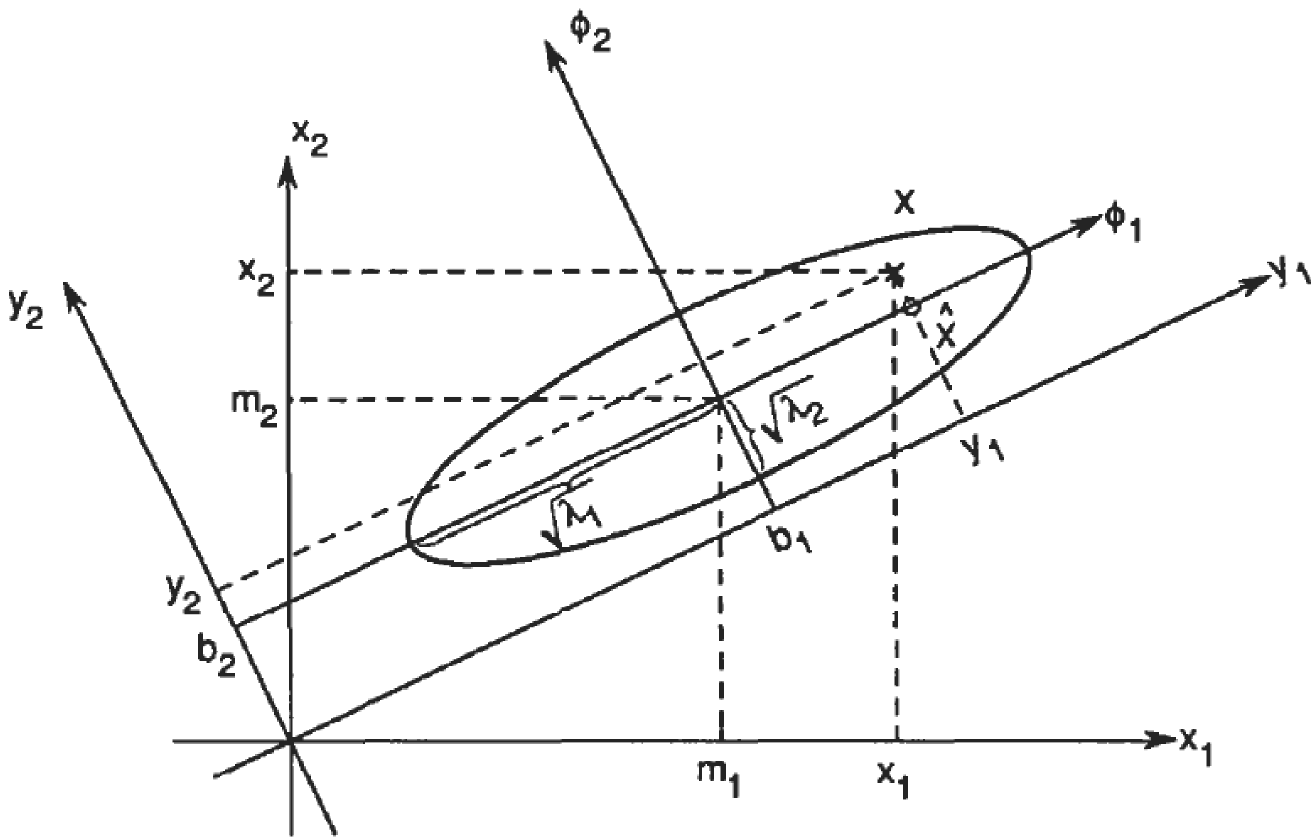


Fig. 5. Two-dimensional mappings of the Iris dataset (+: Iris Setosa; *: Iris Versicolor; o: Iris Virginica). (a) PCA, (b) Fisher Mapping, (c) Sammon Mapping, and (d) Kernel PCA with second order polynomial kernel.

تحلیل مولفه های اساسی
Principal Component Analysis(PCA)

- عناوین مختلف تبدیل در حوزه های مختلف:
  - (PCA)در بازشناسی الگو
  - Hotelling transformation در پردازش تصویر
  - Karhunen-Loeve transform در پردازش سیگنال
- آشکار ترین کاربری قابل مشاهده:
  - eigen faces – بازشناسی چهره
- ایده محوری: یافتن جهت(جهت هایی) ازبر دار های یکه متعامد سازنده محور های مختصات، که داده ها حول آن کمترین پراکندگی (و در نتیجه بیشترین تجمع اطلاعات را داشته باشند.

