CHAPTER 5 – FEATURE SELECTION

The ultimate goal in designing a classifier is to exhibit a good generalization performance. That is, to have a good error performance when dealing with data outside the training set.

A classifier may be designed to have very small error rate over the training data set, yet its generalization performance can be very poor.

It turns out that, in order to design a classifier with good generalization performance, the number of training data, N, must be large enough w.r. to its complexity.

For a large class of classifiers the complexity is directly related to the number of the features, i.e., the dimensionality of the feature space, l .

There are cases, however, where the complexity of the classifier does not depend on the dimensionality of the feature space, e.g., Support Vector Machines.

In any case, reducing the number of features is always necessary to get rid of uninformative features, or features that carry redundant information.

Prior to any feature selection, data preprocessing is a necessary step.

❖ Data Preprocessing

- Outlier removal: An outlier is defined as a point that lies very far from the mean of the corresponding random variable. Such points result in large errors during training. If such points are the result of erroneous measurements, they have to be removed.
- Data normalization: Features with large values have large influence compared to others with small values, although this may not necessarily reflect a respective significance towards the design of the classifier.

A common technique is to normalize each feature via the respective estimate of the mean and variance. That is for the k^{th} feature

$$\bar{x}_k = \frac{1}{N}\sum_{i=1}^{N} x_{ik} \quad , \; k = 1,2,\ldots,l$$

$$\sigma_k^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_{ik} - \bar{x}_k)^2$$

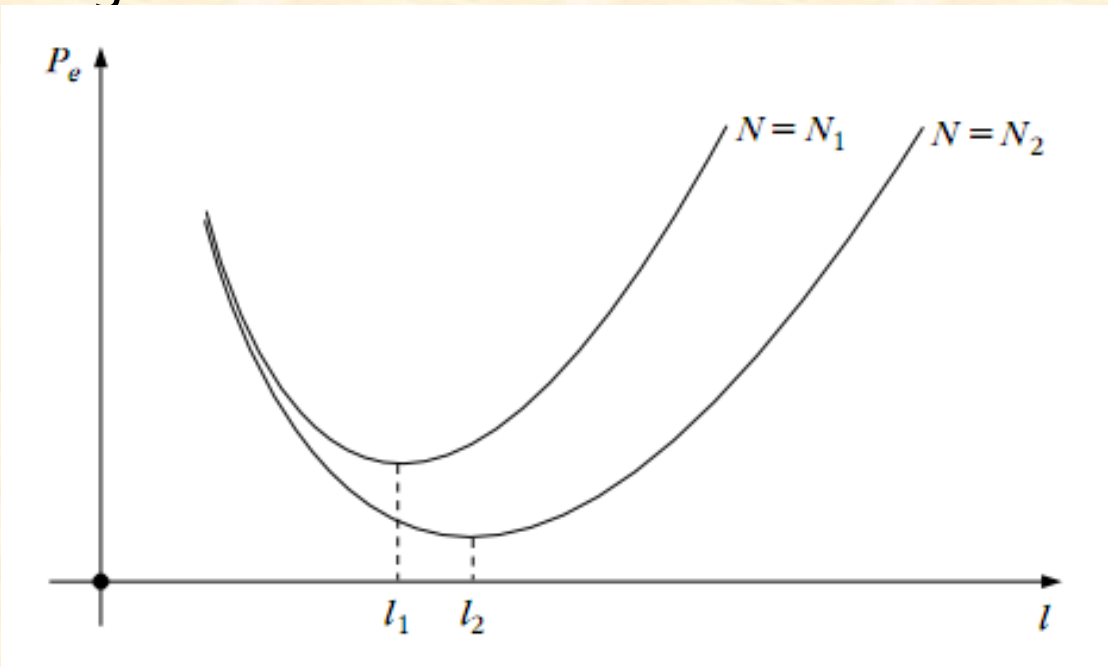
$$\hat{x}_{ik} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}$$

- Missing data: Given N training feature vectors, in some of them the values of certain features may be missing. The missing values can be completed by a number of methods, e.g.,
  - By embedding zeros.
  - By their unconditional mean.
  - By their conditional mean.
  - By more advanced techniques stemming from the theory of incomplete data (e.g., using EM arguments).

❖ The Peaking Phenomenon

If, in an ideal world, the class pdfs were known, then increasing the number of features would be beneficial.

