Pattern Classification

EET3053

Lecture 05: Dimensionality Reduction

Dr. Kundan Kumar
Associate Professor
Department of ECE

Faculty of Engineering (ITER)
S'O'A Deemed to be University, Bhubaneswar, India-751030

Topics to be covered

- Dimensionality Problem
- Dimensionality/Feature reduction
  - Principal component analysis
  - Linear discriminant analysis
    - Fisher Linear discriminant
    - Multiple Discriminant Analysis
- Feature Selection

Dimensionality Problem

Introduction

- In practical multicategory applications, it is not unusual to encounter problems involving tens or hundreds of features.
- Intuitively, it may seem that each feature is useful for at least some of the discriminations.
- In general, if the performance obtained with a given set of features is inadequate, it is natural to consider adding new features.
- Even though increasing the number of features increases the complexity of the classifier, it may be acceptable for an improved performance.

Introduction

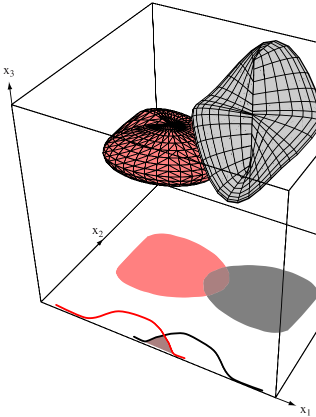


Figure: There is a non-zero Bayes error in the one-dimensional x_1 space or the two-dimensional x_1, x_2 space. However, the Bayes error vanishes in the x_1, x_2, x_3 space because of non-overlapping densities.

Problems of Dimensionality

- Unfortunately, it has frequently been observed in practice that, beyond a certain point, adding new features leads to worse rather than better performance.
- This is called the *curse of dimensionality*.
- There are two issues that we must be careful about:
  - How is the classification accuracy affected by the dimensionality (relative to the amount of training data)?
  - How is the complexity of the classifier affected by the dimensionality?

Problems of Dimensionality

- Potential reasons for increase in error include
  - wrong assumptions in model selection,
  - estimation errors due to the finite number of training samples for high-dimensional observations (overfitting).
- Potential solutions include
  - reducing the dimensionality,
  - simplifying the estimation.

Problems of Dimensionality

- Dimensionality can be reduced by
  - redesigning the features,
  - selecting an appropriate subset among the existing features,
  - combining existing features.
- Estimation errors can be simplified by
  - assuming equal covariance for all classes (for the Gaussian case),
  - using regularization,
  - using prior information and a Bayes estimate,
  - using heuristics such as conditional independence,
  - $\ldots$.

Problem of Dimensionality

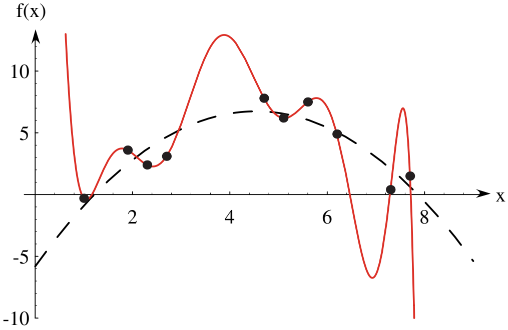


Figure: The "training data" (black dots) were selected from a quadratic function plus Gaussian noise, i.e, $f(x) = ax^2 + bx + c + \varepsilon$ where $p(\varepsilon) \approx N(0, \sigma^2)$. The 10th degree polynomial shown fits the data perfectly, but we desire instead the second-order function $f(x)$, since it would lead to better predictions for few samples.

Problem of Dimensionality

- All of the commonly used classifiers can suffer from the curse of dimensionality.
- While an exact relationship between the probability of error, the number of training samples, the number of features, and the number of parameters is very difficult to establish, some guidelines have been suggested.
- It is generally accepted that using at least ten times as many training samples per class as the number of features $(n/d > 10)$ is a good practice.

Feature/Dimensionality Reduction

Component Analysis and Discriminants

- One way of coping with the problem of high dimensionality is to reduce the dimensionality by combining features.
- Issues in feature/dimensionality reduction:
  - Linear vs. non-linear transformations.
  - Use of class labels or not (depends on the availability of training data).
- Linear combinations are particularly attractive because they are simple to compute and are analytically tractable.
- Linear methods project the high-dimensional data onto a lower dimensional space.
- Advantages of these projections include
  - reduced complexity in estimation and classification,
  - ability to visually examine the multivariate data in two or three dimensions.

Component Analysis and Discriminants

- Given $\mathrm{x} \in \mathbb{R}^d$, the goal is to find a linear transformation A that gives $\mathrm{y} = \mathrm{A}^T\mathrm{x}$, $\mathrm{y} \in \mathbb{R}^{d'}$ where $d' < d$.