مثال دیداری مولفه های اساسی



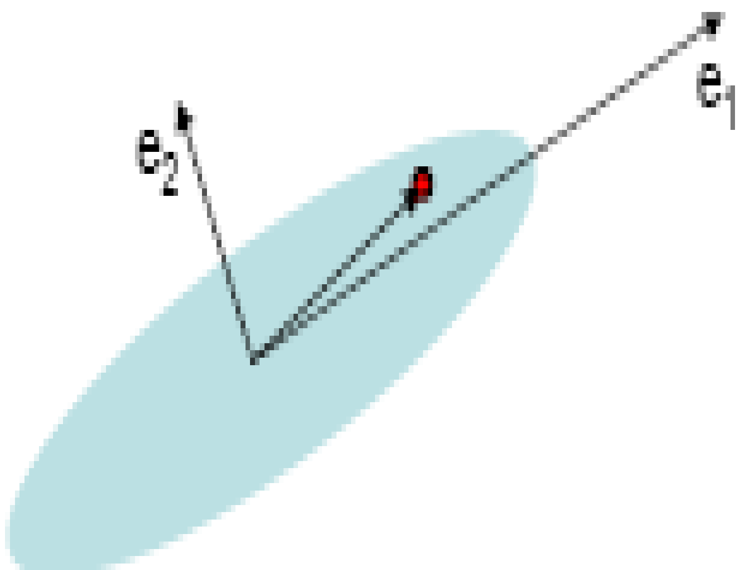
ارتباط با مفهوم فاصله Mahanalobis

$$l_i(x) = (x-\mu_i)^T \sum_i^{-1} (x-\mu_i)$$

- این کمیت رساننده این مفهوم است المانهای ماتریس کوواریانس، مانند ضریب وزنی برای مشخصه ها عمل می کنند.
- اگر ماتریس مزبور قطری باشد(با عناصر غیر صفر) نشاندهنده این است که مشخصه ها از هم مستقل بوده و دو نتیجه گیری حاصل می شود:
  - اطلاعات نمونه ها قابل نشان داده شدن با تعداد کمتری از مشخصه ها نیست(هیچ مشخصه ای را نمی توان با ترکیب خطی از مشخصه های دیگر بدست آود).
  - این امکان وجود دارد که سهم(ارزش) هر مشخصه به تنهایی در اطلاعات باز نمایی نمونه ها معین گردد
- اگر ماتریس مزبور قطری نباشد یک راه ساده تر کردن تحلیل این است که آن را به فضایی منتقل کنیم در فضای جدید به یک ماتریس قطری دست پیدا کنیم.

PCA ادامه

- در تحلیل مولفه های اساسی به دنبال یافتن دستگاه مختصاتی (از بردار های یکه) هستیم که:
  - اولا متعامد باشند، یعنی هیچیک از آنها را نتوان بر حسب ترکیبی خطی از دیگران بیان نمود
  - پراکندگی (فاصله) نمونه ها از آنها محور ها حد اقل باشد.
  - در باز سازی نمونه ها (با تعداد مشخصی از مولفه ها) کمترین خطا را به وجود آورند.



تبیین صورت مسئله به شکل ریاضی

- فرض کنیم: $\{ x1,x2, \ldots, xn \}$ نمونه هایی باشند که در فضای d بعدی داده شده باشند .
- هدف این است که : آنها را با بعد کمتری نشان داد $d' < d$
- این بردار در فضای جدید به شکل زیر قابل بازسازی است

$$m + \sum_{i=1}^{d'} a_{ki} e_i \rightarrow x_k$$

  - m: مقدار متوسط نمونه ها
  - ei: i امین بزرگترین بردار ویژه ماتریس کوواریانس
  - aki: تصویر (پروجکشن) نمونه k ام روی بردار ویژه i ام

ادامه....

- بدیهی است که می توان تعداد زیادی از دستگاه محور ها (بردار های یکه) را برای انتقال در نظر گرفت.
- در این شرایط اگر تعدا بردار های هر دستگاه محدود باشد، برای بازسازی سیگنال اولیه، ممکن است خطایی بوجود آید.
- سوال اساسی آن است که کدام دستگاه(مجموعه بردار های یکه) کمترین خطای بازسازی را ایجاد می کند؟
- معمولترین معیار خطای بازسازی مجذور مربعات است:

$$J_{d'}(m,a,e)=\sum_{k=1}^{n}\|(m+\sum_{i=1}^{d'}a_{ki}e_i)-x_k\|^2$$

چگونگی یافتن دسته بردار های متعامد یکه

- این بردار های یکه چه خصوصیتی خواهند داشت.
- اگر این دسته بردار ها را که از روی نمونه های X بدست امده اند با φ نشان دهیم و تصویر نمونه های اولیه روی آنها را Y بنامیم می توان به روابط زیر رسید. و داریم:

$$\varphi_i . \varphi_j = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}$$

