INF5300

Linear feature transforms

- Linear feature transforms
- Principal component analysis (PCA)
- Fisher's linear discriminant analysis

Curriculum: See links to pdfs on course page.

Linear feature transforms

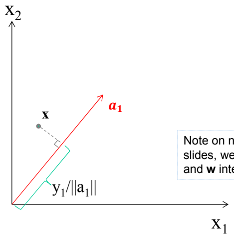
- We create new features by computing linear combinations of our existing features, $x_0, x_1, .. x_{n-1}$:

$$y_1 = \sum_{i=0}^{n-1} a_{i1}x_i, \quad y_2 = \sum_{i=0}^{n-1} a_{i2}x_i, \quad \dots \quad y_m = \sum_{i=0}^{n-1} a_{im}x_i$$

- In matrix notation $\mathbf{y} = \mathbf{A}^T\mathbf{x}$
- If $\mathbf{y}$ has fewer elements than $\mathbf{x}$, we get a feature reduction

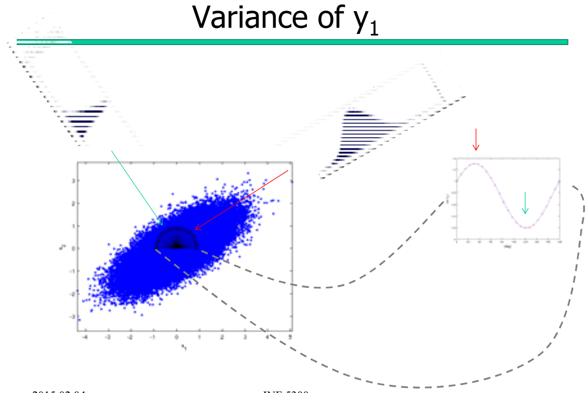
Visualizing the weights in 2D/3D

$$y_1 = \sum_{i=0}^{n-1} a_{i1}x_i = \boldsymbol{a}_1^T\boldsymbol{x}$$



Note on naming: In the slides, we often use **a** and **w** interchangeably

Variance of y_1



Variance of y_1 cont.

- Assume mean of **x** is subtracted

$$\sigma_{y_1}^2 = \frac{1}{N}\sum_i y_i^2$$
$$= \frac{1}{N}\sum_i (\mathbf{w}^T\mathbf{x}_i)^2 = \frac{1}{N}\sum_i \mathbf{w}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{w} = \mathbf{w}^T\Big(\frac{1}{N}\sum_i \mathbf{x}_i\mathbf{x}_i^T\Big)\mathbf{w}$$
$$= \mathbf{w}^T\mathbf{R}\mathbf{w}$$

The sample covariance matrix; **R**

Called σ^2_w on some slides

y_1

Variance and projection residuals

Single sample

Projection onto **w,** assuming |**w**|=1

$$\begin{aligned} \|\vec{x}_i - (\vec{w}\cdot\vec{x}_i)\vec{w}\|^2 &= (\vec{x}_i - (\vec{w}\cdot\vec{x}_i)\vec{w})\cdot(\vec{x}_i - (\vec{w}\cdot\vec{x}_i)\vec{w}) \\ &= \vec{x}_i\cdot\vec{x}_i - \vec{x}_i\cdot(\vec{w}\cdot\vec{x}_i)\vec{w} \\ &\quad -(\vec{w}\cdot\vec{x}_i)\vec{w}\cdot\vec{x}_i + (\vec{w}\cdot\vec{x}_i)\vec{w}\cdot(\vec{w}\cdot\vec{x}_i)\vec{w} \\ &= \|\vec{x}_i\|^2 - 2(\vec{w}\cdot\vec{x}_i)^2 + (\vec{w}\cdot\vec{x}_i)^2\vec{w}\cdot\vec{w} \\ &= \vec{x}_i\cdot\vec{x}_i - (\vec{w}\cdot\vec{x}_i)^2 \end{aligned}$$

«y_i»

«y_i^2»

w·**w**=1

$$MSE(\vec{w}) = \frac{1}{n}\left(\sum_{i=1}^{n}\|\vec{x}_i\|^2 - \sum_{i=1}^{n}(\vec{w}\cdot\vec{x}_i)^2\right)$$

All n samples (not dimensions)

Indie of **w**

σ^2_w

Note: Max variance ↔ min projection residuals!

