Feature Selection and Dimensionality Reduction for Classification

Feature Selection and Dimensionality Reduction

In many applications, we start with many features. Example: the gray value of each pixel as a feature in an image classification problem.

What do we do?

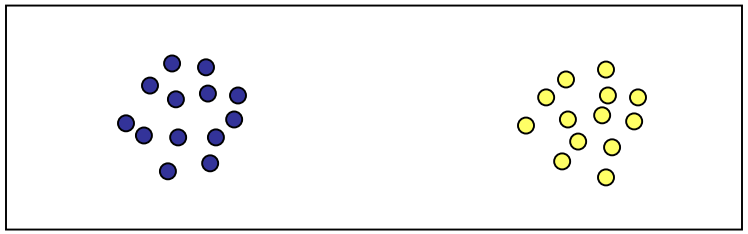
- Select a subset of all m available features, i.e., select l ($l<m$) features to form the feature vectors.
- Generate lower-dimensional "new" feature vectors from the original set of features (dimensionality reduction), e.g., principle component analysis (PCA).

Class Separability Measure

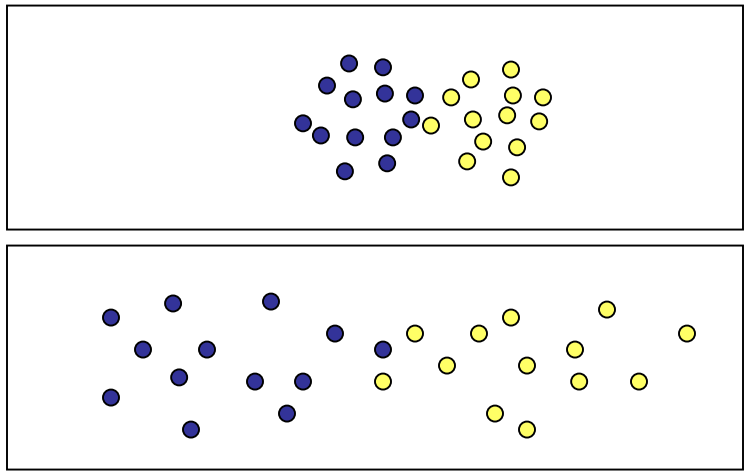
How do we determine which subset of features to use? Example criterion:

Class separability: Measures of how well the training samples of different classes are separated from each other.

Good separability:



Not-so-good separability



Or else, we can use the classification performance (e.g. classification error probability) itself as the criterion. However, this can impose high computational cost, depending on the classifier design.

Scatter Matrices

Within-class scatter matrix:

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i \qquad \text{with} \qquad \Sigma_i = \underset{\boldsymbol{x}\in\omega_i}{E}\left[(\boldsymbol{x}-\boldsymbol{\mu}_i)(\boldsymbol{x}-\boldsymbol{\mu}_i)^T\right]$$

$$\boldsymbol{\mu}_i = \underset{\boldsymbol{x}\in\omega_i}{E}[\boldsymbol{x}]$$

Between-class scatter matrix:

$$S_b = \sum_{i=1}^{M} P_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_0)^T \qquad \text{with} \qquad \boldsymbol{\mu}_0 = \sum_{i=1}^{M} P_i \boldsymbol{\mu}_i = E[\boldsymbol{x}]$$

Mixture scatter matrix:

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

It can be proved that $S_m = S_w + S_b$

Scatter Matrices

The following class separability measures are defined based on the scatter matrices:

$$J_1 = \frac{\mathrm{tr}\{S_m\}}{\mathrm{tr}\{S_w\}} \qquad J_2 = \frac{|S_m|}{|S_w|} = \left|S_w^{-1} S_m\right| \qquad J_3 = \mathrm{tr}\left\{S_w^{-1} S_m\right\}$$

Fisher's discriminant ratio: A class separability measure in 1-D, equiprobable cases:

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

In multi-class cases, we can use the average over all pairs of classes.

FDR can be useful in quantifying the class separability achievable with a single feature or a 1-D projection.

Feature Subset Selection

The goal is to find the best l features that retain as much information for classification in the original m features as possible.