- قرار است هر مولفه از بردار جدید به شکل زیر محاسبه شود:

$$y_i = \varphi_i^T . X$$

- همچنین (در بازسازی کامل با n مولفه):

$$\mathbf{X} = \sum_{i=1}^{n} \mathbf{y}_i \phi_i = \mathbf{\Phi Y} \,,$$

یافتن مولفه های اساسی:

- حال اگر عملا از تعداد کمتری از مولفه های اساسی در بازسازی استفاده شود تخمینی از سیگنال بدست می آید:

$$\hat{\mathbf{X}}(m) = \sum_{i=1}^{m} \mathbf{y}_i \phi_i + \sum_{i=m+1}^{n} b_i \phi_i \ .$$

- میزان خطا:

$$\Delta\mathbf{X}(m) = \mathbf{X} - \hat{\mathbf{X}}(m) = \mathbf{X} - \sum_{i=1}^{m} \mathbf{y}_i \phi_i - \sum_{i=m+1}^{n} b_i \phi_i = \sum_{i=m+1}^{n} (\mathbf{y}_i - b_i) \phi_i \ .$$

- حال اگر معیار را روی کل ابعاداعمال کنیم (و در بهترین شرایط فرض کنیم تعداد نامحدودی بعد داریم) می توانیم مقدار متوسط (امید ریاضی) معیار را محاسبه نمائیم:

یافتن مولفه های اساسی...

- پس:

$$\bar{\varepsilon}^2(m) = E\{\|\Delta \mathbf{X}(m)\|^2\}$$
$$= E\left\{ \sum_{i=m+1}^{n} \sum_{j=m+1}^{n} (\mathbf{y}_i - b_i)(\mathbf{y}_j - b_j)\phi_i^T \phi_j \right\}$$
$$= \sum_{i=m+1}^{n} E\{(\mathbf{y}_i - b_i)^2\} .$$

- درشکل کلی با مشتق گیری:

$$\frac{\partial}{\partial b_i} E\{(\mathbf{y}_i - b_i)^2\} = -2[E\{\mathbf{y}_i\} - b_i] = 0 .$$

b_i

$$b_i = E\{\mathbf{y}_i\} = \phi_i^T E\{\mathbf{X}\} .$$

یافتن مولفه های اساسی...

- و:

$$\bar{\varepsilon}^2(m) = \sum_{i=m+1}^{n} E[(\mathbf{y}_i - E\{\mathbf{y}_i\})^2]$$

$$= \sum_{i=m+1}^{n} \phi_i^T E[(\mathbf{X} - E\{\mathbf{X}\})(\mathbf{X} - E\{\mathbf{X}\})^T]\phi_i$$

$$= \sum_{i=m+1}^{n} \phi_i^T \Sigma_X \phi_i \ .$$

- ملاحظه می شود این معادله، همان معادله مربوط به محاسبه مقادیر یا بردار های ویژه ماتریس کوواریانس است

$$\Sigma_X . \varphi_i = \lambda_i . \varphi_i$$

- پس معیار عبارت خواهد بود از:

$$\bar{\varepsilon}^2(m)_{opt} = \sum_{i=m+1}^{n} \lambda_i \ .$$

محاسبه PCA

- ابتدا مقدار متوسط نمونه ها را محاسبه می کنیم(بردار M بعدی V)
- سپس پراکندگی نمونه ها(N نمونه) را از مقدار متوسط می یابیم.

$$A = [u_1^1 - v, \ldots, u_n^1 - v, \ldots, u_1^p - v, \ldots, u_n^p - v]$$

- در گام بعدی ماتریس کوواریانس را پیدا می کنیم

$$\boxed{C = AA^T}$$

- مقادیر ویژه و بردار های ویژه این ماتریس جواب مسئله هستند.
- اما گاهی (مثلا در تصویر) ابعاد این ماتریس M*M بسیار بزرگ است. و مشکل عملی بوجود می آورد.
- گاهی که تعداد نمونه ها کوچک است، می توان به جای ماتریس C از ماتریس L استفاده کرد که همان مقادیر ویژه را دارد:

