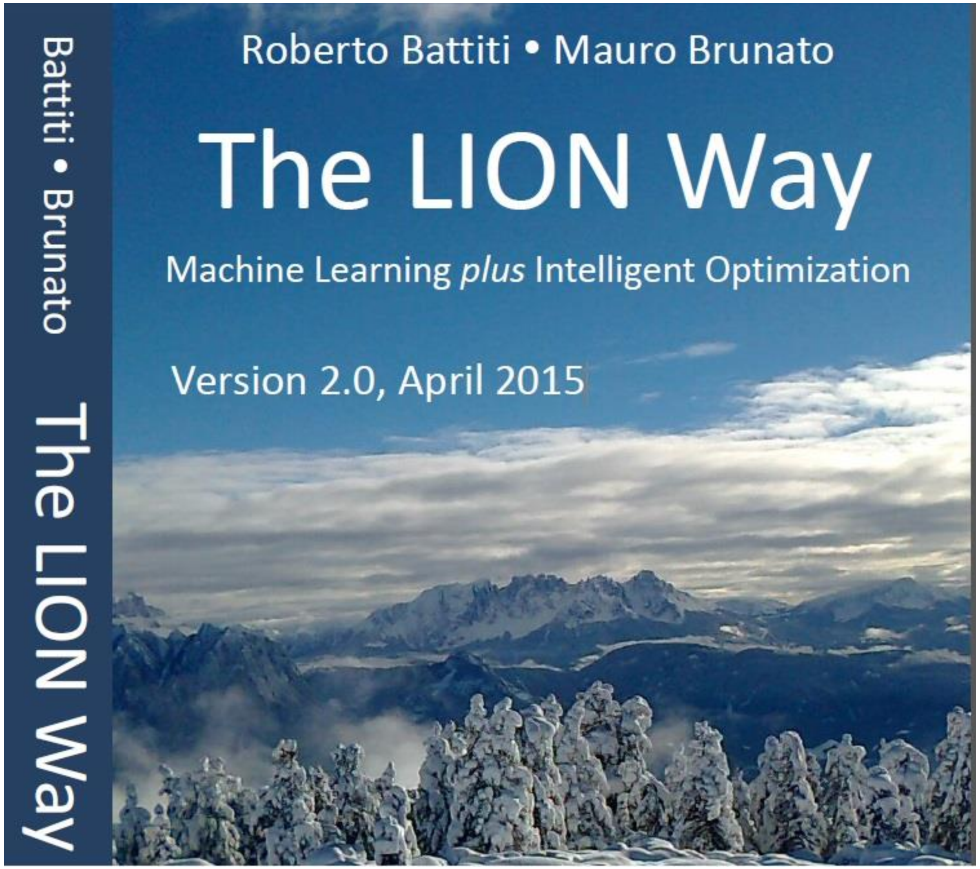
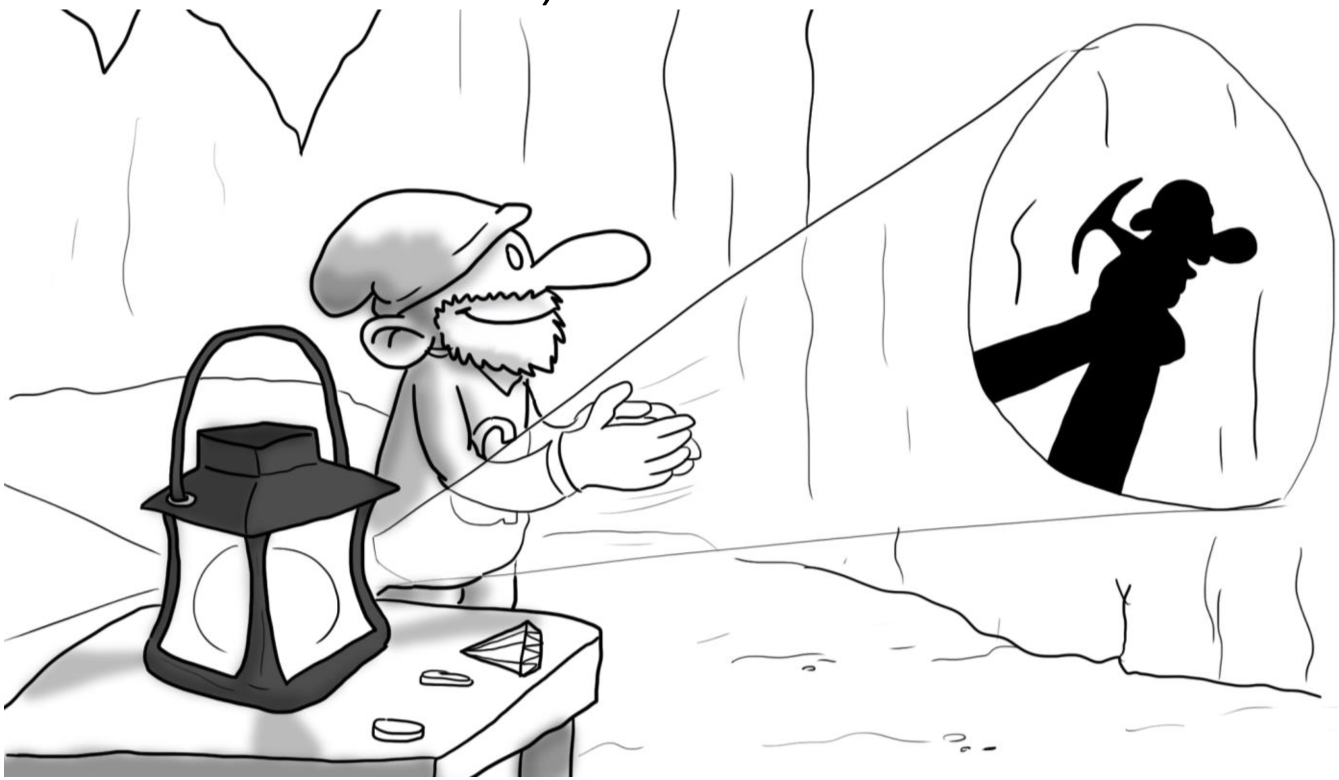