- Two classical approaches for finding optimal linear transformations are:
  - *Principal Components Analysis (PCA)*: Seeks a projection that best represents the data in a least-squares sense.
  - *Multiple Discriminant Analysis (MDA)*: Seeks a projection that best separates the data in a least-squares sense.

Principal Component Analysis

Principal Component Analysis

- Given $\mathrm{x}_1, \mathrm{x}_2, \dots, \mathrm{x}_n \in \mathbb{R}^d$, the goal is to find a d'-dimensional subspace where the reconstruction error of x_i in this subspace is minimized.
- The squared-error criterion function $J_0(\mathrm{x}_0)$ by

$$J_0(\mathrm{x}_0) = \sum_{k=1}^{n} \|\mathrm{x}_0 - \mathrm{x}_k\|^2$$

  and seek the value of x_0 that minimizes J_0
- It is simple to show that the solution to this problem is given by $\mathrm{x}_0 = \mathrm{m}$, where m is the sample mean.

$$\mathrm{m} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{x}_k$$

Principal Component Analysis

- This can be easily verified by writing

$$
\begin{aligned}
J_0(\mathrm{x}_0) &= \sum_{k=1}^{n} \|(\mathrm{x}_0 - \mathrm{m}) - (\mathrm{x}_k - \mathrm{m})\|^2 \\
&= \sum_{k=1}^{n} \|(\mathrm{x}_0 - \mathrm{m})\|^2 - 2\sum_{k=1}^{n} (\mathrm{x}_0 - \mathrm{m})^T (\mathrm{x}_k - \mathrm{m}) + \sum_{k=1}^{n} \|(\mathrm{x}_k - \mathrm{m})\|^2 \\
&= \sum_{k=1}^{n} \|(\mathrm{x}_0 - \mathrm{m})\|^2 - 2(\mathrm{x}_0 - \mathrm{m})^T \sum_{k=1}^{n} (\mathrm{x}_k - \mathrm{m}) + \sum_{k=1}^{n} \|(\mathrm{x}_k - \mathrm{m})\|^2 \\
&= \sum_{k=1}^{n} \|(\mathrm{x}_0 - \mathrm{m})\|^2 + \underbrace{\sum_{k=1}^{n} \|(\mathrm{x}_k - \mathrm{m})\|^2}_{\text{independent of } \mathrm{x}_0}
\end{aligned}
$$

- Since the second sum is independent of x_0, So the above expression is obviously minimized by the choice of $\mathrm{x}_0 = \mathrm{m}$.

Principal Component Analysis

- The sample mean is a zero-dimensional representation of the data set. It is simple, but it does not reveal any of the variability in the data.
- One-dimensional representation by projecting the data onto a line running through the sample mean.
- Let e be a unit vector in the direction of the line. Then equation of line will be

$$\mathrm{x} = \mathrm{m} + a\mathrm{e}$$

  where a is any real value, corresponds to the distance of any point x form the mean m.
- If $\mathrm{x}_k = \mathrm{m} + a_k\mathrm{e}$, then we can find optimal set of coefficients a_k by minimizing the squared-error criterion function.

Principal Component Analysis

- Squared-error criterion function

$$
\begin{aligned}
J_1(a_1, a_2, \ldots, a_n, \mathrm{e}) &= \sum_{k=1}^{n} \|(\mathrm{m} + a_k \mathrm{e}) - \mathrm{x}_k\|^2 \\
&= \sum_{k=1}^{n} \|a_k \mathrm{e} - (\mathrm{x}_k - \mathrm{m})\|^2 \\
&= \sum_{k=1}^{n} a_k^2 \|\mathrm{e}\|^2 - 2\sum_{k=1}^{n} a_k \mathrm{e}^T (\mathrm{x}_k - \mathrm{m}) + \sum_{k=1}^{n} \|(\mathrm{x}_k - \mathrm{m})\|^2
\end{aligned}
$$

- Recognize that $||e|| = 1$, partially differentiating with respect to a_k, and setting the derivative to zero, we obtain

$$
a_k = \mathrm{e}^T (\mathrm{x}_k - \mathrm{m})
$$

- Geometrically, this result merely says that we obtain a least-squares solution by projecting the vector x_k onto the line in the direction of e that passes through the sample mean.