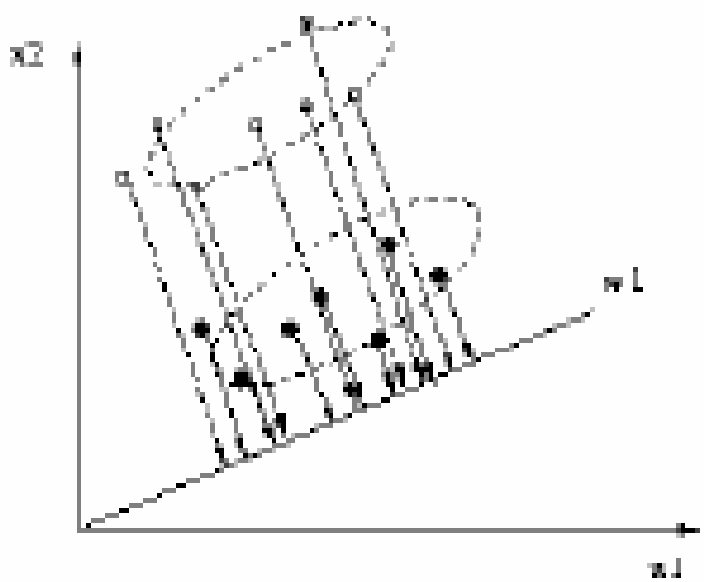
$$\boxed{L = A^T A}$$

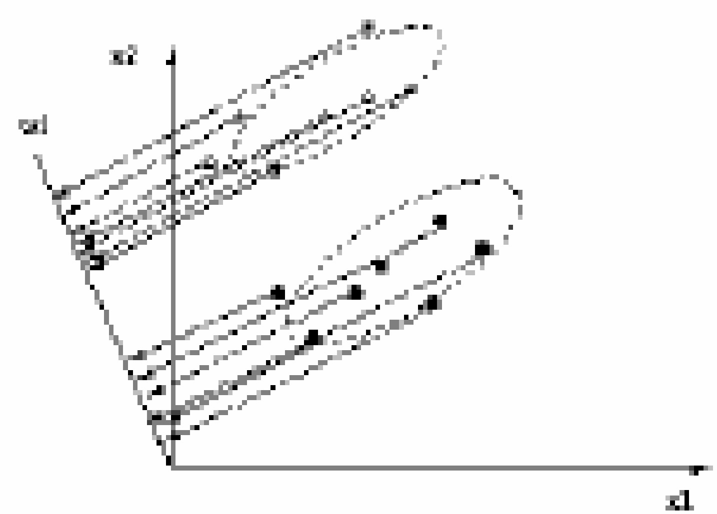
تحلیل جداساز خطی
Linear Discriminant Analysis

- اهداف متفاوتی از مرحله استخراج مشخصه ها وجود دارد که مهمترین آنها عبارت است از : کاهش ابعاد و رسیدن به شرایط راحت تر برای طبقه بندی classification
- روش های "هسته محور" kernel based بر این ایده مبتنی هستند که اطلاعات ورودی را در فضایی با شرایط بهتر نگاشت داده وبه راحتی طبقه بندی بیافزایند.
- یکی از این ایده ها این است که همه داده ها و اطلاعات آنها روی یک خط نگاشت پیدا نماید.
- بعبارت دیگر تصویر نمونه ها (که با بردار نشان داده شده اند) روی یک (یا چند) خط، به عنوان مشخصه feature پذیرفته می شود.

نمایش ایده
Linear Discriminant Analysis

to find a vector w, and project the n samples on this axis $y = w^t\mathrm{x} = \langle w, \mathrm{x} \rangle$, so that the projected samples are well separated.