Criterion function

- Goal: Find transform minimizing representation error
- We start with a single weight-vector, **w**, giving us a single feature, y_1
- Let J(**w**) = σ_w^2
- Now, let's find $\max_{\mathbf{w}} J(\mathbf{w})$ s.t. $||\mathbf{w}|| = 1$

As we learned on the previous slide, maximizing this is equivalent to minimizing representation error

Maximizing variance of y_1

$$\mathscr{L}(\mathbf{w},\lambda) \equiv \sigma_{\mathbf{w}}^2 - \lambda(\mathbf{w}^T\mathbf{w} - 1)$$
$$\frac{\partial L}{\partial \lambda} = \mathbf{w}^T\mathbf{w} - 1$$
$$\frac{\partial L}{\partial \mathbf{w}} = 2\mathrm{R}\mathbf{w} - 2\lambda\mathbf{w}$$

Lagrangian function for maximizing σ^2_w with the constraint $\mathbf{w}^T\mathbf{w}$=1

⇓ Equating zero

$$\mathbf{w}^T\mathbf{w} = 1$$
$$\mathrm{R}\mathbf{w} = \lambda\mathbf{w}$$

Unfamiliar with Lagrangian multipliers? See link to short note on the course's web page – very useful!

The maximizing **w** is an eigenvector of R!

And σ^2_w=λ! [Why?]

Eigenvectors of covariance matricesh

Real-valued, symmetric, «n-dimensional» covariance matrix

$$\mathbf{R} = \lambda_1 \mathbf{a}_1 \mathbf{a}_1' + \lambda_2 \mathbf{a}_2 \mathbf{a}_2' + \ldots + \lambda_n \mathbf{a}_n \mathbf{a}_n'$$

Eigenvalue (let's say largest)

Eigenvector corresponding to λ_1

Smallest eigenvalue

$\mathbf{a}^T_i \mathbf{a}_j = 0$ for $i \neq j$

Remember: λ_i=var of $\mathbf{x}^T \mathbf{a}_i$

Variance of multiple variables

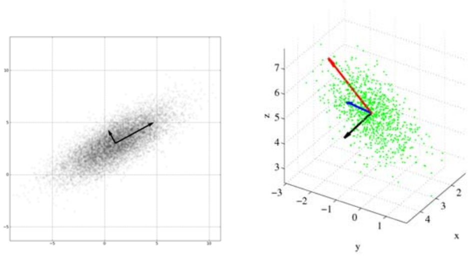
$$\sigma^2_{y1+y2} = \frac{1}{N} \sum_i (\mathbf{w}_1^T \mathbf{x}_i + \mathbf{w}_2^T \mathbf{x}_i)^2 = \ldots = \mathbf{w}_1^T \mathbf{R} \mathbf{w}_1 + \mathbf{w}_2^T \mathbf{R} \mathbf{w}_2 + 2\mathbf{w}_1^T \mathbf{R} \mathbf{w}_2$$

=0 if y_1 and y_2 are uncorrelated, e.g. if $\mathbf{w}_1$ and $\mathbf{w}_2$ are eigenvectors of **R**

That is, on the previous slide $\mathbf{a}^T_i \mathbf{R} \mathbf{a}_j = 0$ for $i \neq j$

- If the weight-vectors yield uncorrelated features, their combined variance is the sum of each one's
- If $\mathbf{w}_1$ is the principle eigenvector of **R**, which $\mathbf{w}_2$ giving an uncorrelated feature would you choose to maximize σ^2_{y1+y2}?
- Say $\mathbf{w}_1$ and $\mathbf{w}_2$ are the two principle eigenvectors of **R** on the previous slide; what ratio of the total variance would they have?

Example of distributions and eigenvectors



(Illustration courtesy of «the Internet»)

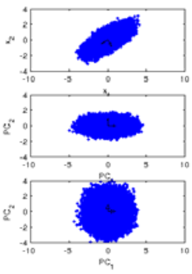
Principal component transform (PCA)

- Place the *m* «principle» eigenvectors (the ones with the largest eigenvalues) along the columns of A
- Then the transform $\mathbf{y} = \mathbf{A}^T \mathbf{x}$ gives you the *m* first principle components
- The *m*-dimensional **y**
  - have uncorrelated elements
  - retains as much variance as possible
  - gives the best (in the mean-square sense) description of the original data (through the «image»/projection/reconstruction **Ay**)