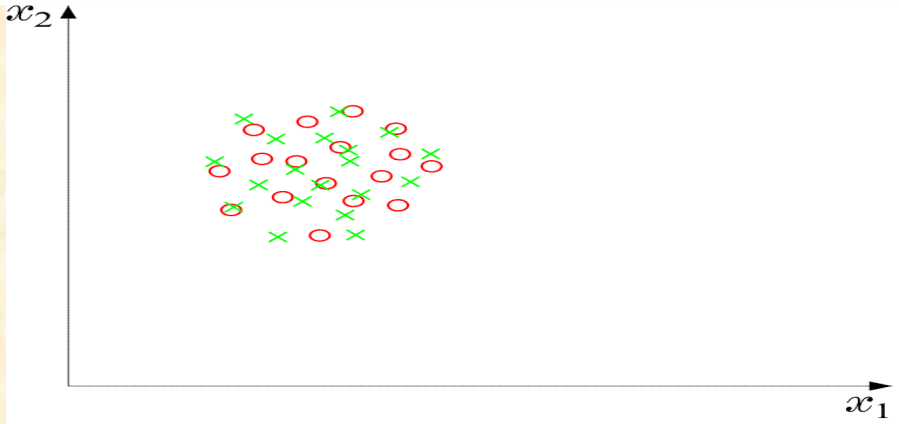
In practice, the general trend is that for a finite number of training points, increasing the number of features initially improves the generalization error rate, but after a certain value, the generalization error rate increases.



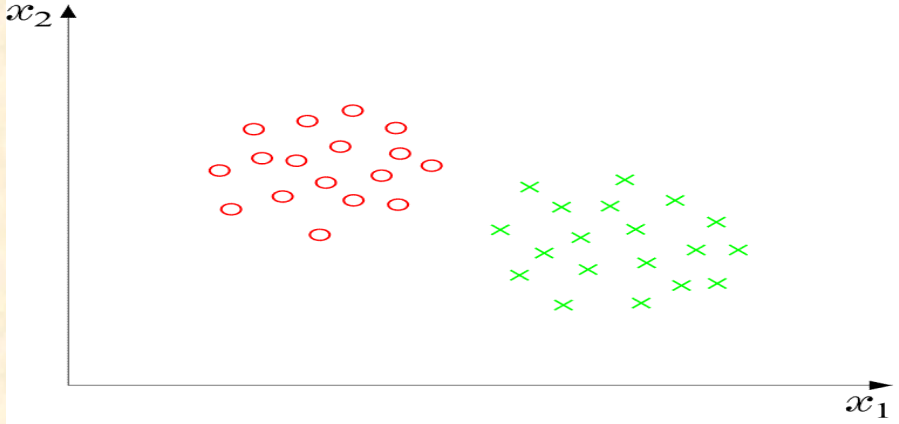
- The main goals in feature selection:
  - Select the "optimum" number l of features
  - Select the "best" l features
- Large l has a three-fold disadvantage:
  - High computational demands
  - Low generalization performance
  - Poor error estimates

- Given N
  - l must be large enough to learn
    - what makes classes different
    - what makes patterns in the same class similar
  - l must be small enough not to learn what makes patterns of the same class different

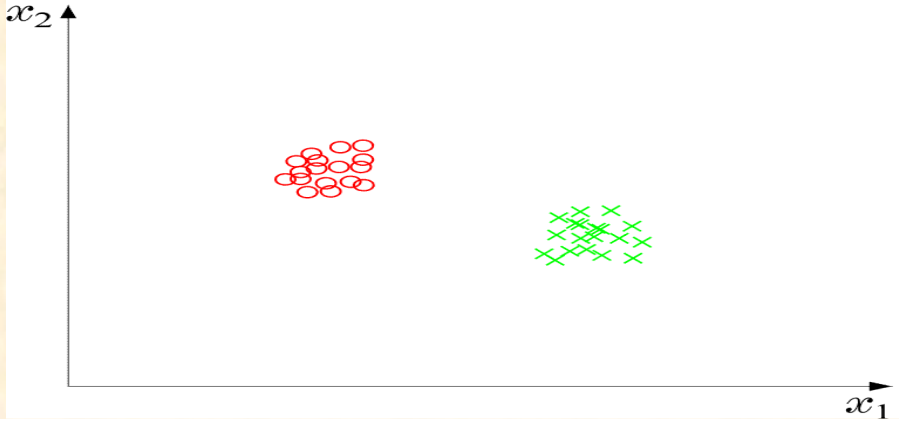
- Once l has been decided, choose the l most informative features
  - Best: **Large** between class distance,
    Small within class variance