Choosing the best l individual features

- Pro: Simple and fast
- Con: Loss of information of the correlation between features

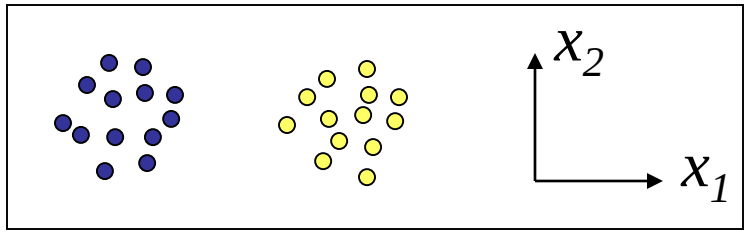
Choosing the best l-feature subset

- Pro: Retains the information of the correlation between features
- Con: Impractical to find the "optimal" subset (too many combinations to test)

Scalar Feature Selection

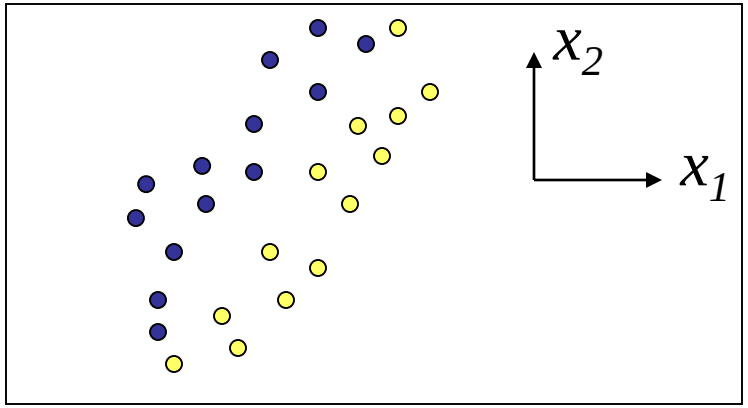
The idea is to test each of the m available features individually, and select the best l of them according to some criterion, such as a class separability measure.

Example:



Feature x_1 is better than feature x_2.

Problem with correlated features:



Neither feature is good individually, but they are useful when used together.

Sequential Feature Selection

Greedy approaches for (sub-optimal) feature subset selection:

Let X_k be the k-element feature subset selected.
Let C be the criterion used.
The goal is to choose l of m features.

Sequential Forward Selection

Start with $X_0=\{\}$
For k = 1 to l
  For all $x \notin X_{k-1}$, compute C for $\{x\} \cup X_{k-1}$
  $X_k \leftarrow$ the $\{x\} \cup X_{k-1}$ that gives the best C

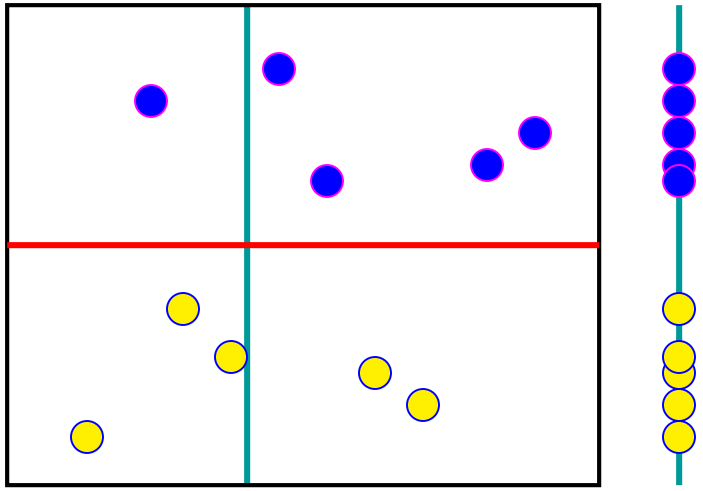
Sequential Backward Selection

Start with $X_l=\{\text{all } m \text{ features}\}$
For $k = m-1$ to l (decreasing)
  For all $x \in X_{k+1}$, compute C for $X_{k+1} \backslash \{x\}$
  $X_k \leftarrow$ the $X_{k+1} \backslash \{x\}$ that gives the best C

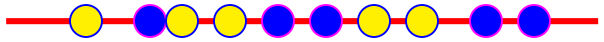
Fisher's Linear Discriminant

Goal: To maximize a class separability measure while projecting the feature vectors onto a one-dimensional subspace of the original m-dimensional feature space.

Example 2-D, 2-class case:



Good separability



Low separability

Let $\boldsymbol{w}$ be the vector of the direction of projection:

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

Fisher's Linear Discriminant

For each class, the mean and variance are also transformed similarly:

$$\mu_i = \boldsymbol{w}^T \boldsymbol{\mu}_i \quad \text{and} \quad \sigma^2 = \boldsymbol{w}^T \Sigma_i \boldsymbol{w}$$

For 2-class cases, we can easily see that FDR becomes

$$FDR(\boldsymbol{w}) = \frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}}$$

To maximize it, we set

$$\frac{\partial \left\{ (\boldsymbol{w}^T S_w \boldsymbol{w})^{-1} (\boldsymbol{w}^T S_b \boldsymbol{w}) \right\}}{\partial \boldsymbol{w}} = 0$$

We can then obtain

$$S_b \boldsymbol{w} = \left(\frac{\boldsymbol{w}^T S_b \boldsymbol{w}}{\boldsymbol{w}^T S_w \boldsymbol{w}} \right) S_w \boldsymbol{w} \quad \Rightarrow \quad S_b \boldsymbol{w} = \lambda S_w \boldsymbol{w}$$