Note: The eigenvectors themselves can often give interesting information

PCA is also known as Karhunen-Loeve transform

PCA and rotation and «whitening»



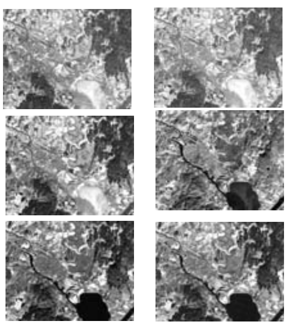
If we use all eigenvectors in the transform, $\mathbf{y} = \mathbf{A}^t\mathbf{x}$, we simply rotate our data so that our new features are uncorrelated, i.e., cov($\mathbf{y}$) is a diagonal matrix.

If we as a next step scale each feature by their σ, $\mathbf{y} = \mathbf{D}^{(-1/2)}\mathbf{A}^t\mathbf{x}$, where $\mathbf{D}$ is a diagonal matrix of eigenvalues (i.e., variances), we get cov($\mathbf{y}$)=$\mathbf{I}$. We say that we have «whitened» the data.

PCA and multiband images

- We can compute the principal component transform for an image with *n* bands
- Let $\mathbf{X}$ be an Nxn matrix having a row for each image sample
- Sample covariance matrix $\boldsymbol{R} = \frac{1}{N}\boldsymbol{X}^T\boldsymbol{X}$
- Place the (sorted) eigenvectors along the columns of $\mathbf{A}$
- $\mathbf{Y}$=$\mathbf{XA}$ will then contain the image samples, however most of the variance is in the «bands» with the lowest index (corresponding to the largest eigenvalues), and the new features are uncorrelated

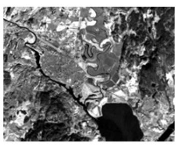
PCA example – original image

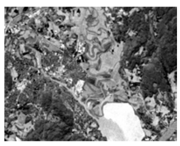


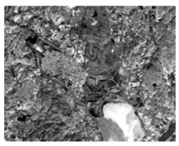
- Satellite image from Kjeller
- 6 spectral bands with different wavelengths

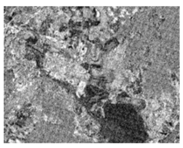
1	Blue	0.45-0.52	Max. penetration of water
2	Green	0.52-0.60	Vegetation and chlorophyll
3	Red	0.63-0.69	Vegetation type
4	Near-IR	0.76-0.90	Biomass
5	Mid-IR	1.55-1.75	Moisture/water content in vegetation/soil
7	Mid-IR	2.08-2.35	Minerals

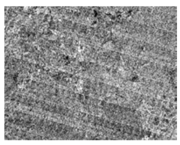
Example cont: Principal component images

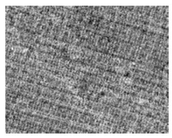

Principal component 1


Principal component 2


Principal component 3


Principal component 4

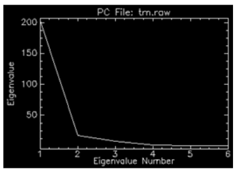

Principal component 5


Principal component 6

Example cont: Inspecting the eigenvalues

The mean-square representation error we get with m of the N PCA-components is given as

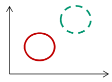
$$E\left[\|x-\hat{x}\|^2\right] = \sum_{i=1}^{N-1}\lambda_i - \sum_{i=1}^{m}\lambda_i = \sum_{i=m}^{N-1}\lambda_i$$



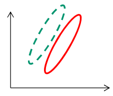
Plotting λ_i will give indications on how many features are needed for representation

PCA and classification

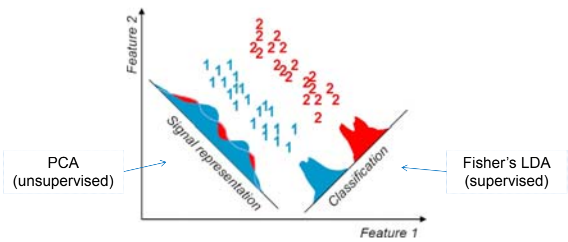
- Reduce overfitting by detecting directions/components without any/very little variance
- Sometimes high variation means useful features for classification:



- .. and sometimes not:



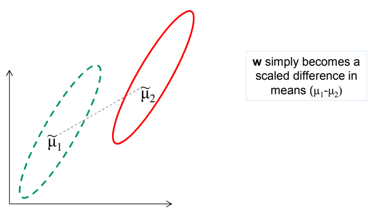
Intro to Fisher's linear discriminant



Criterion function - a first attempt

- To find a good projection vector, we need to define a measure of separation between the projections . This will be the criterion function *J*(***w***)
- A naive choice would be projected mean difference, $J(w) = |\tilde{\mu}_1 - \tilde{\mu}_2|^2$, s.t. |**w**|=1.