Principal Component Analysis

- The solution to the problem involves the scatter matrix S defined by

$$\mathrm{S} = \sum_{k=1}^{n} (\mathrm{x}_k - \mathrm{m})(\mathrm{x}_k - \mathrm{m})^T$$

- Scatter matrix is n times the sample covariance matrix.
- Substitute a_k in the cost function

$$\begin{aligned}
J_1(\mathrm{e}) &= \sum_{k=1}^{n} a_k^2 - 2\sum_{k=1}^{n} a_k^2 + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2 \\
&= -\sum_{k=1}^{n} \left[\mathrm{e}^T(\mathrm{x}_k - \mathrm{m})\right]^2 + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2 \\
&= -\sum_{k=1}^{n} \mathrm{e}^T(\mathrm{x}_k - \mathrm{m})(\mathrm{x}_k - \mathrm{m})^T\mathrm{e} + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2 \\
&= -\mathrm{e}^T\mathrm{S}\mathrm{e} + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2
\end{aligned}$$

Principal Component Analysis

- So the resulting cost function

$$J_1(\mathrm{e}) = -\mathrm{e}^T \mathrm{S}\mathrm{e} + \sum_{k=1}^{n} \|\mathrm{x}_k - \mathrm{m}\|^2$$

- Use Lagrange multipliers to maximize $\mathrm{e}^T \mathrm{S}\mathrm{e}$ subject to the constraint that $\|\mathrm{e}\| = 1$.
- Letting λ be the undetermined multiplier, we differentiate

$$u = \mathrm{e}^T \mathrm{S}\mathrm{e} - \lambda(\mathrm{e}^T \mathrm{e} - 1)$$
$$\frac{\partial u}{\partial \mathrm{e}} = 2\mathrm{S}\mathrm{e} - 2\lambda \mathrm{e}$$
$$\mathrm{S}\mathrm{e} = \lambda \mathrm{e}$$

- In particular, because $\mathrm{e}^T \mathrm{S}\mathrm{e} = \lambda \mathrm{e}^T \mathrm{e} = \lambda$, it follows that to maximize $\mathrm{e}^T \mathrm{S}\mathrm{e}$, so select the eigenvector corresponding to the largest eigenvalue of the scatter matrix.

Principal Component Analysis

- To find the best one-dimensional projection of the data (best in the least-sum-of-squared-error sense), we project the data onto a line through the sample mean in the direction of the eigenvector of the scatter matrix having the largest eigenvalue.

- This result can be readily extended from 1-D to a d'-D projection.

$$\mathrm{x} = \mathrm{m} + \sum_{i=1}^{d'} a_i \mathrm{e}_i$$

where $d' \leq d$.

Principal Component Analysis

- It is not difficult to show that the criterion function

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathrm{m} + \sum_{i=1}^{d'} a_{ki} \mathrm{e}_i \right) - \mathrm{x}_k \right\|^2$$

is minimized when the vector $\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_{d'}$ are the d' eigenvector of the scatter matrix having the largest eigenvalues.

- Because the scatter matrix is real and symmetric, these eigenvectors are orthogonal.

- The coefficients a_i are the components of x in that basis, and are the principal components.

Principal Component Analysis

- Given $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n \in \mathbb{R}^d$, the goal is to find a d'-dimensional subspace where the reconstruction error of x_i in this subspace is minimized.
- The squared error criterion function $J_0(\mathrm{x}_0)$ can be minimized by selecting $\mathrm{x}_0 = \mathrm{m}$, where m is the sample mean.
- The sample mean is a zero-dimensional representation of the data set. It is simple, but it does not reveal any of the variability in the data.
- We must consider at least one-dimensional representation of data by choosing

$$\mathrm{x} = \mathrm{m} + a\mathrm{e}$$

  and compute the optimal value of a such that the squared error criterion function J_1 is minimum.
- We obtained the solution as

$$a_k = \mathrm{e}^T(\mathrm{x}_k - \mathrm{m})$$

Principal Component Analysis

- Given $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n \in \mathbb{R}^d$, the goal is to find a d'-dimensional subspace where the reconstruction error of x_i in this subspace is minimized.
- The criterion function for the reconstruction error can be defined in the least squares sense as

$$J_{d'} = \sum_{k=1}^{n} \left\| \left(\mathrm{m} + \sum_{i=1}^{d'} a_{ki} \mathrm{e}_i \right) - \mathrm{x}_k \right\|^2$$

where $\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_{d'}$ are the bases for the subspace (stored as the columns of A) and a_i is the projection of x_i onto that subspace.

Principal Component Analysis

- It can be shown that $J_{d'}$ is minimized when $\mathrm{e}_1, \mathrm{e}_2, \ldots, \mathrm{e}_{d'}$ are eigenvectors corresponding to first d' largest eigenvalues of scatter matrix.

$$S = \sum_{k=1}^{n} (\mathrm{x}_k - \mathrm{m})(\mathrm{x}_k - \mathrm{m})^T$$

- The coefficients $a = (a_i, \ldots, a_{d'})^T$ are called the principal components.
- When the eigenvectors are sorted in descending order of the corresponding eigenvalues, the greatest variance of the data lies on the first principal component, the second greatest variance on the second component, and so on.