ROBERTO BATTITI, MAURO BRUNATO.
The LION Way: Machine Learning plus *Intelligent Optimization*.
LIONlab, University of Trento, Italy,
Apr 2015

http://intelligent-optimization.org/LIONbook

Dimensionality reduction by linear transformations (projections)

You, who are blessed with shade as well as light, you, who are gifted with two eyes, endowed with a knowledge of perspective, and charmed with the enjoyment of various colors, you, who can actually see an angle, and contemplate the complete circumference of a Circle in the happy region of the Three Dimensions – how shall I make it clear to you the extreme difficulty which we in Flatland experience in recognizing one another's configuration?
(Flatland - 1884 -Edwin Abbott Abbott)



Dimensionality reduction

- Mapping entities to two or three dimensions helps the visual analysis of the data
- The mapping has to preserve most of the information present in the original data.
- The objective is to organize entities in two-three dimensions so that similar objects are near to each other and dissimilar objects are far from each other.

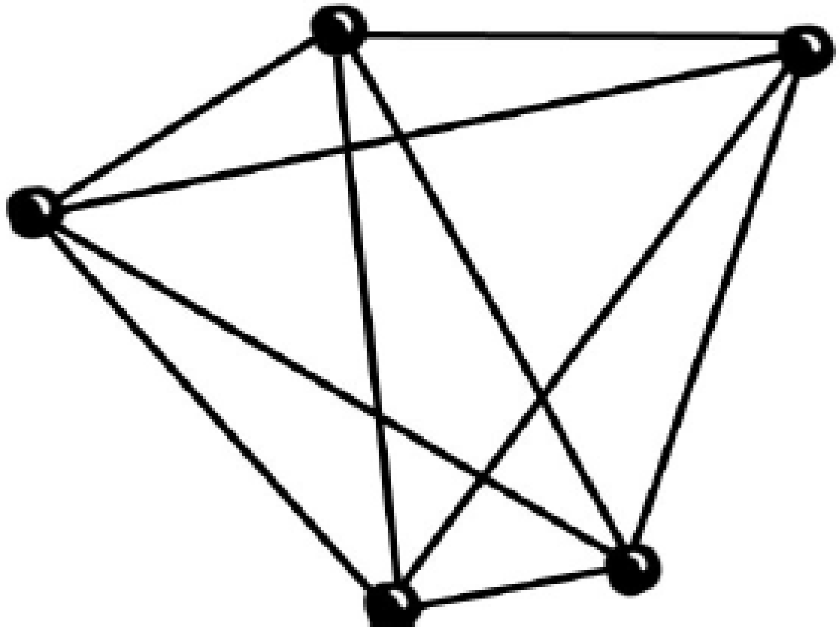
- In general, reducing dimensions to a smaller number (but greater than 2 o 3) can be very useful for subsequent machine learning steps

