LIGO

DISCRIMINANT ANALYSIS USING MEASURES OF DATA DEPTH

PROBAL CHAUDHURI

probal@isical.ac.in

THEORETICAL STATISTICS AND MATHEMATICS UNIT

INDIAN STATISTICAL INSTITUTE, KOLKATA

LIGO

Contents

Contents of this talk are based on

- Ghosh, A. K. and Chaudhuri, P. (2005) On data depth and distribution free discriminant analysis using separating surfaces. **Bernoulli, 11, 1-27**.
- Ghosh, A. K. and Chaudhuri, P. (2005) On maximum depth and related classifiers. **Scandinavian Journal of Statistics, 32, 327-350**.

PDF versions of these articles are available at

http://www.geocities.com/ghosh_anilk

LIGO

Some examples

- **Fisher's Iris data**: Four measurements (sepal length, sepal width, petal length, petal width) are taken on three different types of Iris flower : Setosa, Virginica and Versicolor.

LIGO

Some examples

- **Fisher's Iris data**: Four measurements (sepal length, sepal width, petal length, petal width) are taken on three different types of Iris flower : Setosa, Virginica and Versicolor.
- **Crab data**: Information is available on body depth and four other measurements on carapace of two different species of rock crabs, which were marked by 'orange' and 'blue' colors. As the preserved specimens lost their colors, morphological study was needed to classify the museum materials.

LIGO

Some examples

- **Fisher's Iris data**: Four measurements (sepal length, sepal width, petal length, petal width) are taken on three different types of Iris flower : Setosa, Virginica and Versicolor.
- **Crab data**: Information is available on body depth and four other measurements on carapace of two different species of rock crabs, which were marked by 'orange' and 'blue' colors. As the preserved specimens lost their colors, morphological study was needed to classify the museum materials.
- **Biomedical data**: Four different measurements are taken on each of 209 blood samples collected from normal people as well as carriers of a genetic disorder. 15 observations, which have missing values, have been removed before the analysis.

LIGO

Some examples (contd..)

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from normal people as well as chemical diabetic and overt diabetic patients.

LIGO

Some examples (contd..)

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from normal people as well as chemical diabetic and overt diabetic patients.
- **Vowel recognition data**: A number of speakers spoke some words formed by ‘h’ followed by a vowel and then followed by ‘d’. Two lowest resonant frequencies of a speaker’s vocal tract were noted for 10 different vowels.

LIGO

Some examples (contd..)

- **Diabetes data** : Measurements are taken on fasting plasma glucose level, steady state plasma glucose level, glucose area, insulin area and relative weight from normal people as well as chemical diabetic and overt diabetic patients.
- **Vowel recognition data**: A number of speakers spoke some words formed by ‘h’ followed by a vowel and then followed by ‘d’. Two lowest resonant frequencies of a speaker’s vocal tract were noted for 10 different vowels.
- **Synthetic data (Ripley, 1994)**: Each of the two classes is an equal mixture of two bivariate normal distributions, which have the same dispersion matrix but different location parameters.

LIGO

Mathematical Formulation

- **Training data** : $(\mathbf{x}_n, c_n),\ n = 1, 2, \ldots, N$.
  Vector of measurement variables : $\mathbf{x}_n \in R^d$,
  Class labels :$c_n \in \{1, 2, \ldots, J\}$.

LIGO

Mathematical Formulation

- **Training data** : $(\mathbf{x}_n, c_n),\ n = 1, 2, \ldots, N.$
  Vector of measurement variables : $\mathbf{x}_n \in R^d$,
  Class labels :$c_n \in \{1, 2, \ldots, J\}$.
- **Decision rule** : $d(\mathbf{x})\ :\ R^d \rightarrow \{1, 2, \ldots, J.\}$

LIGO

Mathematical Formulation

- **Training data** : $(\mathbf{x}_n, c_n),\ n = 1, 2, \ldots, N.$
  Vector of measurement variables : $\mathbf{x}_n \in R^d$,
  Class labels :$c_n \in \{1, 2, \ldots, J\}.$
- **Decision rule** : $d(\mathbf{x})\ :\ R^d \rightarrow \{1, 2, \ldots, J.\}$
- **Bayes rule** : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$
  $f_j(\mathbf{x})$: density functions, π_j : prior probabilities.
  Can be used only when the class densities and prior probabilities are known.

LIGO

Mathematical Formulation

- **Training data** : $(\mathbf{x}_n, c_n),\ n = 1, 2, \ldots, N.$
  Vector of measurement variables : $\mathbf{x}_n \in R^d$,
  Class labels :$c_n \in \{1, 2, \ldots, J\}.$
- **Decision rule** : $d(\mathbf{x})\ :\ R^d \rightarrow \{1, 2, \ldots, J.\}$
- **Bayes rule** : $d_B(\mathbf{x}) = arg\max_j p(j \mid \mathbf{x}) = arg\max_j \pi_j f_j(\mathbf{x})$
  $f_j(\mathbf{x})$: density functions, π_j : prior probabilities.
  Can be used only when the class densities and prior probabilities are known.
- **Estimation of f_j's**
  - **Parametric** : LDA, QDA.
  - **Nonparametric** : Kernels, Nearest neighbors, Neural nets, Classification trees, Support vector machines.

LIGO

Data depth

Data depth measures the centrality of a multivariate observation w.r.t. a multivariate distribution.