Bad choice



Not bad choice



Good choice

- The basic philosophy
  - Discard individual features with poor information content
  - The remaining information rich features are examined jointly as vectors
- Feature Selection based on statistical Hypothesis Testing
  - The Goal: For each individual feature, find whether the values, which the feature takes for the different classes, differ significantly. This is based on the values of an appropriately chosen parameter . That is, answer
    - $H_1 : \theta = \theta_0$:The values differ significantly
    - $H_0 : \theta \neq \theta_0$:The values do not differ significantly

    If they do not differ significantly reject feature from subsequent stages.

❖ Class Separability Measures

The emphasis, so far, was on individually considered features. However, such an approach cannot take into account existing correlations among the features. That is, two features may be rich in information, but if they are highly correlated we need not consider both of them. To this end, in order to search for possible correlations, we consider features jointly as elements of vectors. To this end:

- Discard poor in information features, by means of a statistical test.

- Choose the maximum number, ℓ, of features to be used. This is dictated by the specific problem (e.g., the number, N, of available training patterns and the type of the classifier to be adopted).

- Combine remaining features to search for the "best" combination. To this end:

  - Use different feature combinations to form the feature vector. Train the classifier, and choose the combination resulting in the best classifier performance.

    A major disadvantage of this approach is the high complexity. Also, local minima, can give misleading results.

  - Adopt a class separability measure and choose the best feature combination against this cost.

- Class separability measures: Let $\underline{x}$ be the current feature combination vector.
  - Divergence. To see the rationale behind this cost, consider the two – class case. Obviously, if on the average the value of $\ln \frac{p(\underline{x}\,|\,\omega_1)}{p(\underline{x}\,|\,\omega_2)}$ is close to zero, then $\underline{x}$ should be a poor feature combination. Define:
    - $D_{12} = \int_{-\infty}^{+\infty} p(\underline{x}\,|\,\omega_1) \ln \frac{p(\underline{x}\,|\,\omega_1)}{p(\underline{x}\,|\,\omega_2)} d\underline{x}$
    - $D_{21} = \int_{-\infty}^{+\infty} p(\underline{x}\,|\,\omega_2) \ln \frac{p(\underline{x}\,|\,\omega_2)}{p(\underline{x}\,|\,\omega_1)} d\underline{x}$
    - $d_{12} = D_{12} + D_{21}$

    d_{12} is known as the divergence and can be used as a class separability measure.

– For the multi-class case, define d_{ij} for every pair of classes ω_i, ω_j and the average divergence is defined as

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} P(\omega_i) P(\omega_j) d_{ij}$$

– Some properties:

$$d_{ij} \geq 0$$

$$d_{ij} = 0, \text{if } i = j$$

$$d_{ij} = d_{ji}$$

– Large values of d are indicative of good feature combination.

- Scatter Matrices. These are used as a measure of the way data are scattered in the respective feature space.
  - Within-class scatter matrix

$$S_w = \sum_{i=1}^{M} P_i S_i$$

where

$$S_i = E\left[(\underline{x} - \underline{\mu}_i)(\underline{x} - \underline{\mu}_i)^T\right]$$

and

$$P_i \equiv P(\omega_i) \approx \frac{n_i}{N}$$

n_i the number of training samples in ω_i.

Trace $\{S_w\}$ is a measure of the average variance of the features.

- Between-class scatter matrix

$$S_b = \sum_{i=1}^{M} P_i \left(\underline{\mu}_i - \underline{\mu}_0\right)\left(\underline{\mu}_i - \underline{\mu}_0\right)^T$$

$$\underline{\mu}_0 = \sum_{i=1}^{M} P_i \underline{\mu}_i$$

Trace $\{S_b\}$ is a measure of the average distance of the mean of each class from the respective global one.