How to measure dissimilarity

- Dissimilarity measures can be explicitly given for any pair of entities (external dissimilarity) or can be derived from an internal representation of the entities (i.e. from their coordinates) from a distance metric.
- In some cases the information given to the system consists of *both* **coordinates and relationships** (think about labeling entities after a clustering is performed)

External dissimilarity structure.

- N entities, characterized by some mutual dissimilarities δ_{ij} can represented by a graph



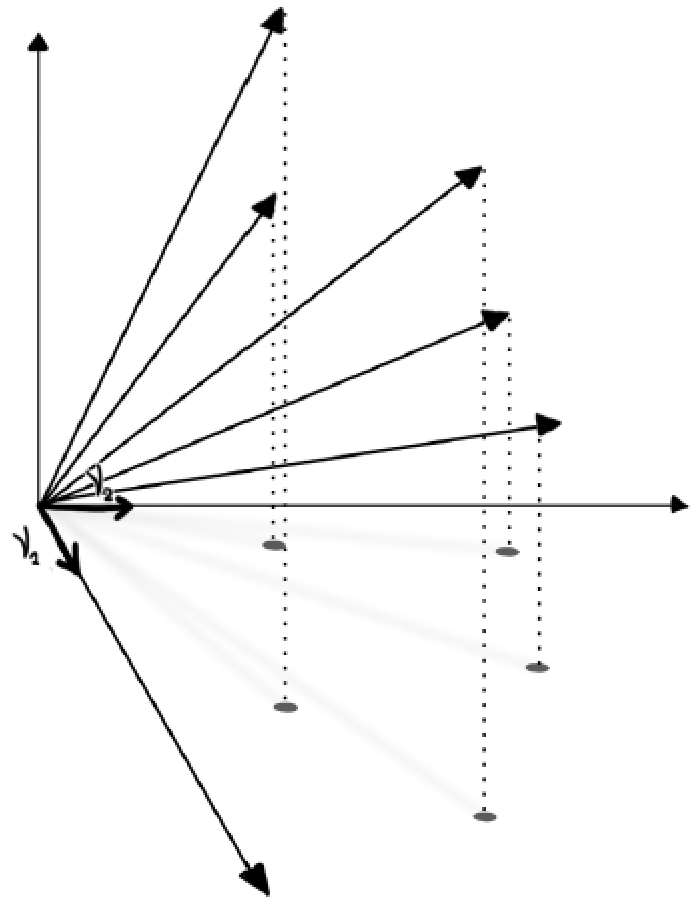
- Each node represents an entity, and a connection with weight δ_{ij} is present between two nodes, if and only if a distance δ_{ij} is defined for the corresponding entities.

Notation

Assumption: data is **centered** (the mean of each coordinate is zero), otherwise subtract mean.

1. n number of vectors (entities)
2. m dimension of each vector
3. **X** n x m matrix storing the entities (one row for each entity: $X_{i\alpha}$ α-th coordinates of item i)
4. Latin indices i j range over the data items, while Greek indices α,β range over the coordinates
5. **S** is the the m x m biased covariance matrix: $S = 1/n\ X^T X$

Linear projection



A projection: each dotted line connecting a vector to its projection is perpendicular to the plane defined by v1 and v2 (in this case the direction vectors are the X and Y axes, in general a projection can be on any plane identified by two linearly independent vectors).

Linear projection (2)

- Linear transformation L of the items to a space of dimension p
- L is represented by a p x m matrix, acting on vector x by matrix multiplication
  $y = Lx$
- The p rows $v^1, \ldots, v^p$ in R^m of L are called direction vectors, and have unit norm
- Each coordinate in the transformed p-dimensional space is obtained by projecting the original vector x onto v^α .
- The coordinate vectors are given by $x^1 = Xv^1$,..., $x^p = Xv^p$.