A generalized eigenvalue problem

Fisher's Linear Discriminant

If S_w is invertible, we get $$S_w^{-1} S_b \boldsymbol{w} = \lambda \boldsymbol{w}$$

However, the vector $S_b\boldsymbol{w}$ is always along the $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ direction, i.e.,

$$S_b\boldsymbol{w} = \alpha(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

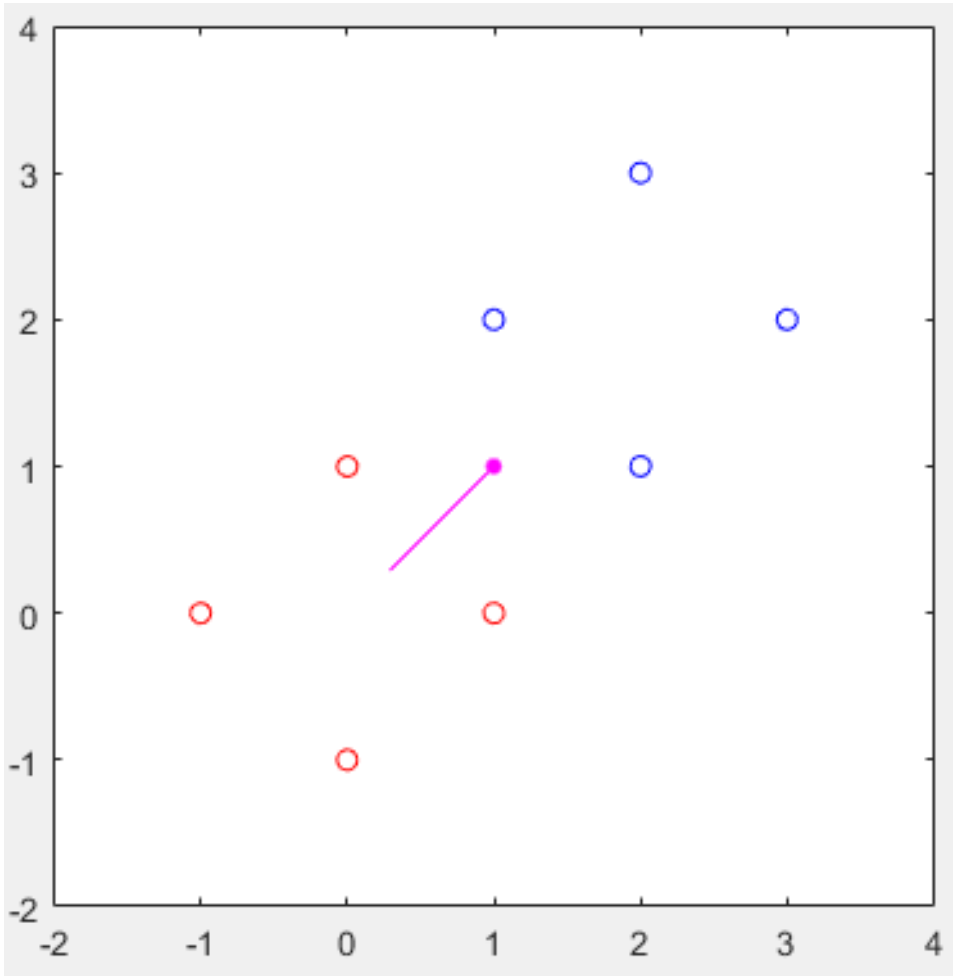
We can then obtain A as $$\boldsymbol{w} = S_w^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

We are only interested in the direction, so we drop all scalar factors of $\boldsymbol{w}$.