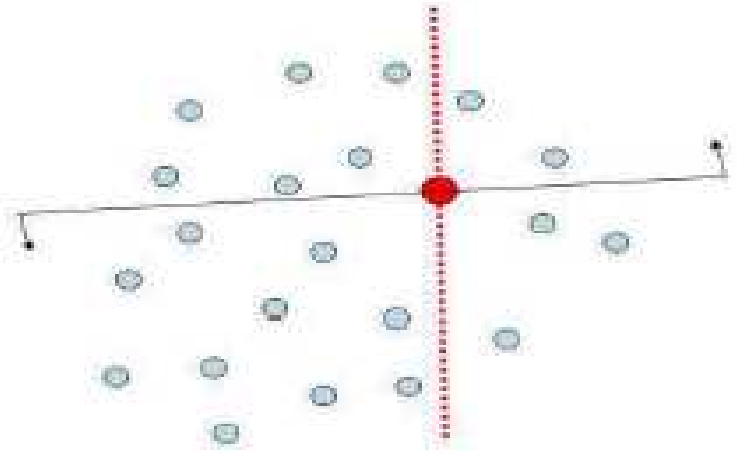
LIGO

Data depth

Data depth measures the centrality of a multivariate observation w.r.t. a multivariate distribution.

- Half-space depth (Tukey, 1975)

$$HD(F, \mathbf{x}) = \inf_{H} \left\{ P_F(H) : H \text{ is a closed half space in } R^d, \text{ and } \mathbf{x} \in H \right\}$$
$$= 1 - \sup_{\boldsymbol{\alpha}} P\{\boldsymbol{\alpha}'(\mathbf{X} - \mathbf{x}) > 0\} \text{ where } \mathbf{X} \sim F$$



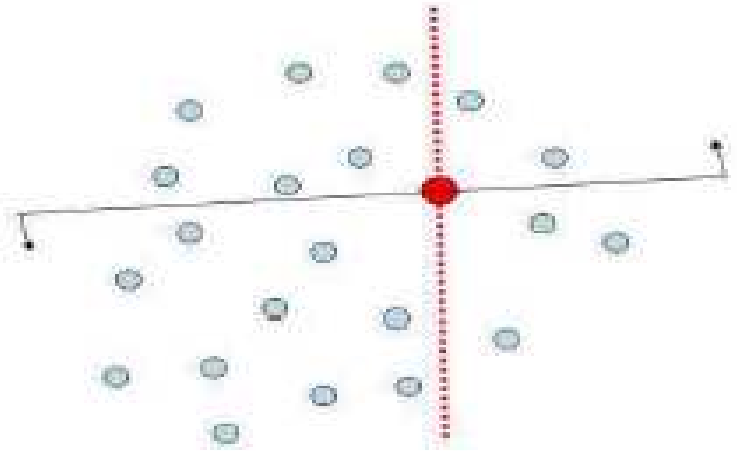
LIGO

Data depth

Data depth measures the centrality of a multivariate observation w.r.t. a multivariate distribution.

- Half-space depth (Tukey, 1975)

$$HD(F, \mathbf{x}) = \inf_{H} \left\{ P_F(H) : H \text{ is a closed half space in } R^d, \text{ and } \mathbf{x} \in H \right\}$$
$$= 1 - \sup_{\boldsymbol{\alpha}} P\{\boldsymbol{\alpha}'(\mathbf{X} - \mathbf{x}) > 0\} \text{ where } \mathbf{X} \sim F$$



How is half-space depth related to linear classification ?

LIGO Half-space depth and linear separation

- Measure of linear separation

$$\sup_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} P\{\boldsymbol{\alpha}^{'}(\mathbf{X}_1 - \mathbf{X}_2) > 0\}.$$

LIGO Half-space depth and linear separation

- Measure of linear separation

$$\sup_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} P\{\boldsymbol{\alpha}^{'}(\mathbf{X}_1 - \mathbf{X}_2) > 0\}.$$

- $\sup_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) = 1 - HD(G, \mathbf{0})$,
  where G denotes the distribution of $\mathbf{X}_1 - \mathbf{X}_2$.

LIGO Half-space depth and linear separation

- Measure of linear separation

$$\sup_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} P\{\boldsymbol{\alpha}^{'}(\mathbf{X}_1 - \mathbf{X}_2) > 0\}.$$

- $\sup_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) = 1 - HD(G, \mathbf{0})$,
  where G denotes the distribution of $\mathbf{X}_1 - \mathbf{X}_2$.

- If f_1 and f_2 are unimodal, elliptically symmetric, and $f_1(\mathbf{x}) = f_2(\mathbf{x} - \boldsymbol{\theta})$ for some location shift parameter $\boldsymbol{\theta}$,

$$arg\max_{\boldsymbol{\alpha}} U(\boldsymbol{\alpha}) \equiv arg\max_{\boldsymbol{\alpha}} \frac{\boldsymbol{\alpha}^{'}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^{'}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{'}\Sigma\boldsymbol{\alpha}}.$$

Linear classification using H-depth

- Empirical version :

$$\sup_{\boldsymbol{\alpha}} U_{n_1,n_2}(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} \Big\{ \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(\boldsymbol{\alpha}^{'} (\mathbf{x}_{1i} - \mathbf{x}_{2j}) > 0 \right) \Big\}.$$

Linear classification using H-depth

- Empirical version :

$$\sup_{\boldsymbol{\alpha}} U_{n_1,n_2}(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} \Big\{ \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(\boldsymbol{\alpha}^{'} (\mathbf{x}_{1i} - \mathbf{x}_{2j}) > 0 \right) \Big\}.$$

$1 - \sup_{\boldsymbol{\alpha}} U_{n_1,n_2}(\boldsymbol{\alpha})$ = H-depth of the origin w.r.t. the data cloud formed by the differences of observations.