- Mixture scatter matrix

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

It turns out that:

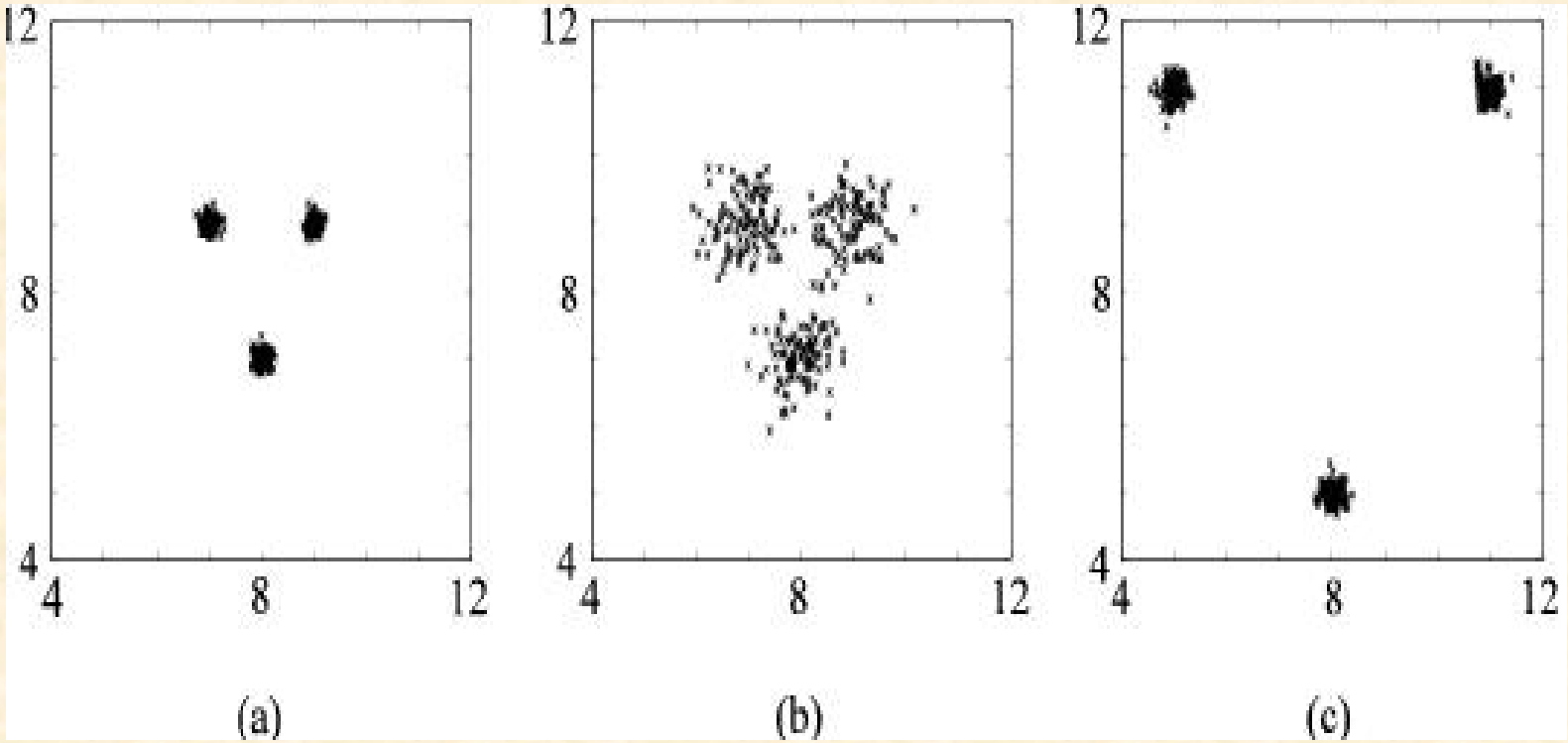
$$S_m = S_w + S_b$$

➢ Measures based on Scatter Matrices.

- $J_1 = \dfrac{\text{Trace}\{S_m\}}{\text{Trace}\{S_w\}}$
- $J_2 = \dfrac{|S_m|}{|S_w|} = |S_w^{-1} S_m|$
- $J_3 = \text{Trace}\{S_w^{-1} S_m\}$
- Other criteria are also possible, by using various combinations of S_m, S_b, S_w.

The above J_1, J_2, J_3 criteria take high values for the cases where:

- Data are clustered together within each class.
- The mean values for the various classes are far.


12
8
4
4
8
12
(a)
12
8
4
4
8
12
(b)
12
8
4
4
8
12
(c)

- Fisher's discriminant ratio. In one dimension and for two equiprobable classes the determinants become:

$$|S_w| \propto \sigma_1^2 + \sigma_2^2$$

$$|S_b| \propto (\mu_1 - \mu_2)^2$$

and

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

known as Fisher's ratio.