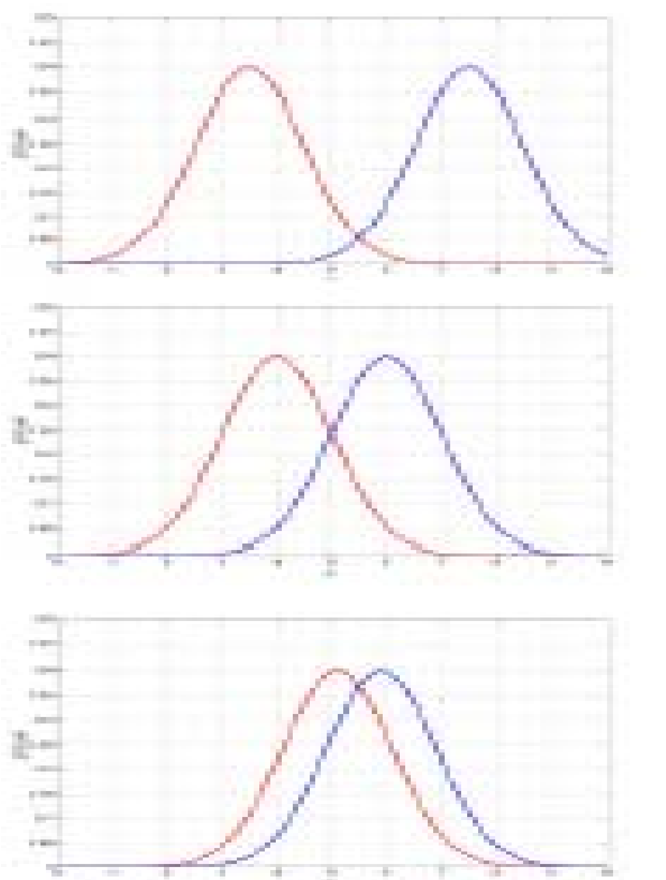




سوال اساسی:

- کدام خط می تواند بهترین شرایط را برای طبقه بندی مهیا کند؟
- در این بحث فعلا ما در مورد یک مسئله دو کلاسه تمرکز می کنیم بدون اینکه از کلیت بحث بکاهیم.
- برای شروع بحث فرض می کنیم توانسته باشیم نگاشت مزبور را انجام دهیم. در این شرایط ما با یک مسئله دو کلاسه با مشخصه یک بعدی سر و کار داریم.

راحتی و سختی طبقه بندی پس از نگاشت ورودی ها روی یک خط



- راحتی و سختی تفکیک اعضای دو کلاس در حالت یک بعدی به دو مفهوم بستگی دارد:
  - میزان فاصله دو اعضای دو کلاس از یکدیگر (هر چه بیشتر باشد طبقه بندی راحت تر است)
  - میزان نزدیکی اعضای یک کلاس به یکدیگر (هر چه بیشتر باشد طبقه بندی راحت تر است)

معیار های کمی برای مفاهیم دو گانه

- برای مفاهیم مزبور (نزدیکی درون کلاسی و فاصله بین کلاسی) چه معیار های کمی می توان پیشنهاد کرد؟
- اگر فرض کنیم که توزیع مقادیر نگاشت یافته روی خط، دارای توزیع نرمال برای دو کلاس مختلف است، در اینصورت می توانیم پارامتر های کمی توزیع های کلاس ها را برای تبیین مفاهیم مزبور بکار گیریم:
  - فاصله بین دو کلاس را می توان با فاصله مقدار متوسط mean مرتبط شمرد
  - نزدیکی اعضای یک کلاس به یکدیگر را می توان با پراکندگی (واریانس) مربوط دانست

بهترین خط:

- بنابراین ایده اصلی استخراج مشخصه ها با دارا بودن مقادیر سنجش شده اولیه این است که:
  - خطی را پیدا کنیم که پس از نگاشت هر نقطه (بردار اولیه مقادیر سنجش شده) روی آن، دو کلاس جدید روی آن خط بوجود بیاید به طوریکه :
    - فاصله میانگین های این دو کلاس حد اکثر باشد
    - میزان پراکندگی اعضای هر یک از دو کلاس حد اقل باشد.
- با توجه به این اینکه خط مزبور یگانه است (مثلا با فرض اینکه خط مزبور از مبدا می گذرد، آن را با یک بردار وزنی W می توان نشان داد)، فی الواقع به یک پروسه بهینه سازی optimization بر خورد می کنیم.

ملزومات بهینه سازی

- تعریف یک معیار
  - در بر دارنده هر دو مفهوم کم بودن فاصله درون کلاسی و زیاد بودن فاصله بین دو کلاس باشد
  - با خط مزبور (W) مرتبط باشد تا بر اساس بهینه سازی آن بهترین خط پیدا شود
- ارائه روشی برای جستجو در فضای مزبور
  - روش های فرمولی گرادیان (مانند مشتق گیری و مساوی صفر قرار دادن)
  - انواع دیگر جستجو (اتفاقی مانند جستجوی ژنتیک)

تبیین صورت مسئله

- آنچه در اختیار است دو دسته از بردار ها (X) در دو کلاس مختلف است(d بعدی).
- آنچه مجهول است خطی مانند W است که قرار است تصویر هر بردار روی آن خط به عنوان مشخصه مورد استفاده قرار گیرد.
- می دانیم که تصویر نقطه (یا بردار) X را می توان با استتفاده از مفهوم ضرب داخلی نمایش داد:
- x=W.X
- با این مفهوم می توان بر اساس الگوریتم زیر کار کرد:

الگوریتم پیشنهادی :

Definition:

The **sample mean** for class ω_i:

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{x \in \chi_i} \mathbf{x}, \qquad i = 1, 2.$$

The **scatter matrix** for class ω_i:

$$\mathbf{S}_i = \sum_{x \in \chi_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t, \qquad i = 1, 2.$$

The **between-class scatter matrix**

$$\mathbf{S}_B = (\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t$$

The **within-class scatter matrix**

$$\mathbf{S}_W = \mathbf{S}_1 + \mathbf{S}_2$$

اما پس از نگاشت:

The **sample mean** of the projected points in class ω_i:

$$\tilde{m}_i = \frac{1}{n_i} \sum_{\mathrm{x} \in \chi_i} w^t \mathrm{x} = w^t \mathrm{m}_i, \qquad i = 1, 2.$$

The **scatter** of the projected points in class ω_i:

$$\tilde{s}_i = \sum_{\mathrm{x} \in \chi_i} (w^t \mathrm{x} - w^t \mathrm{m}_i)^2 = w^t \mathrm{S}_i w, \qquad i = 1, 2.$$

These are 1D variables

تعریف معیار بهینه سازی

Fisher's linear discriminant $y = w^t\mathrm{x}$: choose w to maximize

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} = \frac{w^t S_B w}{w^t S_W w}.$$

i.e. the between-class distance should be as large as possible, and the within class scatter should be as small as possible.

This is a typical criterion used in almost all discriminative methods.

معیار بهینه سازی و نقطه بهینه

The vector w that maximizes the criterion function

$$J(w) = \frac{w^t S_B w}{w^t S_W w}$$

$$w = S_W^{-1}(m_1 - m_2)$$

مشکلات عملی

- مشکل اساسی: نحوست ابعاد
  - فرض کنید برای یک مسئله بازشناسی چهره تنها از ابعاد بسیار کوچک ۶۴×۶۴ استفاده کرده باشیم.
  - در این شرایط بعد تصویر در فضای اولیه عبارت خواهد بود از ۶۴×۶۴=۴۰۹۶ .
  - به تبع ابعاد ماتریش کوواریانس عبارت خواهد بود از ۴۰۹۶×۴۰۹۶=۱۶ مگا بایت
  - بدیهی است با افزایش تعداد نمونه ها با مشکل برخورد می کنیم.

قطری کردن ماتریس ها

-) قطری کردن Sb : ماتریس V را چنان می یابیم که داشته باشیم :
  - VT.Sb.V= Λ به طوریکه VTV=I.
- این امر را می توان با آنالیز مقادیر ویژه انجام داد که در آن هر ستون V یک بردار ویژه Sb است و Λ شامل همه مقادیر ویژه می باشد.
- Sb ممکن است تکین باشد ، برخی از مقادیر ویژه صفر (یا نزدیک صفر) خواهند بود. لازم است که این مقادیر و بردارهای ویژه را حذف نمود چون جهت انتقال با پراکندگی صفر هیچ قدرت جداسازی ندارند اما در عملیات ایجاد اختلال می کنند .
- اگر Y ، n ستون اول V باشد (یک ماتریس n*m) : YT.Sb.Y=Db >0
- که Db یک زیرماتریس اصلی Λ و با سایز m*m می باشد.

ادامه

- فرض کنیم Z=YDb-1/2
- ZT.Sb.Z=I ⇒ (YDb-1/2)TSb(YDb-1/2)=I
- در نتیجه Z موجب کاهش ابعاد از n به m می شود.
- حال ZT.SwZ را قطری می کنیم. UT.ZT.Sw.ZU=Dw که UT.U=I
- Dw ممکن است در قطرش حاوی صفر باشد. از آنجاییکه هدف ماکزیمم کردن نسبت پراکندگی کل به پراکندگی داخل کلاس است ، می توان المانهای قطری Dw را مرتب کرد و برخی مقادیر ویژه را در پایان به همراه بردارهای ویژه متناظر حذف نمود. نگهداشتن ابعاد با مقادیر ویژه کوچک بخصوص صفرها حائز اهمیت می باشد. واین دقیقاً دلیل این قضیه است که ابتدا Sb را به جای Sw قطری می کنیم.