Linear classification using H-depth

- Empirical version :

$$\sup_{\boldsymbol{\alpha}} U_{n_1,n_2}(\boldsymbol{\alpha}) = \sup_{\boldsymbol{\alpha}} \Big\{ \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} I\left(\boldsymbol{\alpha}^{'} (\mathbf{x}_{1i} - \mathbf{x}_{2j}) > 0 \right) \Big\}.$$

$1 - \sup_{\boldsymbol{\alpha}} U_{n_1,n_2}(\boldsymbol{\alpha})$ = H-depth of the origin w.r.t. the data cloud formed by the differences of observations.

- Estimation of discriminating surface :
  - $\boldsymbol{\alpha}$ is estimated by maximizing $U_{n_1,n_2}(\boldsymbol{\alpha})$
  - β is estimated by minimizing the training set misclassification probability of the classifier $\boldsymbol{\alpha}^{'}\mathbf{x} + \beta = 0$

LIGO

A more direct approach

Minimization of training sample misclassification probability

$$\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta) = \pi_1\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\boldsymbol{\alpha}^{'}\mathbf{x}_{1i}+\beta<0\}\right] + \pi_2\left[\frac{1}{n_2}\sum_{i=1}^{n_2} I\{\boldsymbol{\alpha}^{'}\mathbf{x}_{2i}+\beta>0\}\right]$$

LIGO

A more direct approach

Minimization of training sample misclassification probability

$$\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta) = \pi_1\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\boldsymbol{\alpha}'\mathbf{x}_{1i}+\beta<0\}\right] + \pi_2\left[\frac{1}{n_2}\sum_{i=1}^{n_2} I\{\boldsymbol{\alpha}'\mathbf{x}_{2i}+\beta>0\}\right]$$

Remark : If the observations of the two classes are completely separable, then $(\boldsymbol{\alpha}^*,\beta^*)$ is a minimizer of $\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta)$ iff $\boldsymbol{\alpha}^*$ maximizes $U_{n_1,n_2}(\boldsymbol{\alpha})$.

LIGO

A more direct approach

Minimization of training sample misclassification probability

$$\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta) = \pi_1\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\boldsymbol{\alpha}'\mathbf{x}_{1i}+\beta<0\}\right] + \pi_2\left[\frac{1}{n_2}\sum_{i=1}^{n_2} I\{\boldsymbol{\alpha}'\mathbf{x}_{2i}+\beta>0\}\right]$$

Remark : If the observations of the two classes are completely separable, then $(\boldsymbol{\alpha}^*,\beta^*)$ is a minimizer of $\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta)$ iff $\boldsymbol{\alpha}^*$ maximizes $U_{n_1,n_2}(\boldsymbol{\alpha})$.

- This classifier is closely related to the idea of regression depth (Rousseeuw and Hubert, 1999; JASA)

LIGO

A more direct approach

Minimization of training sample misclassification probability

$$\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta) = \pi_1\left[\frac{1}{n_1}\sum_{i=1}^{n_1} I\{\boldsymbol{\alpha}'\mathbf{x}_{1i}+\beta<0\}\right] + \pi_2\left[\frac{1}{n_2}\sum_{i=1}^{n_2} I\{\boldsymbol{\alpha}'\mathbf{x}_{2i}+\beta>0\}\right]$$

Remark : If the observations of the two classes are completely separable, then $(\boldsymbol{\alpha}^*,\beta^*)$ is a minimizer of $\Delta_{n_1,n_2}(\boldsymbol{\alpha},\beta)$ iff $\boldsymbol{\alpha}^*$ maximizes $U_{n_1,n_2}(\boldsymbol{\alpha})$.

- This classifier is closely related to the idea of regression depth (Rousseeuw and Hubert, 1999; JASA)
- This classification method was also discussed in Christmann and Rousseeuw (2001), Christmann, Fischer and Joachims (2002).

LIGO

Traditional linear and quadratic classifiers

- Primarily motivated by normal distribution.
- Sample moments are used to estimate the discriminating surface.
- Highly sensitive to outliers, fail in the case of heavy-tailed distributions.

LIGO

Traditional linear and quadratic classifiers

- Primarily motivated by normal distribution.
- Sample moments are used to estimate the discriminating surface.
- Highly sensitive to outliers, fail in the case of heavy-tailed distributions.

Breakdown properties of traditional methods, effect of outliers on misclassification rates and classification with robust estimates of location and scatter parameters were discussed in