❖ Ways to combine features:

Trying to form all possible combinations of ℓ features from an original set of m selected features is a computationally hard task. Thus, a number of suboptimal searching techniques have been derived.

➢ Sequential forward selection. Let x_1, x_2, x_3, x_4 the available features (m=4). The procedure consists of the following steps:

- Adopt a class separability criterion (could also be the error rate of the respective classifier). Compute its value for ALL features considered jointly $[x_1, x_2, x_3, x_4]^T$.

- Eliminate one feature at a time and for each of the possible resulting combinations, that is $[x_1, x_2, x_3]^T$, $[x_1, x_2, x_4]^T$, $[x_1, x_3, x_4]^T$, $[x_2, x_3, x_4]^T$, compute the class reparability criterion value C. Select the best combination, say $[x_1, x_2, x_3]^T$.

- From the above selected feature vector, eliminate one feature and for each of the resulting combinations, $[x_1, x_2]^T$, $[x_2, x_3]^T$, $[x_1, x_3]^T$, compute C and select the best combination.

The above selection procedure shows how one can start from m features and end up with the "best" ℓ ones. Obviously, the choice is suboptimal. The number of required calculations is:

$$1+\frac{1}{2}\left((m+1)m-\ell(\ell+1)\right)$$

In contrast, a full search requires:

$$\binom{m}{\ell}=\frac{m!}{\ell!(m-\ell)!}$$

operations.

- Sequential backward selection. Here the reverse procedure is followed.
  - Compute C for each feature. Select the "best" one, say x_1
  - For all possible $2\mathrm{D}$ combinations of x_1, i.e., $[x_1, x_2]$, $[x_1, x_3]$, $[x_1, x_4]$ compute C and choose the best, say $[x_1, x_3]$.
  - For all possible $3\mathrm{D}$ combinations of $[x_1, x_3]$, e.g., $[x_1, x_3, x_2]$, etc., compute C and choose the best one.

The above procedure is repeated till the "best" vector with ℓ features has been formed. This is also a suboptimal technique, requiring:

$$\ell m - \frac{\ell(\ell-1)}{2}$$

operations.

➢ Floating Search Methods

The above two procedures suffer from the nesting effect. Once a bad choice has been done, there is no way to reconsider it in the following steps.

In the floating search methods one is given the opportunity in reconsidering a previously discarded feature or to discard a feature that was previously chosen.

The method is still suboptimal, however it leads to improved performance, at the expense of complexity.

Optimal Feature Generation

- In general, feature generation is a problem-dependent task. However, there are a few general directions common in a number of applications. We focus on three such alternatives.
- Fisher's Linear Discriminant (the two-class case)
  - Let the feature vectors live in an m-dimensional space,

$$\underline{x} \in \Re^m$$

  - The goal: Generate a feature y, as a linear combination of the components of $\underline{x}$, i.e.

$$y = w_1 x_1 + w_2 x_2 + \ldots + w_m x_m = \underline{w}^T \underline{x}$$

    so that the two classes are best separated.
  - Alternatively: Find the hyperplane $\underline{w}$, so that, after projecting $\underline{x}$ onto $\underline{w}$, we achieve maximum class separability according to a criterion.

- The criterion: Maximize $FDR = \dfrac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$

  Thus, we seek for the direction $\underline{w}$, for which
  - The mean values are as far as possible.
  - The classes are as compact as possible (small variances).
- If $\underline{\mu}_1$, $\underline{\mu}_2$ are the mean values in R^m, the respective means after projection are

$$\mu_i = \underline{w}^T \underline{\mu}_i \ , \ i = 1,2$$

or for equiprobable classes

$$(\mu_1 - \mu_2)^2 = \underline{w}^T (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{w}$$

$$\propto \underline{w}^T S_b \underline{w}$$

where S_b is the between-class scatter matrix

$$S_b = \frac{1}{2} \sum_{i=1}^{2} (\underline{\mu}_1 - \underline{\mu}_0)(\underline{\mu}_1 - \underline{\mu}_0)^T$$

$$\text{and} \quad \underline{\mu}_0 = \frac{1}{2}(\underline{\mu}_1 + \underline{\mu}_2)$$

$$\text{or} \quad S_b \propto (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T$$

- The respective variances are

$$\sigma_i^2 = E[(y-\mu_i)^2] = E[\underline{w}^T(\underline{x}-\mu_i)(\underline{x}-\mu_i)^T\underline{w}]$$
$$= \underline{w}^T\Sigma_i\underline{w}$$

where Σ_i is the respective covariance matrix .