Orthogonal projections

- If the direction vectors $v^1, \ldots, v^p$ are **mutually orthogonal** and with unit norm: $v^i \bullet v^i = \delta_{ij}$, we have an orthogonal projection
- Example: a selection of a subset of the original coordinates (in this case $v^i = (0, 0, \ldots, 1, \ldots 0)$)
- The visualization is simple because it shows **genuine properties of the data**
- On the contrary, nonlinear transformations may deform the original data distribution in arbitrary and potentially very complex and counter-intuitive ways
- Linear projection are efficient both computationally and with respect to storage requirements.

Principal Component Analysis (PCA)

- PCA finds the orthogonal projection that **maximizes the sum of all squared pairwise distances** between the projected data elements.
- Let dist_{ij}^{p} be the distance between the projections of two data points i and j

$$\text{dist}_{ij}^{p} = \sqrt{\sum_{\alpha=1}^{p}((X\nu^{\alpha})_i - (X\nu^{\alpha})_j)^2},$$

- PCA maximizes $\sum_{i<j}(\text{dist}_{ij}^{p})^2.$

Maximize variance,
Spread points as much as possible

Principal Component Analysis (2)

- With a projection, distances can only decrease, from Pythagora's theorem

$$\max_{\nu^1,\ldots,\nu^p} \sum_{i<j} (\mathrm{dist}^p_{ij})^2 \leq \sum_{i<j} (\mathrm{dist}_{ij})^2.$$

- The objective of PCA is to approximate as much as possible the original sum of squared distances, in a space of smaller dimension, by projection.

Principal Component Analysis (3)

- Define the n x n unit Laplacian matrix L^u, as
  $L^u_{ij} = (n \bullet \delta_{ij} - 1)$
  (a key tool for describing pairwise relationships)

- The optimization problem becomes

$$\max_{\nu^1,\ldots,\nu^p} \quad \sum_{\alpha=1}^{p} (\nu^\alpha)^T X^T L^u X \nu^\alpha$$

$$\text{subject to} \quad \nu^\alpha \cdot \nu^\beta = \delta_{\alpha,\beta}, \qquad \alpha, \beta = 1, \ldots, p.$$

- The solution is given by the p eigenvectors with largest eigenvalues of the m x m matrix $X^T L^u X$

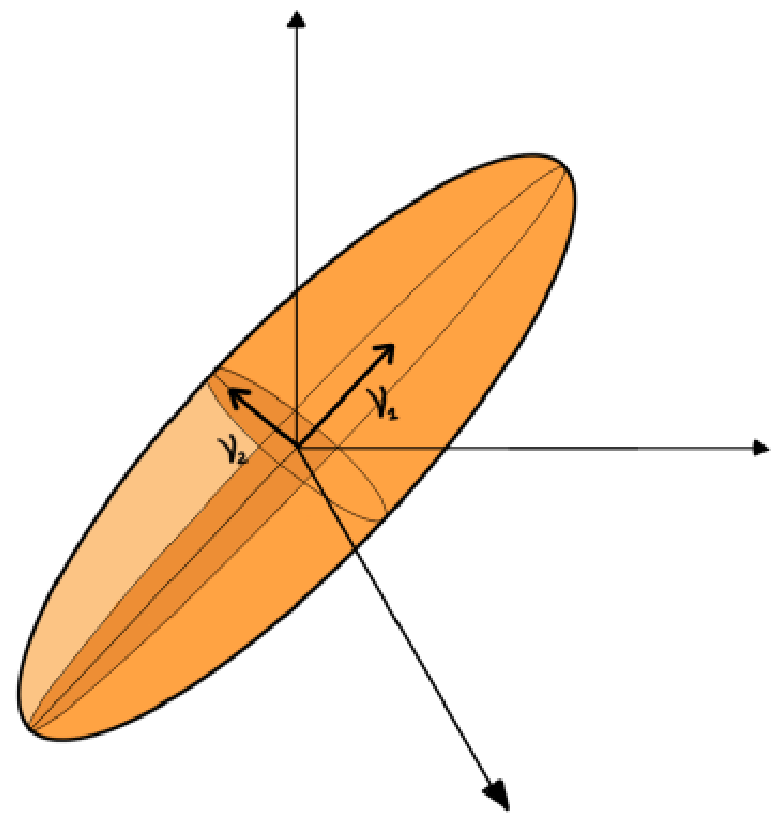
Principal Component Analysis (4)