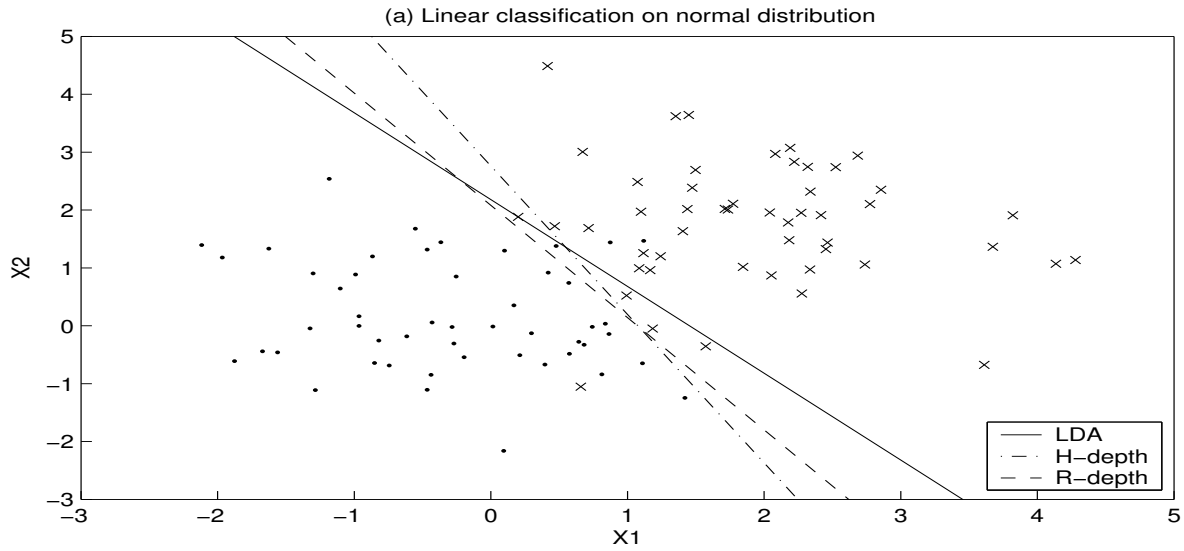
- Hawkins and McLachlan (1977,JASA)
- Randles, Brossit, Ransberg, Hogg (1978, JASA)
- Croux and Dehon (2001, Can. J. Statist.)
- Hubert and Van Driessen (2004, CSDA).

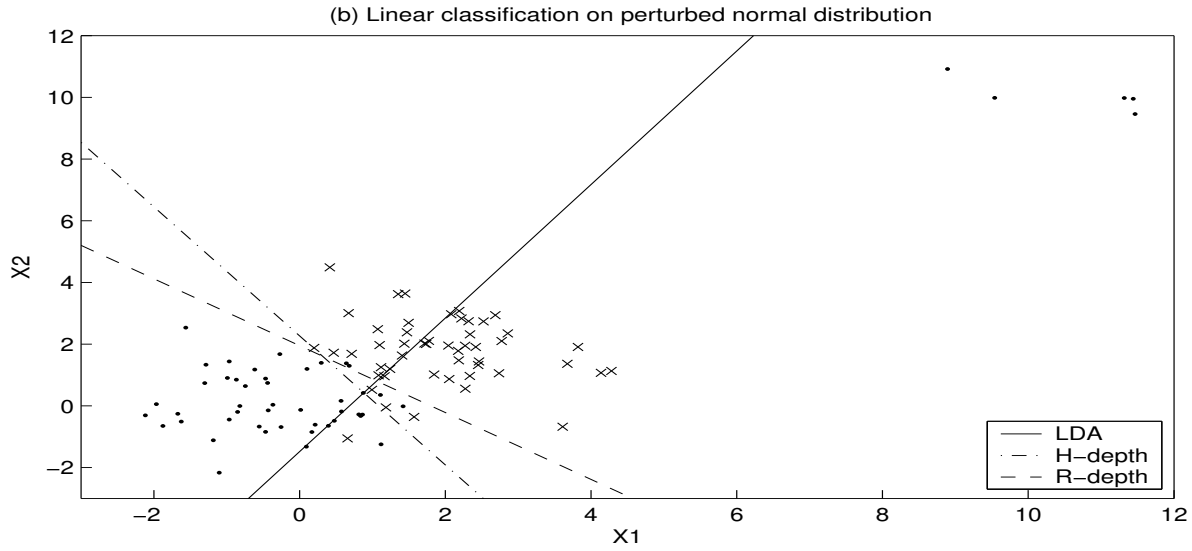
LIGO

Depth-based classifiers

- Distribution free approach.
- Distributional geometry of the data cloud is used to estimate the discriminating surface.
- Robust against outliers and heavy-tailed distributions.

Linear classification for normal and perturbed normal populations





LIGO

Computational issues

Depth computation in higher dimension

- Replace the indicator $I\{x > 0\}$ by $1/(1 + e^{-tx})$ with large t. It allows the use of derivatives to find out the direction of steepest descent/ascent and thereby makes the algorithm computationally efficient

LIGO

Computational issues

Depth computation in higher dimension

- Replace the indicator $I\{x > 0\}$ by $1/(1 + e^{-tx})$ with large t. It allows the use of derivatives to find out the direction of steepest descent/ascent and thereby makes the algorithm computationally efficient
- As an alternative to this, for depth computation one may also use
  - Algorithm of Rousseeuw and Struyf (1998)
  - Probabilistic search algorithms using annealing or genetic algorithms (Chakraborty and Chaudhuri, 2003, 2004).

LIGO Results Based on Simulation Studies

$$\boldsymbol{\mu}_1 = (0,0), \quad \boldsymbol{\mu}_2 = (1,1), \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}$$

	n	LDA	H-depth		R-depth	
			Exact	Approx.	Exact	Approx.
normal	30	24.80 (0.15)	25.85 (0.20)	25.79 (0.19)	26.11 (0.24)	26.06 (0.22)
	50	24.40 (0.10)	25.21 (0.14)	25.19 (0.15)	25.44 (0.15)	25.42 (0.13)
	100	24.21 (0.10)	24.80 (0.10)	24.72 (0.13)	25.11 (0.12)	24.88 (0.13)
cauchy	30	42.54 (0.91)	33.20 (0.29)	33.20 (0.28)	32.75 (0.25)	32.61 (0.24)
	50	43.81 (0.95)	32.45 (0.26)	32.51 (0.24)	32.45 (0.25)	32.50 (0.27)
	100	41.95 (0.98)	31.78 (0.15)	31.80 (0.15)	31.77 (0.15)	31.59 (0.14)
perturb	30	50.07 (0.41)	29.97 (0.29)	30.05 (0.29)	30.21 (0.26)	29.94 (0.21)
normal	50	50.75 (0.53)	29.15 (0.15)	28.96 (0.15)	29.21 (0.16)	29.20 (0.16)
	100	50.28 (0.53)	28.55 (0.12)	28.65 (0.13)	28.86 (0.13)	28.70 (0.12)

Bayes risk : 23.98, 30.40 and 22.71

Nonlinear classification & classification between multiple populations

Nonlinear classification :

- Project the observations into vector space of nonlinear functions
- Perform linear classification on that projected space.