This criterion does not consider variance

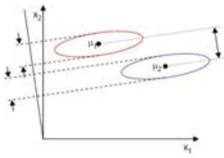


A criterion function including variance

- Fisher's solution: Maximize a function that represents the difference between the means, scaled by a measure of the within-class scatter
- Define classwise scatter (scaled variance)
  $\tilde{s}_i^2 = \sum_{y \in \omega_i} (y - \tilde{\mu}_i)^2$
- $\tilde{s}_1^2 + \tilde{s}_2^2$ is *within class scatter*
- Fisher's criterion is then

$$J(\mathbf{w}) = \frac{|\tilde{\mu}_1 - \tilde{\mu}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}$$

- We look for a projection where examples from the same class are close to each other, while at the same time projected mean values are as far apart as possible



Scatter matrices – M classes

- Within-class scatter matrix:

$$S_w = \sum_{i=1}^{M} P(\omega_i) S_i$$

$$S_i = E[(x - \mu_i)(x - \mu_i)_T]$$

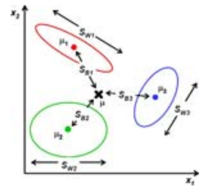
Weighted average of each class' sample covariance matrix

- Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P(\omega_i)(\mu_i - \mu)(\mu_i - \mu)^T$$

$$\mu = \sum_{i=1}^{M} \mu_i$$

Sample covariance matrix for the means

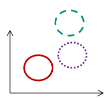


Fisher criterion in terms of within-class and between-class scatter matrices:

$$J(\mathbf{w}) = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$$

Multiple classes, $\mathbf{S}_w = \sigma^2 \mathbf{I}$

- If $\mathbf{S}_w = \sigma^2 \mathbf{I}$, we can fix $||\mathbf{w}||=1$ and make the denominator in J(**w**) independent of **w** → J(**w**) guided by the spread of the means ($\mathbf{S}_b$) only:



$$J(\mathbf{w}) = \mathbf{w}^T \mathbf{S}_b \mathbf{w}$$

We should know how to maximize this s.t. |**w**|=1 by now!

- Weight-vector giving maximum separability is given by principal eigenvector of $\mathbf{S}_b$
  - Second best (and orthogonal to first) by next-to-principal
  - … etc. for higher dimensional settings
  - … until a maximum of M-1 dimensions (number of classes minus one) [If classes are «isotropically» Gaussian distributed, all discriminatory information is in this subspace!]

General $\mathbf{S}_w$ I/II

- We saw that $\mathbf{S}_w = \mathbf{I}$ gave Fisher criterion independent of $\mathbf{S}_w$, and only dependent on $\mathbf{S}_b$
- We can get there by «whitening» the data before applying the Fisher criterion
  - Whitening data by rotation and scaling -> No general loss as distribution overlap does not change
- We must find $\mathbf{y} = \mathbf{A}^T\mathbf{x}$ that yields $\mathbf{S}_{w\mathbf{y}} = \mathbf{I}$
  - We have seen that PCA gives uncorrelated data, per-feature scaling can give unit variance per feature:
  - $\mathbf{y} = \mathbf{D}^{-1/2}\mathbf{A}^T\mathbf{x}$, where $\mathbf{A}$ has eigenvectors of $\mathbf{S}_w$ as columns, and $\mathbf{D}$ is a diagonal matrix with corresponding eigenvalues

$$\mathbf{S}_{w_y} = \frac{1}{N}\sum_i (\mathbf{D}^{-1/2}\mathbf{A}^T\mathbf{x}_i)(\mathbf{D}^{-1/2}\mathbf{A}^T\mathbf{x}_i)^T = \mathbf{D}^{-1/2}\mathbf{A}^T\mathbf{S}_w\mathbf{A}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}\mathbf{D}\mathbf{D}^{-1/2} = \mathbf{I}$$