- For centered coordinates, the following holds true: $X^T L^u X = n^2 S$. That is, the matrix is **proportional to the covariance** matrix, hence:

- The solutions for PCA is obtained by finding the **eigenvectors** of the **covariance** matrix with largest eigenvalue

Meaning of PCA

- PCA transforms a number of possibly correlated variables into a smaller number of uncorrelated variables (principal components).
- The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible
- PCA minimizes the mean squared error incurred when approximating the data with their projections



Principal component analysis, data are described by an **ellipsoid**. The first eigenvector is in the direction of the longest principal axis, the second eigenvector lies in a plane perpendicular to the first one, along the longest axis of the two-dimensional ellipse.

Limitations of PCA

PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance.

- The computational cost is related to solving the eigenvalues-eigenvectors for a matrix of dimension m x m and is not related to the number of points n, hence it is **very fast if the dimension is limited**

Limitations:

- Having a larger **variance** is not always related to having a larger **information content** (it can be a side-effect of the choice of physical units)
- PCA is sensitive to **outliers** (the sum of squared distances in involved in the optimization)
- If PCA is used for feature selection for classification, its main limitation is that it **makes no use of the class label** of the feature vector

Weighted PCA: combining coordinates *and* relationships

- in some cases additional information is available in the form of relationships between (some) entities
- extend the PCA approach to incorporate additional information:
- minimize a weighted sum of the squared projected distances (d_{ij} represents the weight)

$$\sum_{i<j} d_{ij} \cdot (\text{dist}^p_{ij})^2.$$

Optimization problem for weighted PCA

- we can now assign to the problem an n x n Laplacian matrix L^d

$$L_{ij}^d = \begin{cases} \sum_{j=1}^n d_{ij} & \text{if } i = j \\ -d_{ij} & \text{otherwise} \end{cases}$$

- The optimal projection are obtained as the direction vectors given by the p highest eigenvectors of the matrix $X^T L^d X$.