- Finally,

$$FDR \propto \frac{\underline{w}^T S_b \underline{w}}{\underline{w}^T S_w \underline{w}}$$

where S_w is the within-class scatter matrix.

$$S_w = \frac{1}{2}(\Sigma_1 + \Sigma_2)$$

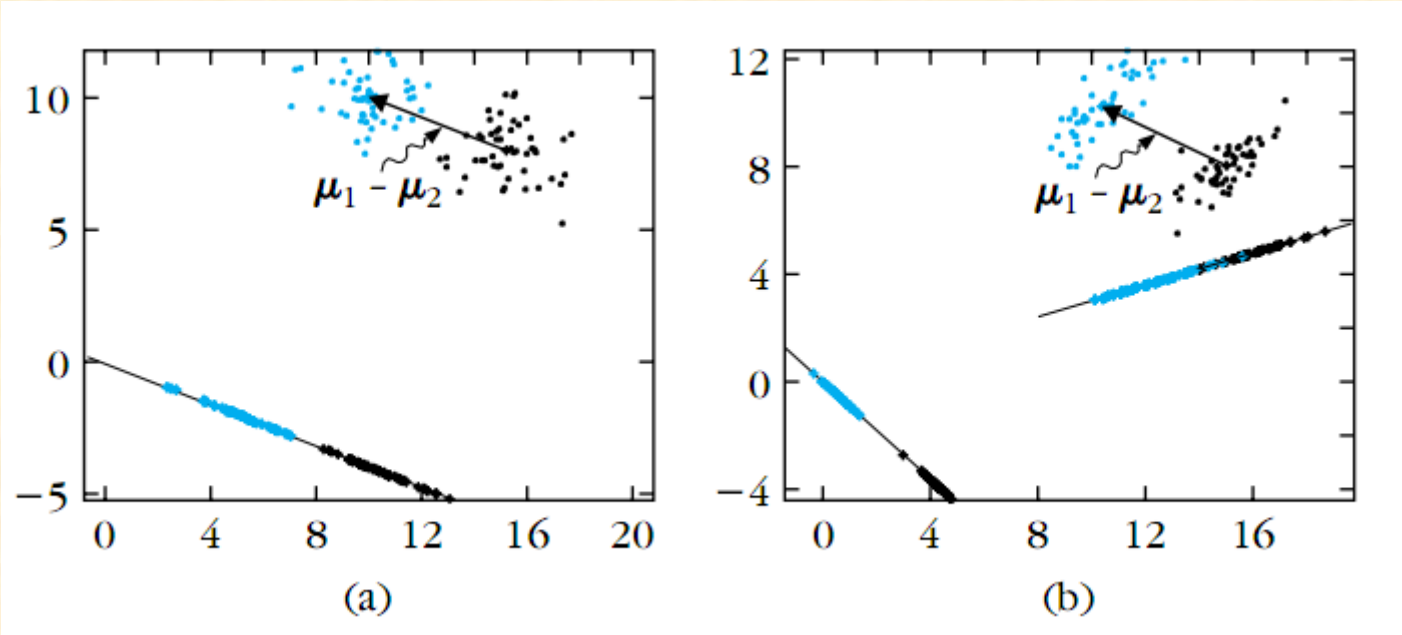
- Maximizing we get that

$$S_b \underline{w} = \lambda S_w \underline{w}$$

or $$\lambda S_w \underline{w} \propto (\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)^T \underline{w}$$

$$\propto (\underline{\mu}_1 - \underline{\mu}_2)$$

or $$\underline{w} = S_w^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$$



❖ Fisher's Linear Discriminant (the many-class case)

- The goal: Given an original set of m measurements $\underline{x} \in \Re^m$, compute $\underline{y} \in \Re^\ell$, by the linear transformation $\underline{y} = A^T \underline{x}$,

so that the J_3 scattering matrix criterion involving S_w, S_b is maximized. A^T is an $\ell x m$ matrix.

➢The basic steps in the proof:

- $J_3 = \text{trace}\{S_w^{-1} S_m\}$
- $S_{yw} = A^T S_{xw} A,\ S_{yb} = A^T S_{xb} A,$
- $J_3(A)=\text{trace}\{(A^T S_{xw} A)^{-1}\ (A^T S_{xb} A)\}$
- Compute A so that $J_3(A)$ is maximum.