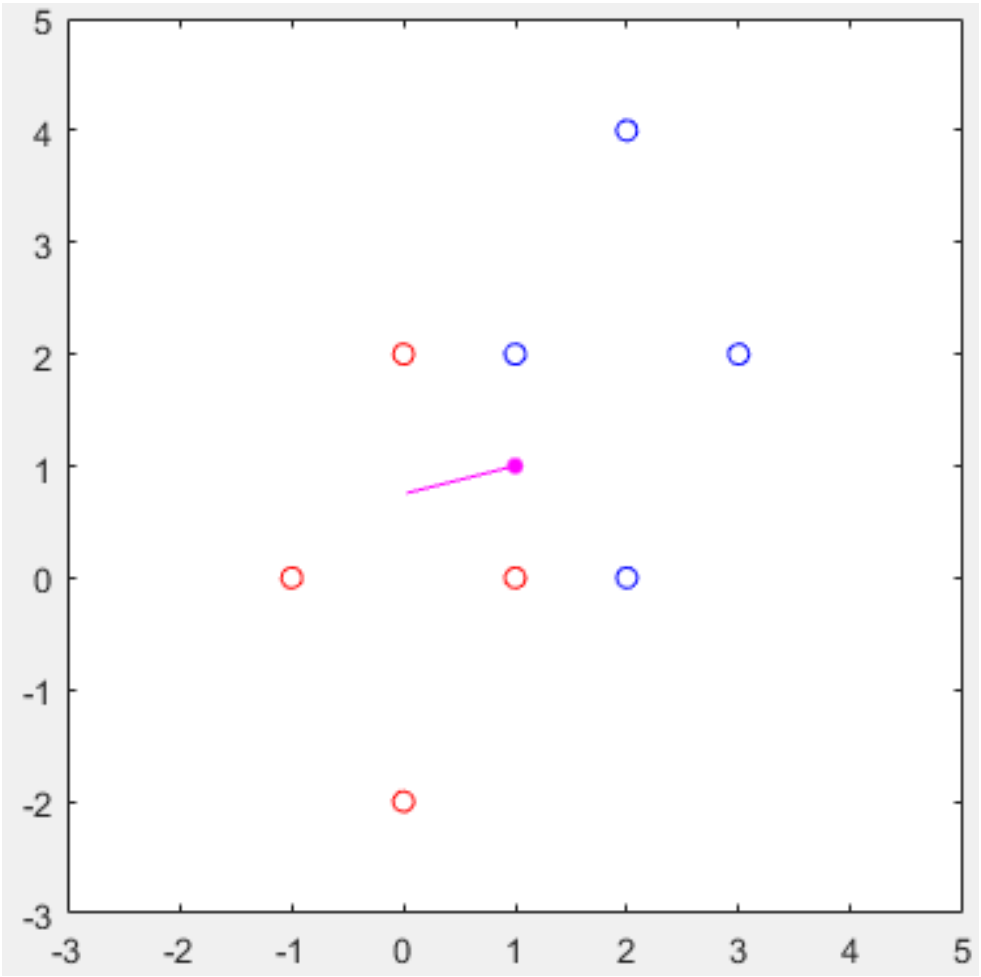
FLD gives a 1-D projection direction based on class separability and can be used as a linear classifier with a separately specified threshold. It can also be directly evaluated using, say, ROC curves.

FLD Example

$$X_1 = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 1 & -1 & 0 & 0 \end{bmatrix} \quad X_2 = \begin{bmatrix} 2 & 2 & 3 & 1 \\ 1 & 3 & 2 & 2 \end{bmatrix}$$



$$X_1 = \begin{bmatrix} 0 & 0 & 1 & -1 \\ 2 & -2 & 0 & 0 \end{bmatrix} \quad X_2 = \begin{bmatrix} 2 & 2 & 3 & 1 \\ 0 & 4 & 2 & 2 \end{bmatrix}$$



Linear Discriminant Analysis

Goal: To maximize a class separability measure while projecting the feature vectors onto a l-dimensional subspace of the original m-dimensional feature space.

The class separability measure used: $$J = \mathrm{tr}\left\{ S_w^{-1} S_b \right\}$$

Let A be the matrix representing the projection: $$\boldsymbol{y} = A^T \boldsymbol{x}$$

Scatter matrices are then transformed:

$$S_{yw} = A^T S_{xw} A \quad \text{and} \quad S_{yb} = A^T S_{xb} A$$

Linear Discriminant Analysis

The condition $\frac{\partial J_y}{\partial A} = 0$ now yields $S_{xw}^{-1} S_{xb} A = A S_{yw}^{-1} S_{yb}$

Using a matrix B that simultaneously diagonalizes S_{yw} and S_{yb}:

$B^T S_{yw} B = I$ and $B^T S_{yb} B = D$ (D is diagonal)

We obtain $(S_{xw}^{-1} S_{xb}) C = CD$ where $C=AB$.

This again is an eigenvalue problem. (We end up with C while looking for A; this is ok because the value of J is the same using either one as the transformation matrix.)

Linear Discriminant Analysis

The rank of S_{xb} is limited to $M-1$ (i.e., at most $M-1$ non-zero eigenvalues). So is $S_{xw}^{-1}S_{xb}$. As a result, we can project the data onto a subspace of at most $M-1$ dimensions, i.e., $l \leq M-1$. The special case $M=2$ gives FLD.

Let us now express J_y, which we want to maximize, using the sum of eigenvalues of $S_{xw}^{-1}S_{xb}$:

$$J_y = \mathrm{tr}\{D\}$$

When using $l < M-1$, we discard some eigenvectors of $S_{xw}^{-1}S_{xb}$ from C and their corresponding eigenvalues from D. The value of J_y is then the sum of the remaining eigenvalues. So we maximize J_y by projecting to the l-dimensional subspace containing the eigenvectors associated with the l largest eigenvalues.