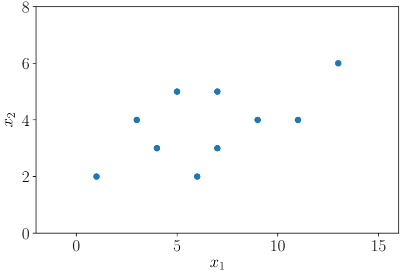
Example to be solved

Question: Given the following sets of feature vector belonging to two classes ω_1 and ω_2 which is Gaussian distributed.

$$\left\{ \begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 4 \\ 3 \end{pmatrix}, \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \begin{pmatrix} 7 \\ 5 \end{pmatrix} \right\} \in \omega_1$$

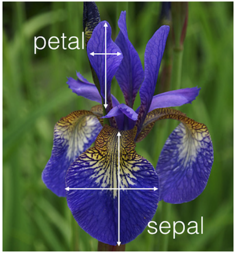
$$\left\{ \begin{pmatrix} 6 \\ 2 \end{pmatrix}, \begin{pmatrix} 9 \\ 4 \end{pmatrix}, \begin{pmatrix} 7 \\ 3 \end{pmatrix}, \begin{pmatrix} 11 \\ 4 \end{pmatrix}, \begin{pmatrix} 13 \\ 6 \end{pmatrix} \right\} \in \omega_2$$

Find out the best direction of the line of projection that best represent the data in one-dimensional feature space.



Examples: Iris dataset representation

- "Iris" dataset is very famous dataset used for data analysis problems (classification, feature reduction, and many more)
- Available on the UCI machine learning repository `https://archive.ics.uci.edu/ml/datasets/Iris`.
- The iris dataset contains measurements for 150 iris flowers from three different species.
  - Iris-setosa ($n_1 = 50$)
  - Iris-versicolor ($n_2 = 50$)
  - Iris-virginica ($n_3 = 50$)
- And the four features of in Iris dataset are:
  - sepal length in cm
  - sepal width in cm
  - petal length in cm
  - petal width in cm



Examples: Iris data representation

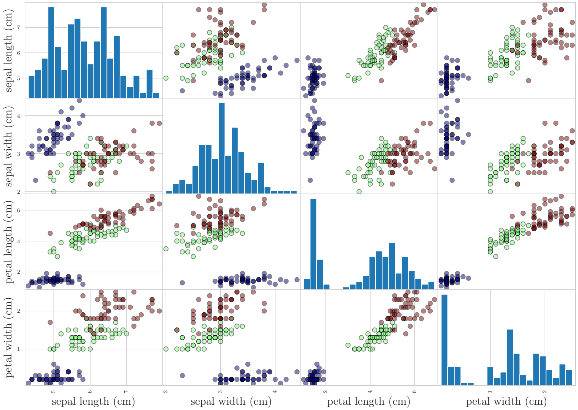
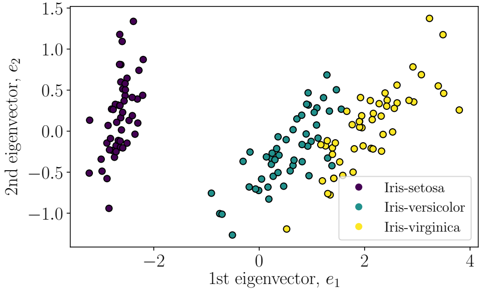


Figure: Scatter plot of the iris data. Diagonal cells show the histogram for each feature. Other cells show scatters of pairs of features x_1, x_2, x_3, x_4 in top-down and left-right order. Red, green and blue points represent samples for the setosa, versicolor and virginica classes, respectively.

Examples: Iris data representation



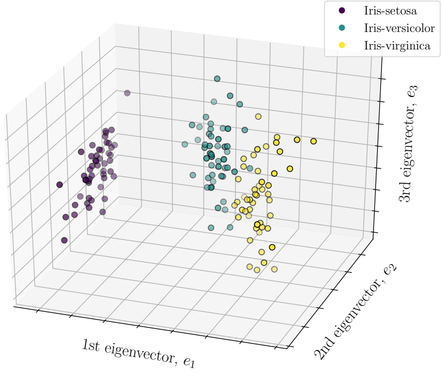


Figure: Scatter plot of the projection of the iris data onto the first two and the first three principal axes.

Linear Discriminant Analysis

Fisher Linear Discriminant

- PCA seeks directions that are efficient for representation, *discriminant analysis* seeks directions that are efficient for discrimination.