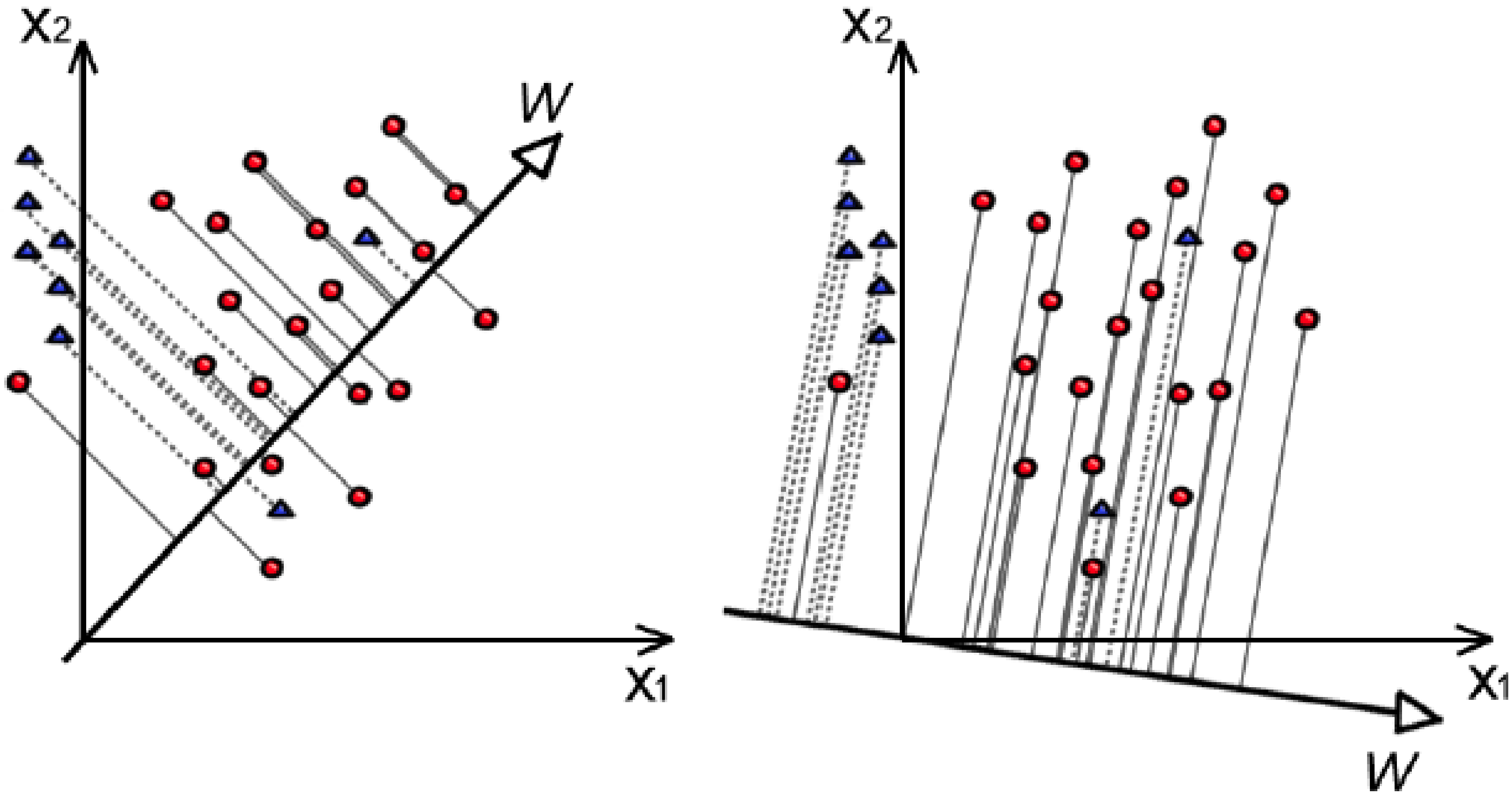
Tuning dissimilarity values to create variations of PCA

- In normalized PCA $d_{ij} = 1/disti_{ij}$ to discount large original distances in the optimization (useful to increase **robustness with respect to outliers**)
- In supervised PCA, with data labeled as belonging to different classes, we can set dissimilarities d_{ij} to a small value if i and j belong to the same class, to a value 1 if they belong to different classes (more important to put at large distances points of *different* clusters)

Fisher linear discrimination by ratio optimization

- Fisher analysis deals with finding **a vector** $\mathbf{v}_F$ such that, when the original vectors are **projected** onto it, values of the different classes are separated in the best possible way.
- A nice separation is obtained when the sample **means** of the projected points are as different as possible, when normalized with respect to the average **scatter** of the projected points.

Fisher linear discrimination



Fisher linear discrimination (triangles belong to one class, circles to another one): the one-dimensional projection on the left mixes projected points from the two classes, while the projection on the right best separates the projected sample means w.r.t. the projected scatter.

Fisher linear discriminant: formal definition

- Let n_i be the number of points in the i-th cluster, let μ_i and S_i be the mean vector and the biased covariance matrix for the i-th cluster
- The matrix $\bar{S}_{\text{within}} = \frac{1}{n}\sum_{i=1}^{c} n_i S_i$ is the average within-cluster covariance matrix
- and $S_{\text{between}} = \frac{1}{n}\sum_{i=1}^{c} n_i \mu_i \mu_i^T$ is the average between cluster covariance matrix.

Fisher linear discriminant: formal definition (2)

- Fisher linear discriminant is defined as the linear function y = $v^T x$ for which the following ratio is maximized:

$$\frac{\nu^T S_{\text{between}}\ \nu}{\nu^T S_{\text{within}}\ \nu}.$$

- We want to maximally separate the clusters (the role of the numerator where the projections of the means count) and keep the clusters as compact as possible (the role of the denominator).