General $\mathbf{S}_w$ II/II

- Let $\mathbf{B} = \mathbf{D}^{-1/2}\mathbf{A}^T$ (the whitening transform)
- $\mathbf{S}_b$ becomes after whitening step:

$$\mathbf{S}_{by} = \mathbf{B}\mathbf{S}_b\mathbf{B}^T$$

- Ignoring the denominator (which is now independent of **w**), we get
  - $J_y(\mathbf{w}) = \mathbf{w}^T\mathbf{S}_{by}\mathbf{w} = \mathbf{w}^T\mathbf{B}\mathbf{S}_b\mathbf{B}^T\mathbf{w}$, s.t. $|\mathbf{w}|=1$
- The weight-vectors, $\mathbf{w}^*$, maximizing separation are now given by the principal eigenvectors of $\mathbf{B}\mathbf{S}_b\mathbf{B}^T$ (in the whitened space)

Set $J_y(\mathbf{w}^*)=J(\mathbf{w})$ to see this

- In the original space, $\mathbf{w} = \mathbf{B}^T\mathbf{w}^* = \mathbf{A}\mathbf{D}^{-1/2}\mathbf{w}^*$

Solving Fisher more directly

- Alternatively, you can notice that

$$J(\mathbf{w}) = \frac{\mathbf{w}^T\mathbf{S}_b\mathbf{w}}{\mathbf{w}^T\mathbf{S}_w\mathbf{w}}$$

- .. is a «generalized Rayleigh quotient» and look up the solution for its maximum, which is the principal eigenvector of

$$\mathbf{S}_w^{-1}\mathbf{S}_b$$

- The following solutions (orthogonal in $\mathbf{S}_w$, i.e., $\mathbf{w}_i^T\mathbf{S}_w\mathbf{w}_j=0$, for $i\neq j$) are the next principal eigenvectors

Note that the obtained **w**s are identical (up to scaling) to those from the two-step procedure from the previous slides

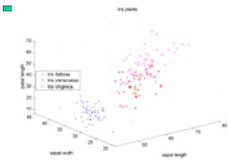
Comments on Fisher's discriminant

- In general, projection of the original feature vector to a lower dimensional space is associated with some loss of information
  - Keeping all M-1 dimensions gives you no reduction in classification performance for a Gaussian classifier with equal class-covariance matrices (LDA)
- Although the projection is optimal with respect to the given criterion (J(**w**)), the criterion itself might not be suitable for a given data set / classifier
- Minimizing J(**w**) is not equivalent to minimizing the classification error

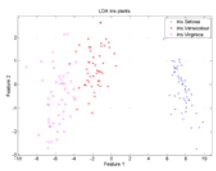
Limitations of Fisher's discriminant

- It produces at most M-*1* feature projections
- Its criterion function is based on class-wise distributions of limited complexity (all classes have a similarly-shaped Gaussian distribution)
- It will fail when the discriminatory information is not in the mean but in the variance of the data

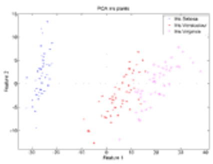
Fisher's discriminant example



← Original 3D data



Best 2 Fisher's



Best 2 PCA

Summary

- PCA (unsupervised)
  - Max variance <-> min projection error
  - Eigenvectors of sample cov.mat. / scatter matrix
- Fisher's linear discriminant (supervised)
  - Maximizes spread of means while minimizing intra-class spread
  - $\mathbf{S}_{wy}=\mathbf{I}$ and «whitening of data»
  - Eigenvectors of $\mathbf{S}_w^{-1}\mathbf{S}_b$
  - At most nClasses-1 features
  - Limitations

Literature on pattern recognition

- A review on statistical pattern recognition (still good fourteen years on):
  - A. Jain, R. Duin and J. Mao: Statistical pattern recognition: a review, IEEE Trans. Pattern analysis and Machine Intelligence, vol. 22, no. 1, January 2001, pp. 4--
- Classical PR-books
  - R. Duda, P. Hart and D. Stork, Pattern Classification, 2. ed. Wiley, 2001
  - B. Ripley, Pattern Recognition and Neural Networks, Cambridge Press, 1996.
  - S. Theodoridis and K. Koutroumbas, Pattern Recognition, Academic Press, 2006.