➢The solution:

- Let B be the matrix that diagonalizes simultaneously matrices S_{yw}, S_{yb} , i.e:

$$B^T S_{yw} B = I\ ,\ B^T S_{yb} B = D$$

where B is a $\ell x \ell$ matrix, and D a $\ell x \ell$ diagonal matrix.

- Let $C=AB$ an $mx\ell$ matrix. If A maximizes $J_3(A)$ then

$$\left(S_{xw}^{-1}S_{xb}\right)C = CD$$

The above is an eigenvalue-eigenvector problem. For an M-class problem, $S_{xw}^{-1}S_{xb}$ is of rank M-1.

- If $\ell=M$-1, choose C to consist of the M-1 eigenvectors, corresponding to the non-zero eigenvalues.

$$\underline{y} = C^T \underline{x}$$

The above guarantees maximum J_3 value. In this case: $J_{3,x} = J_{3,y}$.

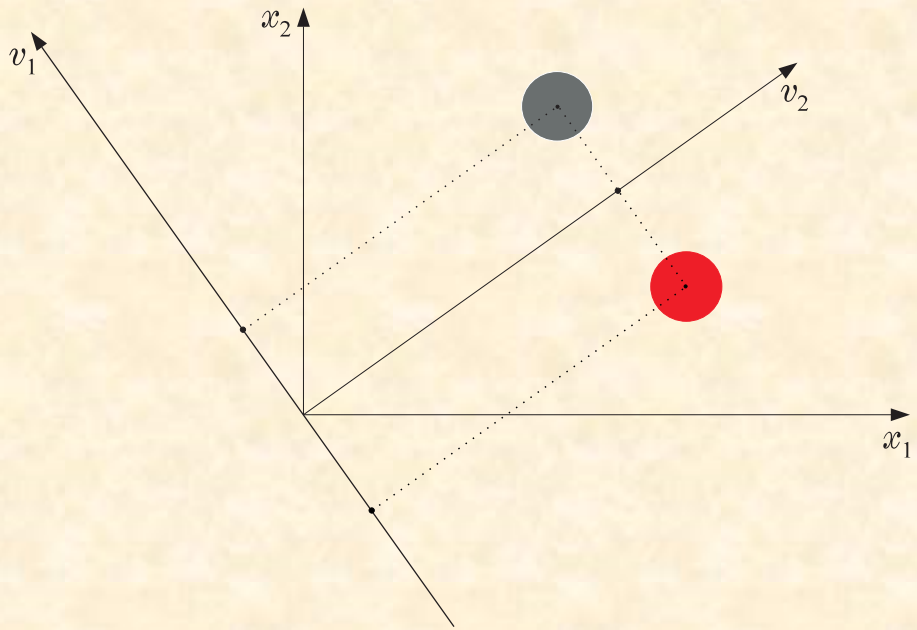
- For a two-class problem, this results to the well known Fisher's linear discriminant

$$\underline{y} = \left(\underline{\mu}_1 - \underline{\mu}_2\right)S_{xw}^{-1}\underline{x}$$

For Gaussian classes, this is the optimal Bayesian classifier, with a difference of a threshold value .

- If $\ell < M-1$, choose the ℓ eigenvectors corresponding to the ℓ largest eigenvectors.
- In this case, $J_{3,y} < J_{3,x}$, that is there is loss of information.

- Geometric interpretation. The vector $\underline{y}$ is the projection of $\underline{x}$ onto the subspace spanned by the eigenvectors of $S_{xw}^{-1} S_{xb}$.



➢ Bayesian Information Criterion (BIC)

Let N the size of the training set, $\underline{\theta}_m$ the vector of the unknown parameters of the classifier, K_m the dimensionality of $\underline{\theta}_m$, and m runs over all possible models.

- The BIC criterion chooses the model by minimizing:

$$BIC = -2L\left(\hat{\underline{\theta}}_m\right) + K_m \ln N$$

  - $L\left(\hat{\underline{\theta}}_m\right)$ is the log-likelihood computed at the ML estimate $\hat{\underline{\theta}}_m$, and it is the performance index.
  - $K_m \ln N$ is the model complexity term.

- Akaike Information Criterion:

$$AIC = -2L\left(\hat{\underline{\theta}}_m\right) + 2K_m$$