$\mathbf{X} = (X_1, X_2) \quad \rightarrow \quad \mathbf{Z} = (X_1,\ X_2,\ X_1^2,\ X_2^2,\ X_1X_2)$

Linear discrimination $\rightarrow$ Quadratic discrimination

LIGO Nonlinear classification & classification between multiple populations

Nonlinear classification :

- Project the observations into vector space of nonlinear functions
- Perform linear classification on that projected space.

$$\mathbf{X} = (X_1, X_2) \quad \rightarrow \quad \mathbf{Z} = (X_1,\ X_2,\ X_1^2,\ X_2^2,\ X_1X_2)$$

Linear discrimination $\rightarrow$ Quadratic discrimination

Multi-class problems :

- Perform pairwise classification taking each pair of classes.
- Combine the results using voting (Friedman, 1996) or coupling (Hastie and Tibshirani, 1998).

LIGO

Large sample properties of misclassification rates

- Under appropriate conditions, average misclassification rate of the RD-based linear (or nonlinear) classifier asymptotically converges to the best possible average misclassification rate that can be obtained using a linear (or nonlinear) classifier as sample sizes tend to infinity.

LIGO

Large sample properties of misclassification rates

- Under appropriate conditions, average misclassification rate of the RD-based linear (or nonlinear) classifier asymptotically converges to the best possible average misclassification rate that can be obtained using a linear (or nonlinear) classifier as sample sizes tend to infinity.
- Under location shift and elliptic symmetry, the average misclassification rate (with equal priors) for the RD-based linear classifier converges to the optimal Bayes error. In that case, if the Bayes classifier is unique, and $U(\boldsymbol{\alpha})$ has a unique maximizer, the same holds for the HD-based linear classifier.

LIGO

Large sample properties of misclassification rates

- Under appropriate conditions, average misclassification rate of the RD-based linear (or nonlinear) classifier asymptotically converges to the best possible average misclassification rate that can be obtained using a linear (or nonlinear) classifier as sample sizes tend to infinity.
- Under location shift and elliptic symmetry, the average misclassification rate (with equal priors) for the RD-based linear classifier converges to the optimal Bayes error. In that case, if the Bayes classifier is unique, and $U(\boldsymbol{\alpha})$ has a unique maximizer, the same holds for the HD-based linear classifier.
- The above convergence results remain valid even if we do not restrict ourselves to only finite dimensional parametric surfaces but the number of basis functions used in a nonlinear classifier grows at an appropriate rate with the sample size.

LIGO

Results on linear classification

Data sets	LDA	H-depth	R-depth
Vowel	25.26	20.72	19.83
Diabetes	11.12 (0.07)	5.49 (0.06)	6.12 (0.06)
Bio-medical	15.96 (0.07)	10.87 (0.07)	11.03 (0.07)
Synthetic	10.80	10.70	11.50
Crab	5.20 (0.06)	4.85 (0.06)	4.47 (0.06)
Iris	2.18 (0.07)	3.92 (0.10)	3.56 (0.10)

Results on quadratic classification

Data sets	QDA	H-depth	R-depth
Vowel	19.83	19.22	19.53
Diabetes	9.32 (0.06)	6.57 (0.06)	7.09 (0.06)
Bio-medical	12.68 (0.06)	11.61 (0.07)	11.76 (0.06)
Synthetic	10.20	11.00	10.70
Crab	5.89 (0.06)	4.37 (0.06)	4.26 (0.06)
Iris	2.75 (0.09)	3.99 (0.11)	3.43 (0.10)

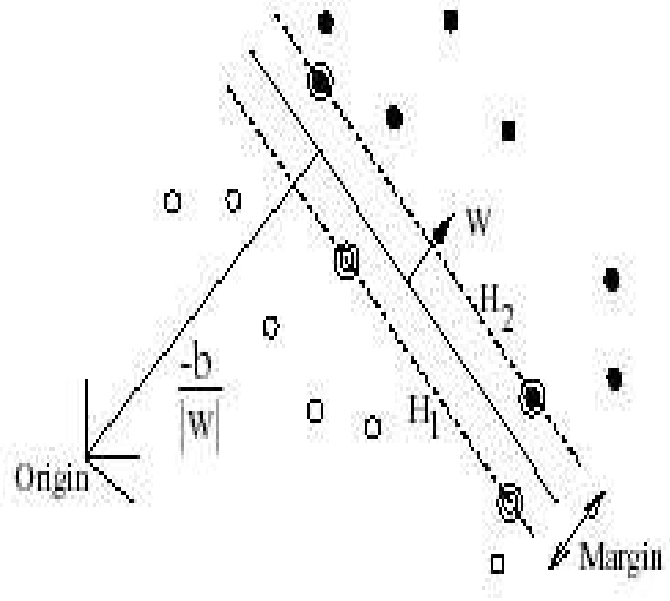
LIGO

Support vector machines (SVM)