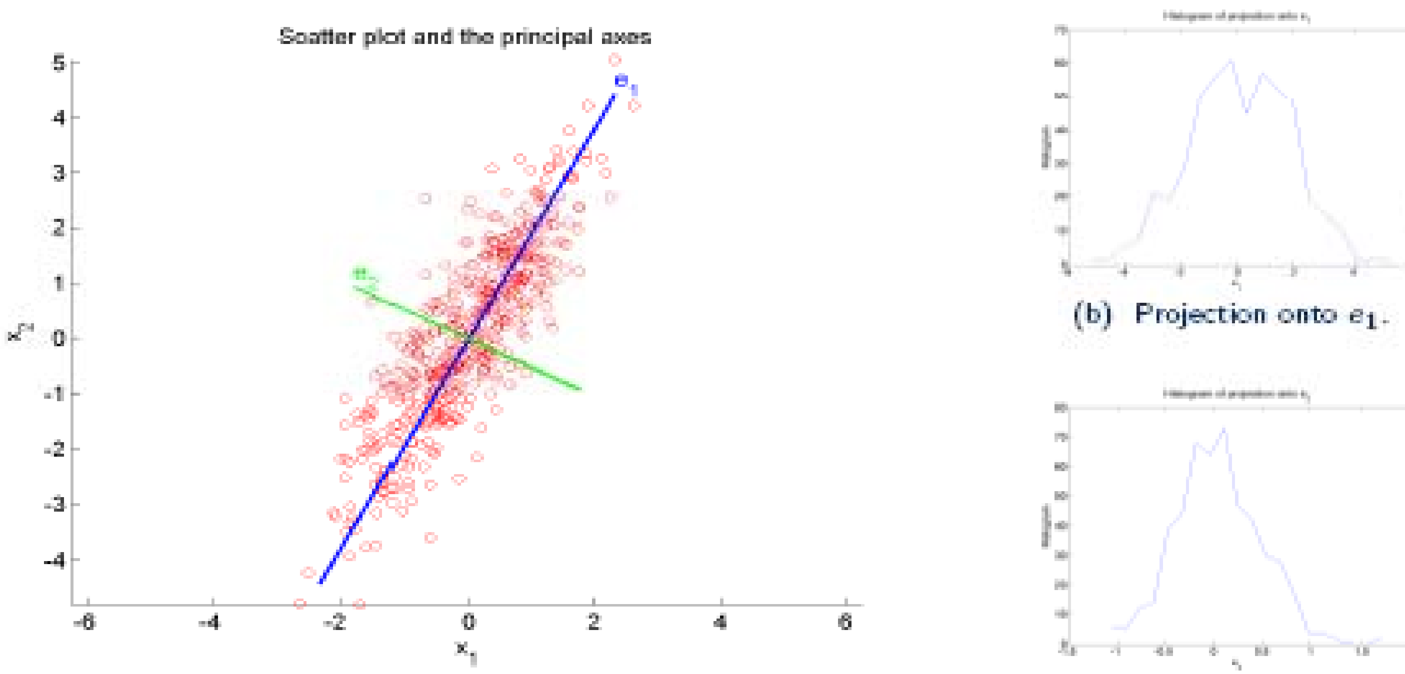
ادامه

- ماتریس LDA را به این شکل تعریف می نماییم.
A=UT.ZT
- A هم صورت و هم مخرج ضابطه فیشر را قطری می نماید.
- A.Sb.AT=I A.Sw.AT=Dw
- ۴) برای مقاصد کلاسه بندی ، باید توجه کرد که A ، Sw را قطری می نماید. در نتیجه تبدیل نهایی به این صورت خواهد بود.
- X* → Dw-1/2AX

چند نکته در خصوص LDA

- با کمی توجه آشکار می شود که نخستین تفاوت این روش با PCA در این است که این روش به قدرت تفکیک تکیه دارد در حالیکه PCA به اطلاعات موجود توجه می کند.
- روش LDA از همان ابتدا از دانش مربوط به کلاس ها (اطلاعات یادگیری با مربی) استفاده نموده است در حالیکه در روش معمول PCAاز این اطلاعات استفاده نمی شود.
- در بسیاری از موارد این روش نتایج بهتری می دهد

