Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data

Addy M. Bolivar Cime

CIMAT

Joint work with
J. S. Marron [1]

[1] UNC at Chapel Hill

Content

Content

Content

Content

Content

Content

HDLSS context

- Data with High-Dimension and Low Sample Size (HDLSS) arise in many fields.
- There is an increasing current interest in the statistical analysis of this kind of data.
- Recently several authors have been interested in the comparison of classical methods for Binary Discrimination Analysis with new ones designed for the HDLSS context.

Geometric representation

- Let $z(d) = (z^{(1)}, \ldots, z^{(d)})^\top$ be a d-dimensional random vector drawn from the multivariate standard normal distribution.
- By Hall, Marron and Neeman (2005) the random vector z lie near the surface of an expanding sphere:

$$\begin{aligned} \| z \| &= \left(\sum_{k=1}^{d} z^{(k)2} \right)^{1/2} \\ &= d^{1/2} + O_p(1), \qquad \text{as } d \to \infty. \end{aligned}$$

If z_1 and z_2 are independent then

$$\| z_1 - z_2 \| = (2d)^{1/2} + O_p(1), \qquad \text{as } d \to \infty,$$

and

$$\text{Angle}(z_1, z_2) = \frac{\pi}{2} + O_p(d^{-1/2}), \qquad \text{as } d \to \infty.$$

Case $d = 3$ and $n = 3$

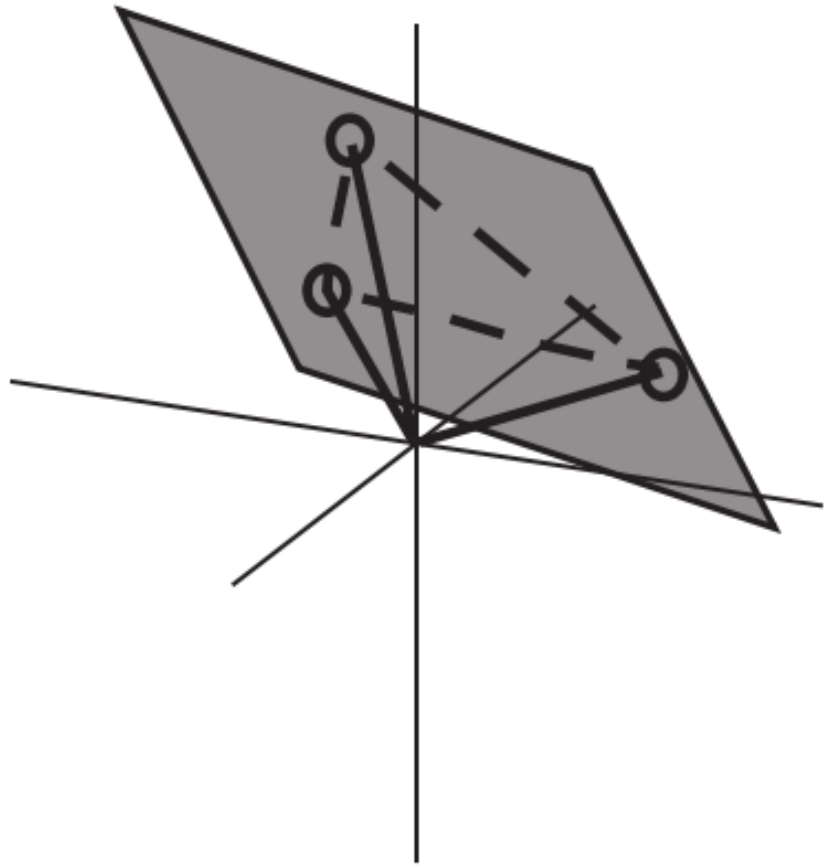


Figure: 3-simplex

Binary discrimination analysis

- Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$.

Binary discrimination analysis

- Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

  where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$.
- We have two classes of data, the classes C_+ and C_- corresponding to the vectors with $w_i = 1$ and $w_i = -1$, respectively.

Binary discrimination analysis

- Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

  where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$.
- We have two classes of data, the classes C_+ and C_- corresponding to the vectors with $w_i = 1$ and $w_i = -1$, respectively.
- Let m and n be the cardinality of C_+ and C_-, respectively.