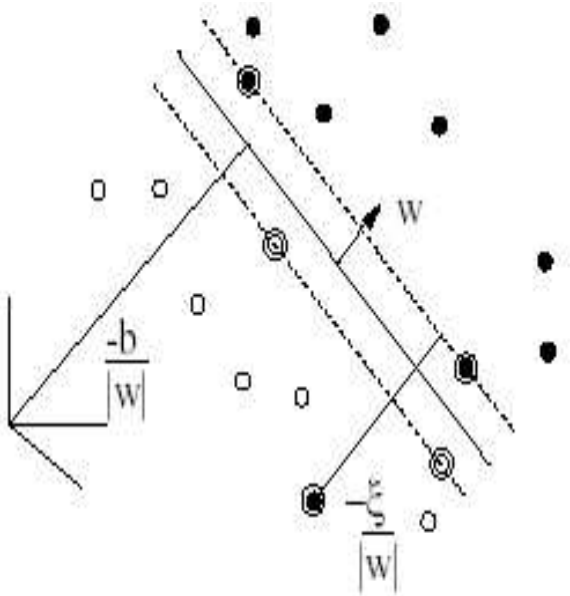
SVM maximizes the distance between two parallel separating hyperplanes $\mathbf{w}'\mathbf{x} + b = 1$ and $\mathbf{w}'\mathbf{x} + b = -1$.

$$\text{Minimize } (\text{w.r.t. } \mathbf{w}, b, \xi) \quad \tfrac{1}{2}\|\mathbf{w}\|^2$$

$$\text{subject to} \quad (\mathbf{w}'\mathbf{x}_i + b - 1)y_i \geq 0 \ \text{ for all } \mathbf{x}_i.$$



Non-separable case



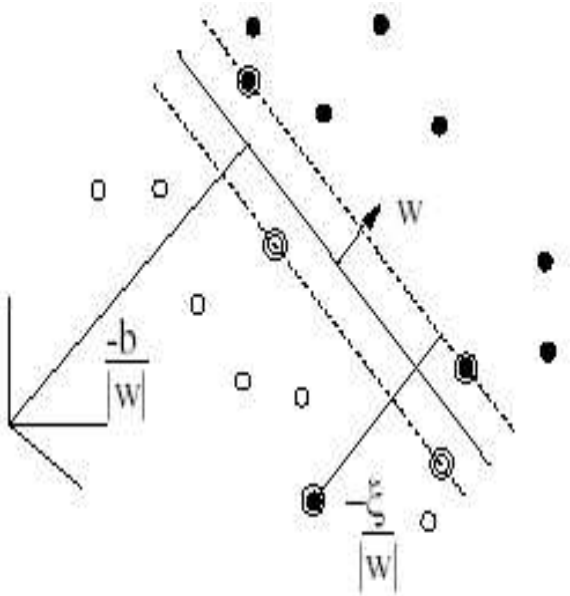
Minimize (w.r.t. $\mathbf{w}, b, \xi$) $\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$

subject to $\mathbf{w}'\mathbf{x}_i + b \geq 1 - \xi_i$ for $y_i = 1$

$\mathbf{w}'\mathbf{x}_i + b \leq -1 + \xi_i$ for $y_i = -1$

$\xi_i \geq 0 \ \ \forall i$

Non-separable case



Minimize (w.r.t. $\mathbf{w}, b, \xi$) $\quad \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \xi_i$

subject to $\quad \mathbf{w}'\mathbf{x}_i + b \geq 1 - \xi_i \quad$ for $y_i = 1$

$\mathbf{w}'\mathbf{x}_i + b \leq -1 + \xi_i \quad$ for $y_i = -1$

$\xi_i \geq 0 \ \ \forall i$

- **Nonlinear classification** : Project the observations into a higher dimensional space and perform linear classification.

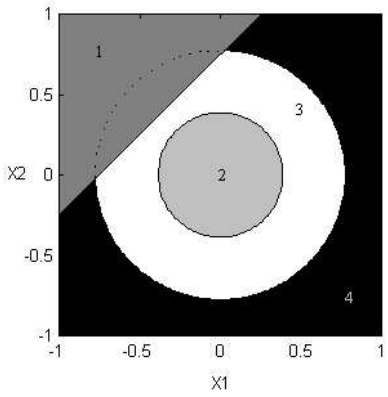
LIGO Simulated example (Christmann, 2002)

Class-1 if $x_2 - x_1 > 0.758$

Class-2 if $x_1^2 + x_2^2 \leq 0.15$

Class-3 if $0.15 < x_1^2 + x_2^2 \leq 0.75$

Class-4 otherwise



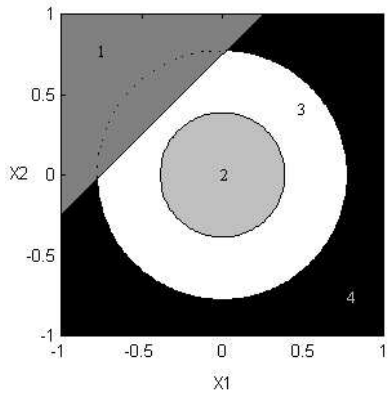
LIGOSimulated example (Christmann, 2002)

Class-1 if $x_2 - x_1 > 0.758$

Class-2 if $x_1^2 + x_2^2 \leq 0.15$

Class-3 if $0.15 < x_1^2 + x_2^2 \leq 0.75$

Class-4 otherwise



- Number of simulations :250

  Size of training samples :300 Size of test samples : 700

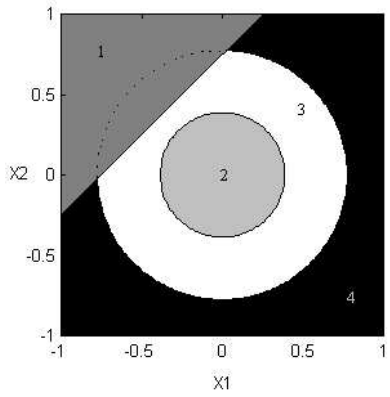
LIGO Simulated example (Christmann, 2002)