مثال: مولفه های اساسی ونگاشت



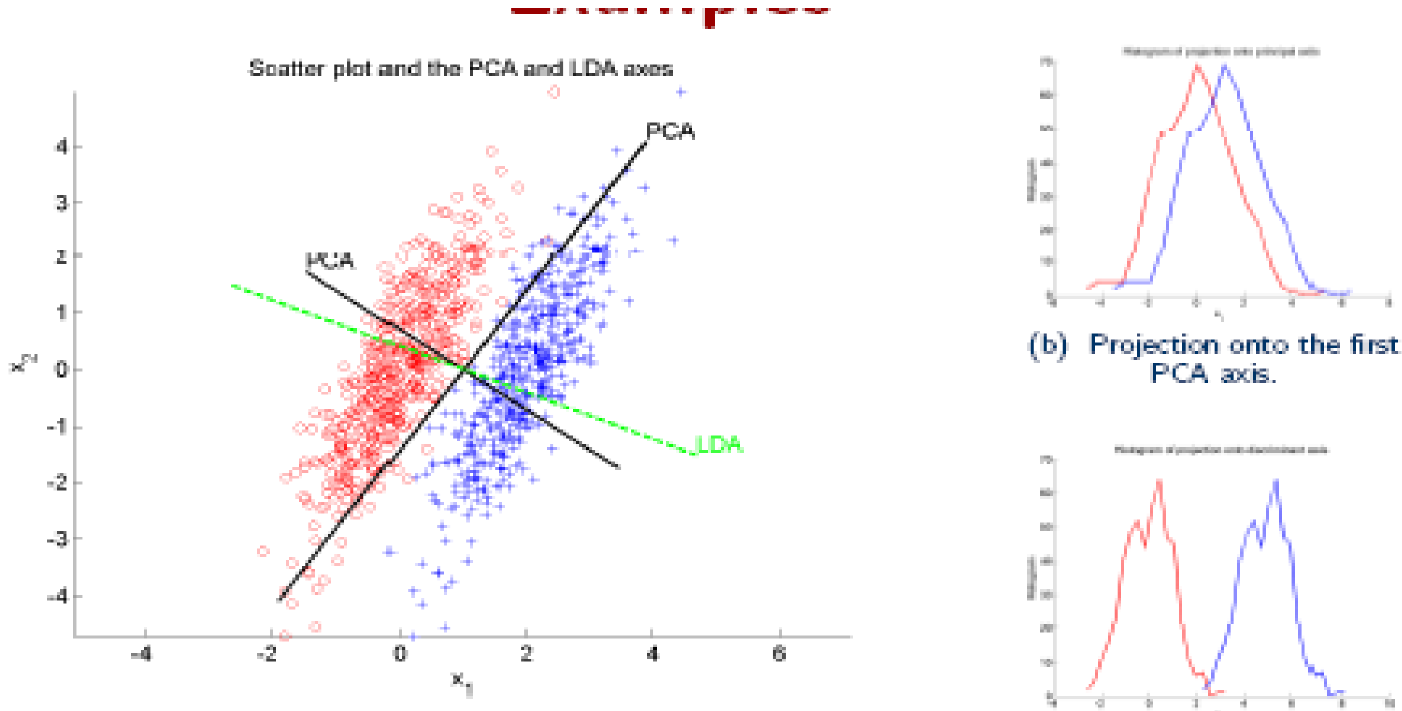
(a) Scatter plot.

(b) Projection onto e_1.

(c) Projection onto e_2.

Figure 7: Scatter plot (red dots) and the principal axes for a bivariate sample. The blue line shows the axis e_1 with the greatest variance and the green line shows the axis e_2 with the smallest variance. Features are now uncorrelated.

مثال مقایسه دو روش



(a) Scatter plot.

(b) Projection onto the first PCA axis.

(c) Projection onto the first LDA axis.

Figure 8: Scatter plot and the PCA and LDA axes for a bivariate sample with two classes. Histogram of the projection onto the first LDA axis shows better separation than the projection onto the first PCA axis.

مثال دیگر

(a) Scatter plot.

(b) Projection onto the first PCA axis.

(c) Projection onto the first LDA axis.

Figure 9: Scatter plot and the PCA and LDA axes for a bivariate sample with two classes. Histogram of the projection onto the first LDA axis shows better separation than the projection onto the first PCA axis.

خلاصه و نتیجه گیری

- LDA بر نگاشت داده ها روی یک خط مبتنی است در حالیکه PCA بر تجمع داده ها حوی یک خط
- LDA به شرایطی توجه دارد که تصویر نمونه ها از هم بهترین تفکیک پذیری را داشته باشند
  - بیشترین پراکندگی بین کلاس های مختلف
  - کمترین پراکندگی در درون هر کلاس
- در هر دو روش بطور معمول با مشکلات مربوط به عملیات ماتریسی مواجه هستیم