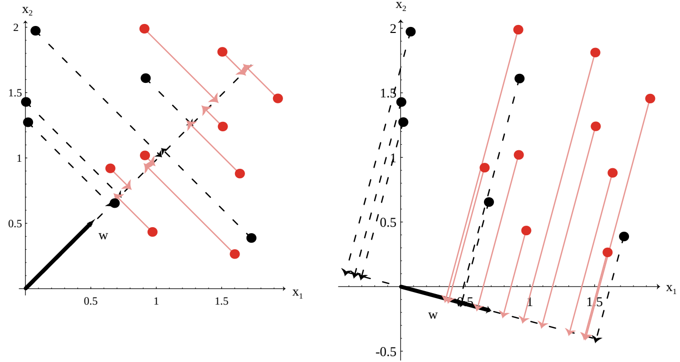


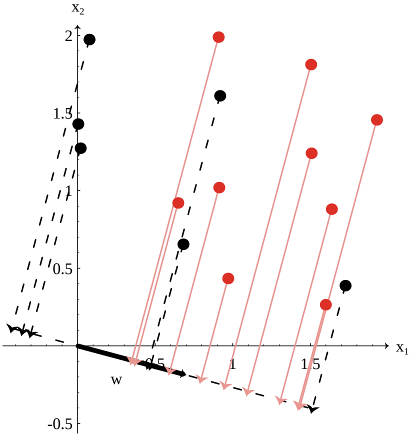
Figure: Projection of the same set of samples onto two different lines in the directions marked as w. The figure on the right shows greater separation between the red and black projected points

Fisher Linear Discriminant

- Suppose $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n \in \mathbb{R}^d$ are divided into two subsets $\mathcal{D}_1$ (n_1 samples) and $\mathcal{D}_2$ (n_2 samples) corresponding to the classes ω_1 and ω_2 respectively, the goal is to find a projection onto a line defined as

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

  such that the points corresponding to $\mathcal{D}_1$ and $\mathcal{D}_2$ are well separated.
- A corresponding set of n samples $y_1, y_2, \ldots, y_n$ divided into the subset $\mathcal{Y}_1$ and $\mathcal{Y}_2$.



Fisher Linear Discriminant

- The criterion function for the best separation can be defined as

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

  where, $\tilde{m}_i$ is the sample mean and $\tilde{s}_i^2$ is the scatter for the projected samples labeled ω_i, given as

$$\tilde{m}_i = \frac{1}{n_i} \sum_{y \in \mathcal{Y}_i} y \qquad \tilde{s}_i^2 = \sum_{y \in \mathcal{Y}_i} (y - \tilde{m}_i)^2$$

- This is called the *Fisher's linear discriminant* with the geometric interpretation that the best projection makes the difference between the means as large as possible relative to the variance.

Fisher Linear Discriminant

- To compute the optimal w, we define the *scatter matrices* S_i

$$\mathbf{S}_i = \sum_{\mathbf{x} \in \mathcal{D}_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t$$

where,

$$\mathbf{m}_i = \frac{1}{n_i} \sum_{\mathbf{x} \in \mathcal{D}_i} \mathbf{x}$$

- The *within-class scatter matrix* S_W

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

and the *between-class scatter matrix* S_B

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

Fisher Linear Discriminant

- Then, the criterion function becomes

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

  This expression is well known in mathematical physics as the generalized Rayleigh quotient.

- A vector w that maximizes $J(\cdot)$ must satisfy

$$\mathrm{S}_B\mathrm{w} = \lambda \mathrm{S}_W \mathrm{w}$$
$$\mathrm{S}_W{}^{-1}\mathrm{S}_B\mathrm{w} = \lambda \mathrm{w}$$

- In this particular case, it is unnecessary to solve for the eigenvalues and eigenvectors of $\mathrm{S}_W{}^{-1}\mathrm{S}_B$ due to the fact that $\mathrm{S}_B\mathrm{w}$ is always in the direction of $\mathrm{m}_1 - \mathrm{m}_2$

Fisher Linear Discriminant

- So we can find the immediate solution as

$$\mathbf{w} = \mathbf{S}_W^{-1}(\mathbf{m}_1 - \mathbf{m}_2)$$

- Note that, $\mathrm{S_W}$ is symmetric and positive semidefinite, and it is usually nonsingular if $n > d$. $\mathrm{S_B}$ is also symmetric and positive semidefinite, but its rank is at most 1.
- Thus, we have obtained w for Fisher's linear discriminant – the linear function yielding the maximum ratio of between-class scatter to within-class scatter.