Class-1 if $x_2 - x_1 > 0.758$

Class-2 if $x_1^2 + x_2^2 \leq 0.15$

Class-3 if $0.15 < x_1^2 + x_2^2 \leq 0.75$

Class-4 otherwise



- Number of simulations :250
  Size of training samples :300 Size of test samples : 700
- Misclassification rates :
  Support Vector Machine : 36% Quadratic Discrim. Anal. : 20.9%
  H-Depth Classifier : 1.58% R-Depth Classifier : 2.81%

LIGO

Nonparametric classification using data depth

Maximum depth classification :

$$\mathrm{d}(\mathbf{x}) = j_o \;\text{ where } D_{n_{j_o}}(j_o, \mathbf{x}) \geq D_{n_j}(j, \mathbf{x}) \;\; \forall j \neq j_o,$$

where n_j = training sample size of j^{th} population

$D_{n_j}(j, \mathbf{x})$ = empirical depth of $\mathbf{x}$ w.r.t. j^{th} population

LIGO

Nonparametric classification using data depth

Maximum depth classification :

$$\mathrm{d}(\mathbf{x}) = j_o \;\text{ where } D_{n_{j_o}}(j_o, \mathbf{x}) \geq D_{n_j}(j, \mathbf{x}) \;\; \forall j \neq j_o,$$

where n_j = training sample size of j^{th} population

$D_{n_j}(j, \mathbf{x})$ = empirical depth of **x** w.r.t. j^{th} population

- Classifiers are not restricted to any specific form of discriminating surface
- Does not require any pairwise treatment for multi-class classification problems.

LIGO

Classification using depth

Mahalanobis depth : $MD(F, \mathbf{x}) = \{1 + (\mathbf{x} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)\}^{-1}$

Use of MD (Mahalanobis, 1936) with common estimate for dispersion matrix leads to linear classification

LIGO

Classification using depth

Mahalanobis depth : $MD(F, \mathbf{x}) = \{1 + (\mathbf{x} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)\}^{-1}$
Use of MD (Mahalanobis, 1936) with common estimate for dispersion matrix leads to linear classification

- Moment based estimates for location parameters and scatter matrix lead to Fisher's LDA, which is not robust.

LIGO

Classification using depth

Mahalanobis depth : $MD(F, \mathbf{x}) = \{1 + (\mathbf{x} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)\}^{-1}$
Use of MD (Mahalanobis, 1936) with common estimate for dispersion matrix leads to linear classification

- Moment based estimates for location parameters and scatter matrix lead to Fisher's LDA, which is not robust.
- For robust discriminant analysis, one needs to plug-in robust estimates of location parameters and scatter matrix.
  - Hawkins and McLachlan (1977,JASA)
  - Randles, Brossit, Ransberg, Hogg (1978, JASA)
  - Croux and Dehon (2001, Can. J. Statist.)
  - Hubert and Van Driessen (2004, CSDA).

LIGO

Classification using depth

Mahalanobis depth : $MD(F, \mathbf{x}) = \{1 + (\mathbf{x} - \boldsymbol{\mu}_F)' \boldsymbol{\Sigma}_F^{-1} (\mathbf{x} - \boldsymbol{\mu}_F)\}^{-1}$
Use of MD (Mahalanobis, 1936) with common estimate for dispersion matrix leads to linear classification

- Moment based estimates for location parameters and scatter matrix lead to Fisher's LDA, which is not robust.
- For robust discriminant analysis, one needs to plug-in robust estimates of location parameters and scatter matrix.
  - Hawkins and McLachlan (1977,JASA)
  - Randles, Brossit, Ransberg, Hogg (1978, JASA)
  - Croux and Dehon (2001, Can. J. Statist.)
  - Hubert and Van Driessen (2004, CSDA).
- Use of other depth functions in classification and clustering
  - Hoberg (2000) used zonoid depth (Mosler, 2002).
  - Jornsten (2004) used spatial depth (Vardi and Zhang, 2000).

LIGO Half-space depth and simplicial depth

- Half-space depth (Tukey, 1975)

$$HD(F, \mathbf{x}) = \inf_{H} \{P_F(H) : H \text{ is a closed half space in } R^d, \text{and } \mathbf{x} \in H\}$$

LIGO Half-space depth and simplicial depth

- Half-space depth (Tukey, 1975)

$$HD(F, \mathbf{x}) = \inf_{H} \{P_F(H) : H \text{ is a closed half space in } R^d, \text{and } \mathbf{x} \in H\}$$

- Simplicial depth (Liu, 1990)

$$SD(F, \mathbf{x}) = P_F\{\mathbf{x} \in S(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1})\},$$

where $S(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1})$ is a d-dimensional simplex formed by $(d+1)$ i.i.d. observations $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{d+1}$ from F

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

- Based on the idea of geometric quantiles (Chaudhuri, 1996; Koltchinskii, 1997).

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

- Based on the idea of geometric quantiles (Chaudhuri, 1996; Koltchinskii, 1997).
- For spherical distribution, it is a decreasing function of Euclidean distance.

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

- Based on the idea of geometric quantiles (Chaudhuri, 1996; Koltchinskii, 1997).
- For spherical distribution, it is a decreasing function of Euclidean distance.
- Computationally efficient than other depth functions.