Fisher linear discriminant for 2 classes

- For the special case of two classes, the Fisher linear discriminant is the linear function $y = w^T x$ for which the following criterion function is maximized:

$$\text{Separation}(\mathbf{w}) = \frac{\|\widetilde{m}_1 - \widetilde{m}_2\|^2}{\widetilde{s}_1^2 + \widetilde{s}_2^2},$$

- where m_i is the sample mean for the projected points

$$\widetilde{m}_i = (1/n_i) \sum_{y \in Class_i} y,$$

- s^2 is the scatter of the projected samples of each class

$$\widetilde{s}_i^2 = \sum_{y \in Class_i} (y - \widetilde{m}_i)^2$$

Fisher linear discriminant for 2 classes (2)

- The solution is:

$$\mathbf{w}_F = (S_w)^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- where m_i is the d-dimensional sample mean for class i and S_w is the sum of the two scatter matrices S_i defined as follows

$$S_i = \sum_{\mathbf{x} \in Class_i} (\mathbf{x}_i - \mathbf{m}_i)(\mathbf{x}_i - \mathbf{m}_i)^T.$$

Fisher discrimination index for selecting features

- Two-way classification problem with input vectors of d dimensions
- Rate the importance of feature i according to the **magnitude of the i-th component of the Fisher vector** w_F
- The largest components in Fisher vector will heuristically identify the most relevant coordinates for separating the classes
- Rationale: if the direction of a coordinate vector is similar to the direction of the Fisher vector, we can use that coordinate to separate the two classes
- Drawbacks: inverting the matrix can be computationally demanding and this measure can be insufficient to order many features

Fisher's linear discriminant analysis (LDA)

- LDA: find p different projection direction instead of just one, maximizing the following ratio:

$$\max_{\nu^1,\dots,\nu^p} \quad \frac{\sum_{\alpha=1}^{p} (\nu^\alpha)^T \; S_{\text{between}} \; \nu^\alpha}{\sum_{\alpha=1}^{p} (\nu^\alpha)^T \; S \; \nu^\alpha}$$

$$\text{subject to} \quad (\nu^\alpha)^T S \nu^\beta = \delta_{\alpha\beta}, \qquad \alpha, \beta = 1, \dots, p.$$

Fisher's linear discriminant analysis (LDA)(2)

- LDA is sensitive to outliers and it does not take the shape and size of clusters into consideration.
- This can be taken into account maximizing the following ratio:

$$\max_{\nu^1,\ldots,\nu^p} \quad \frac{\sum_{i<j} d_{ij}(\mathrm{dist}^p_{ij})^2}{\sum_{i<j} \mathrm{sim}_{ij}(\mathrm{dist}^p_{ij})^2}$$

$$\text{subject to} \quad (\nu^\alpha)^T X^t L^s X \nu^\beta = \delta_{\alpha\beta}, \qquad \alpha,\beta = 1,\ldots,p,$$

where d_{ij} are dissimilarity weights, sim_{ij} are similarity weights and L^s is the Laplacian matrix corresponding to the similarities

$$L^s_{ij} = \begin{cases} \sum_{j=1}^n \mathrm{sim}_{ij} & \text{if } i = j \\ -\mathrm{sim}_{ij} & \text{if } i \neq j \end{cases}.$$

Conclusions

- finding an **optimal** projection requires defining in measurable ways what is meant by optimality
- **unsupervised** (based only on coordinates) and **supervised** information (based on relationships), can be combined to give different weights to different preferences for placing items distant or close.
- What is left is to derive m x m matrices and to use an efficient and numerically stable way to solve an m x m generalized eigenvector problem

GIST

- **Visualizations** help the human unsupervised learning capabilities to extract knowledge from the data.
- They are limited to the two-three dimensions
- A simple way to transform data into two-dimensional views is through **projections**
- **Orthogonal projections** can be intuitively explained as looking at the data from different and distant points of view

GIST(2)

- **Principal Component Analysis (PCA)** identifies the orthogonal projection that spreads the points as much as possible in the projection plane
- But keep in mind that having a larger **variance** is not always related to having the largest **information** content
- If available, relationships can be used to modify PCA (**weighted PCA**) and obtain more meaningful projections.

GIST(3)

- When **class labels** are present, Fisher discrimination projects data so that the ratio of difference in the projected **means** of points belonging to different classes divided by the **intra-class scatter** is maximized.