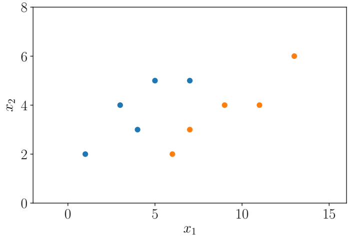
Example to be solved

Question: Given the following sets of feature vector belonging to two classes ω_1 and ω_2 which is Gaussian distributed.

$$\left\{\begin{pmatrix}1\\2\end{pmatrix},\begin{pmatrix}3\\4\end{pmatrix},\begin{pmatrix}4\\3\end{pmatrix},\begin{pmatrix}5\\5\end{pmatrix},\begin{pmatrix}7\\5\end{pmatrix}\right\}\in\omega_1$$

$$\left\{\begin{pmatrix}6\\2\end{pmatrix},\begin{pmatrix}9\\4\end{pmatrix},\begin{pmatrix}7\\3\end{pmatrix},\begin{pmatrix}11\\4\end{pmatrix},\begin{pmatrix}13\\6\end{pmatrix}\right\}\in\omega_2$$

Find out the best direction of the line of projection that best separates the data in one-dimensional feature space.



Multiple Discriminant Analysis

Multiple Discriminant Analysis

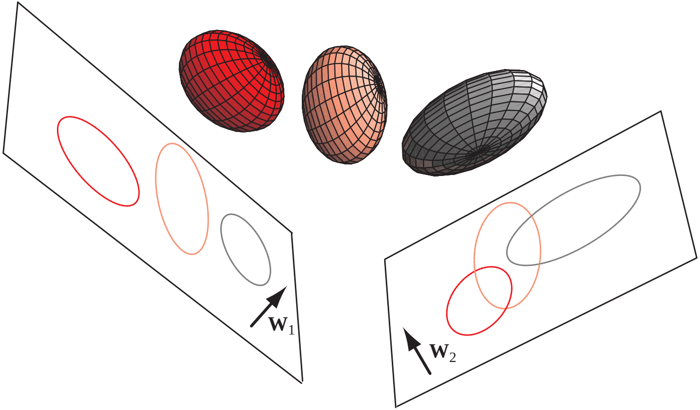


Figure: Three-dimensional distributions are projected onto two-dimensional subspaces, described by a normal vectors W_1 and W_2.

Multiple Discriminant Analysis

- The within-class scatter matrix

$$\mathbf{S}_W = \sum_{i=1}^{c} \mathbf{S}_i$$

where

$$\mathbf{S}_i = \sum_{\mathbf{x}\in\mathcal{D}_i} (\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t \qquad \mathbf{m}_i = \frac{1}{n_i}\sum_{\mathbf{x}\in\mathcal{D}_i} \mathbf{x}$$

- The proper generalization for S_B is not quite so obvious.
- Suppose that we define a *total mean vector* m and a *total scatter matrix* S_T by

$$\mathbf{m} = \frac{1}{n}\sum_{\mathbf{x}} \mathbf{x} = \frac{1}{n}\sum_{i=1}^{c} n_i \mathbf{m}_i \qquad \mathbf{S}_T = \sum_{\mathbf{x}} (\mathbf{x}-\mathbf{m})(\mathbf{x}-\mathbf{m})^t$$

Multiple Discriminant Analysis

- Then we can write

$$\begin{aligned}\mathbf{S}_T &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})(\mathbf{x}-\mathbf{m}_i+\mathbf{m}_i-\mathbf{m})^t \\ &= \sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{x}-\mathbf{m}_i)(\mathbf{x}-\mathbf{m}_i)^t+\sum_{i=1}^{c}\sum_{\mathbf{x}\in\mathcal{D}_i}(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t \\ &= \mathbf{S}_W+\sum_{i=1}^{c}n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t.\end{aligned}$$

- Therefore,

$$\mathrm{S}_T = \mathrm{S}_W + \mathrm{S}_B$$

where

$$\mathbf{S}_B = \sum_{i=1}^{c} n_i(\mathbf{m}_i-\mathbf{m})(\mathbf{m}_i-\mathbf{m})^t$$