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

- Based on the idea of geometric quantiles (Chaudhuri, 1996; Koltchinskii, 1997).
- For spherical distribution, it is a decreasing function of Euclidean distance.
- Computationally efficient than other depth functions.
- Invariant under rotation and homogeneous scale transformation. Standardization of co-ordinate variables is needed by proper re-scaling.

LIGO

Spatial depth (SPD) (Vardi and Zhang, 2000; Serfling, 2002)

$$SPD(F, \mathbf{x}) = 1 - \|E_F\{(\mathbf{x} - \mathbf{X})/\|\mathbf{x} - \mathbf{X}\|\}\|$$

- Based on the idea of geometric quantiles (Chaudhuri, 1996; Koltchinskii, 1997).
- For spherical distribution, it is a decreasing function of Euclidean distance.
- Computationally efficient than other depth functions.
- Invariant under rotation and homogeneous scale transformation. Standardization of co-ordinate variables is needed by proper re-scaling.
- Unlike HD and SD, it does not have the problem of data points having zero depths, and ties do not occur in practice.

LIGO

Asymptotic behavior of error rates

- If $f_1, f_2, \ldots, f_J$ are elliptically symmetric and $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for some g with $g(k\mathbf{x}) \leq g(\mathbf{x})$ for every $\mathbf{x}$ and $k > 1$, in equal prior cases, error rates for half-space depth, simplicial depth, majority depth (Singh, 1991) and projection depth (Stahel, 1981; Donoho, 1982) classifiers converge to the optimal Bayes risk as $\min\{n_1, n_2, \ldots, n_J\} \to \infty$.

LIGO

Asymptotic behavior of error rates

- If $f_1, f_2, \ldots, f_J$ are elliptically symmetric and $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for some g with $g(k\mathbf{x}) \leq g(\mathbf{x})$ for every $\mathbf{x}$ and $k > 1$, in equal prior cases, error rates for half-space depth, simplicial depth, majority depth (Singh, 1991) and projection depth (Stahel, 1981; Donoho, 1982) classifiers converge to the optimal Bayes risk as $\min\{n_1, n_2, \ldots, n_J\} \to \infty$.
- If g is spherical, in equal prior cases, error rate of spatial depth classifier converges to the optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.

LIGO

Asymptotic behavior of error rates

- If $f_1, f_2, \ldots, f_J$ are elliptically symmetric and $f_j(\mathbf{x}) = g(\mathbf{x} - \boldsymbol{\mu}_j)$ for some g with $g(k\mathbf{x}) \leq g(\mathbf{x})$ for every $\mathbf{x}$ and $k > 1$, in equal prior cases, error rates for half-space depth, simplicial depth, majority depth (Singh, 1991) and projection depth (Stahel, 1981; Donoho, 1982) classifiers converge to the optimal Bayes risk as $\min\{n_1, n_2, \ldots, n_J\} \to \infty$.
- If g is spherical, in equal prior cases, error rate of spatial depth classifier converges to the optimal Bayes risk as $min\{n_1, n_2, \ldots, n_J\} \to \infty$.

Follows from the results on convergence of empirical depth contours to population depth contours (Liu, 1990; Nolan, 1992; Donoho and Gasko, 1992; Koltchinskii, 1997; He and Wang, 1999; Zuo and Serfling, 2000, Serfling, 2002)

Results on simulated data sets

$$\boldsymbol{\mu}_1 = (0,0), \quad \boldsymbol{\mu}_2 = (2,2), \quad \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_0 = \begin{bmatrix} 1 & 1 \\ 1 & 4 \end{bmatrix}$$

	$\boldsymbol{\Sigma}$	n	LDA	QDA	SPD	HD	SD
Normal	$\mathbf{I}$	100	8.05(0.09)	8.07(0.09)	8.21(0.09)	8.44(0.10)	8.64(0.11)
		200	7.92(0.09)	7.95(0.08)	8.01(0.09)	8.16(0.09)	8.25(0.09)
	$\boldsymbol{\Sigma}_0$	100	16.07(0.11)	16.14(0.11)	16.92(0.11)	16.86(0.12)	16.88(0.12)
		200	15.99(0.11)	16.04(0.10)	16.89(0.12)	16.30(0.11)	16.40(0.12
Cauchy	$\mathbf{I}$	100	33.05(1.29)	47.81(0.54)	21.84(0.20)	22.65(0.23)	22.75(0.23)
		200	34.42(1.41)	49.37(0.18)	21.05(0.16)	21.91(0.20)	22.00(0.19)
	$\boldsymbol{\Sigma}_0$	100	40.83(1.19)	49.41(0.14)	27.78(0.23)	28.64(0.27)	28.80(0.28)
		200	38.65(1.18)	49.63(0.14)	26.79(0.19)	27.36(0.21)	27.51(0.19)

Bayes risk : 7.87, 15.86, 19.58 and 25.01.

LIGO

Results on benchmark data sets

	Synthetic	Vowel	Salmon
Lin. Disc. Anal.	10.8	25.2	7.54 (0.32)
Quad. Disc. Anal.	10.2	19.8	7.23 (0.32)
Half-space depth	12.8	23.7	7.32 (0.34)
Simplicial depth	13.8	32.7	8.76 (0.41)
Spatial depth	10.5	24.6	7.46 (0.33)
(standardized)	10.5	21.3	7.42 (0.34)
Nearest neighbor	8.7	21.9	8.02 (0.36)
(standardized)	11.7	17.7	8.11 (0.37)