Multiple Discriminant Analysis

- The projection form a d-dimensional space to a $(c-1)$-dimensional space is accomplished by $c-1$ discriminant functions

$$y_i = \mathbf{w}_i^t \mathbf{x} \qquad i = 1, ..., c-1$$

- If the y_i are viewed as components of a vector y and the weight vector w_i are viewed as the columns of a d-by-$(c-1)$ matrix W, then the projection can be written as a single matrix equation

$$\mathbf{y} = \mathbf{W}^t \mathbf{x}$$

Multiple Discriminant Analysis

- The samples $\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_n$ project to a corresponding set of samples $\mathrm{y}_1, \mathrm{y}_2, \ldots, \mathrm{y}_n$, which can be described by their own mean vectors and scatter matrices. Thus

$$\tilde{\mathbf{m}}_i = \frac{1}{n_i} \sum_{\mathbf{y} \in \mathcal{Y}_i} \mathbf{y}$$

$$\tilde{\mathbf{m}} = \frac{1}{n} \sum_{i=1}^{c} n_i \tilde{\mathbf{m}}_i$$

$$\tilde{\mathbf{S}}_W = \sum_{i=1}^{c} \sum_{\mathbf{y} \in \mathcal{Y}_i} (\mathbf{y} - \tilde{\mathbf{m}}_i)(\mathbf{y} - \tilde{\mathbf{m}}_i)^t$$

and

$$\tilde{\mathbf{S}}_B = \sum_{i=1}^{c} n_i (\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})(\tilde{\mathbf{m}}_i - \tilde{\mathbf{m}})^t,$$

Multiple Discriminant Analysis

- It is a straightforward matter to show that

$$\tilde{\mathbf{S}}_W = \mathbf{W}^t \mathbf{S}_W \mathbf{W}$$

and

$$\tilde{\mathbf{S}}_B = \mathbf{W}^t \mathbf{S}_B \mathbf{W}.$$

- These equations show how the within-class and between-class scatter matrices are transformed by the projection to the lower dimensional space.

Multiple Discriminant Analysis: Solution

- The criterion function

$$J(\mathbf{W}) = \frac{|\tilde{\mathbf{S}}_B|}{|\tilde{\mathbf{S}}_W|} = \frac{|\mathbf{W}^t\mathbf{S}_B\mathbf{W}|}{|\mathbf{W}^t\mathbf{S}_W\mathbf{W}|}$$

  the problem of finding a rectangular matrix W that maximized $J(\cdot)$.

- The columns of an optimal W are the generalized eigenvectors that correspond to the largest eigenvalues in

$$(\mathbf{S}_B - \lambda_i\mathbf{S}_W)\mathbf{w}_i = 0$$

Multiple Discriminant Analysis: Observation

$$(\mathbf{S}_B - \lambda_i \mathbf{S}_W)\mathbf{w}_i = 0$$

- If S_W is nonsingular, this can be converted to a conventional eigenvalue problem as before.
- Computation of the inverse of S_W is expensive.
- Instead, one can find the eigenvalues as the roots of the characteristic polynomial

$$|\mathbf{S}_B - \lambda_i \mathbf{S}_W| = 0$$

- and then solve

$$(\mathbf{S}_B - \lambda_i \mathbf{S}_W)\mathbf{w}_i = 0$$

directly for the eigenvectors w_i.

Feature Selection

Feature Selection

- Feature reduction uses a linear or non-linear combination of features.
- An alternative to feature reduction is feature selection that reduces dimensionality by selecting subsets of existing features.
- Benefits of performing feature selection:
  - avoid curse of dimensionality
  - reduce the computational cost
  - improves accuracy
  - avoid overfitting
- The first step in feature selection is to define a criterion function that is often a function of the classification error.
- Note that, the use of classification error in the criterion function makes feature selection procedures dependent on the specific classifier used.

Feature Selection

- The most straightforward approach would require
  - examining all $\binom{d}{m}$ possible subsets of size m,
  - selecting the subset that performs the best according to the criterion function.
- The number of subsets grows combinatorially, making the exhaustive search impractical.
- There are two main types of feature selection algorithms:
  - Wrapper Feature Selection Methods.
  - Filter Feature Selection Methods.

Examples: Iris data representation

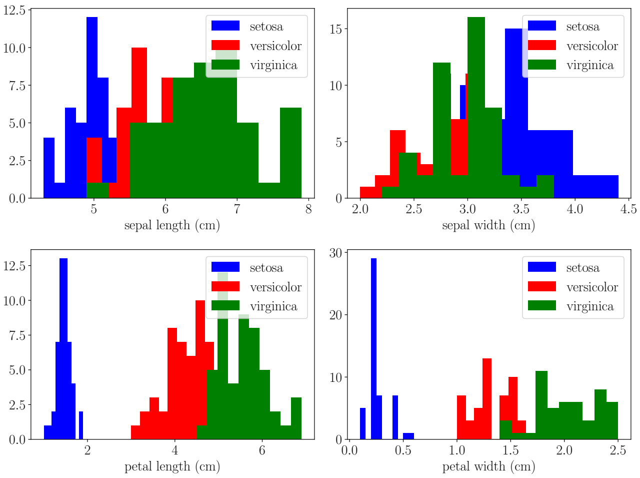


Figure: Histogram plot of Iris features

Examples: Iris data representation

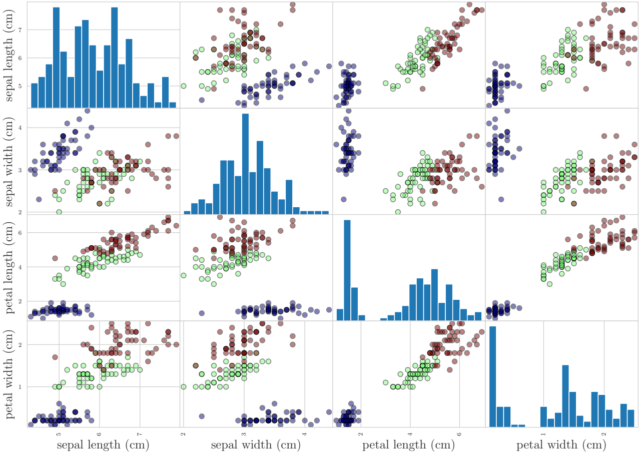
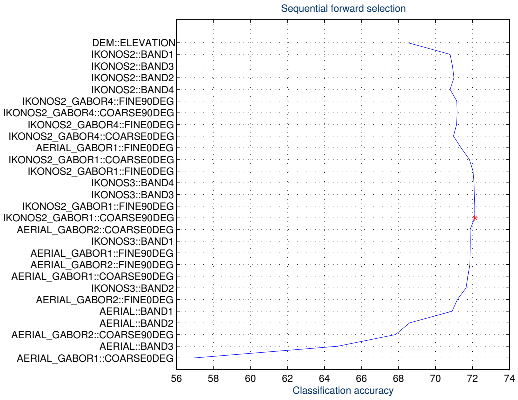


Figure: Scatter plot of the iris data. Off-diagonal cells show scatters of pairs of features x_1, x_2, x_3, x_4.

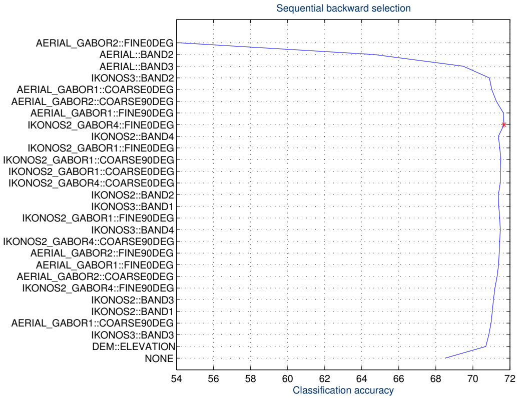
Feature Selection

- *Sequential forward selection:*
  1. First, the best single feature is selected.
  2. Then, pairs of features are formed using one of the remaining features and this best feature, and the best pair is selected.
  3. Next, triplets of features are formed using one of the remaining features and these two best features, and the best triplet is selected.
  4. This procedure continues until all or a predefined number of features are selected.



Feature Selection

- *Sequential backward selection:*
  - First, the criterion function is computed for all d features.
  - Then, each feature is deleted one at a time, the criterion function is computed for all subsets with $d-1$ features, and the worst feature is discarded.
  - Next, each feature among the remaining $d-1$ is deleted one at a time, and the worst feature is discarded to form a subset with $d-2$ features.
  - This procedure continues until one feature or a predefined number of features are left.



Summary

- The choice between feature reduction and feature selection depends on the application domain and the specific training data.
- Feature selection leads to savings in computational costs and the selected features retain their original physical interpretation.
- Feature reduction with transformations may provide a better discriminative ability but these new features may not have a clear physical meaning.

Assignment Problem

Question:

(a) Given the following sets of feature vector belonging to two classes ω_1 and ω_2 which is Gaussian distributed.

$$(1,2)^t,\ (3,5)^t,\ (4,3)^t,\ (5,6)^t,\ (7,5)^t \in \omega_1$$
$$(6,2)^t,\ (9,4)^t,\ (10,1)^t,\ (12,3)^t,\ (13,6)^t \in \omega_2$$

The vector are projected onto a line to represent the feature vectors by a single feature. Find out the best direction of the line of projection that maintains the separability of the two classes.

(b) Assuming the mean of the projected point belonging to ω_1 to be the origin of the projection line, identify the point on the projection line that optimally separates two classes. Assume the classes to be equally probable and the projected features also follow Gaussian distribution.

References

[1] Hart, P. E., Stork, D. G., & Duda, R. O. (2000). Pattern classification. Hoboken: Wiley.

[2] Gose, E. (1997). Pattern recognition and image analysis.

Thank you!