Binary discrimination analysis

- Suppose that we have the following training data set

$$(x_1, w_1), (x_2, w_2), \ldots, (x_N, w_N), \tag{1}$$

  where $x_i \in \mathbb{R}^d$ and $w_i \in \{-1, 1\}$, for $i = 1, 2, \ldots, N$.
- We have two classes of data, the classes C_+ and C_- corresponding to the vectors with $w_i = 1$ and $w_i = -1$, respectively.
- Let m and n be the cardinality of C_+ and C_-, respectively.
- We want to assign a new data point x_0 to one of these classes.

Separable case

- We say that the training data set is linearly separable if there exists a hyperplane for which all the data of the class C_+ are on one side of the hyperplane and all the data of the class C_- are on the other side.
- A hyperplane with such property is called a separating hyperplane of the training data set.

Binary discrimination methods

We consider the following methods based on separating hyperplanes:

- Mean Difference (MD)
- Support Vector Machine (SVM)
- Distance Weighted Discrimination (DWD)
- Maximal Data Piling (MDP)
- Naive Bayes (NB)

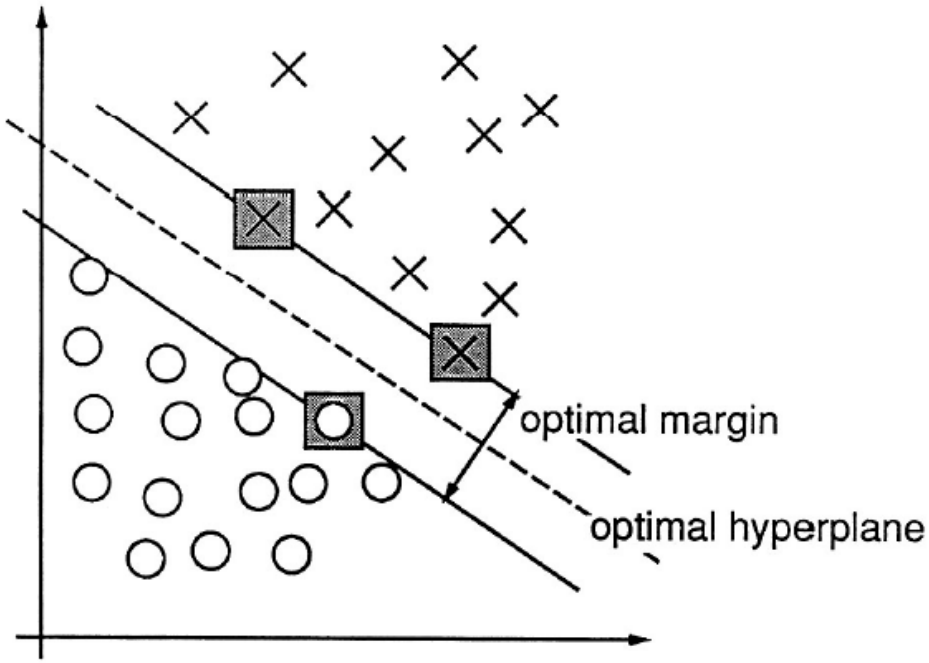
The DWD and MDP methods are designed for the HDLSS context.

- The MD, SVM and DWD methods have been previously studied in Hall, Marron and Neeman (2005), where the **probability of misclassification** of a new data point is considered when the dimension d tends to infinity.
- Marron, Todd and Ahn (2007) observed by simulation that the MD, SVM and DWD methods have asymptotically the same behavior, in terms of **proportion of misclassification** of new data.
- In the present work (Bolivar-Cime, Marron (2011)) we take a different asymptotic viewpoint, based on the **angle between the normal vectors** of the separating hyperplanes.

Support Vector Machine (SVM)

- Vapnik (1982, 1995)
- Cortes and Vapnik (1995)
- Burges (1998)
- Cristianini and Shawe-Taylor (2000)
- Izenman (2008)

Let d_+ and d_- be the shortest distances from the separating hyperplane to the nearest vector in C_+ and C_-, respectively. The margin of the separating hyperplane is defined as $d_0 = d_+ + d_-$.



The vectors closer to the separating hyperplane are called support vectors.

- In the separable case the SVM hyperplane

$$v_0^T x + b_0 = 0,$$

is the unique separating hyperplane with a maximal margin.

- In the separable case the SVM hyperplane

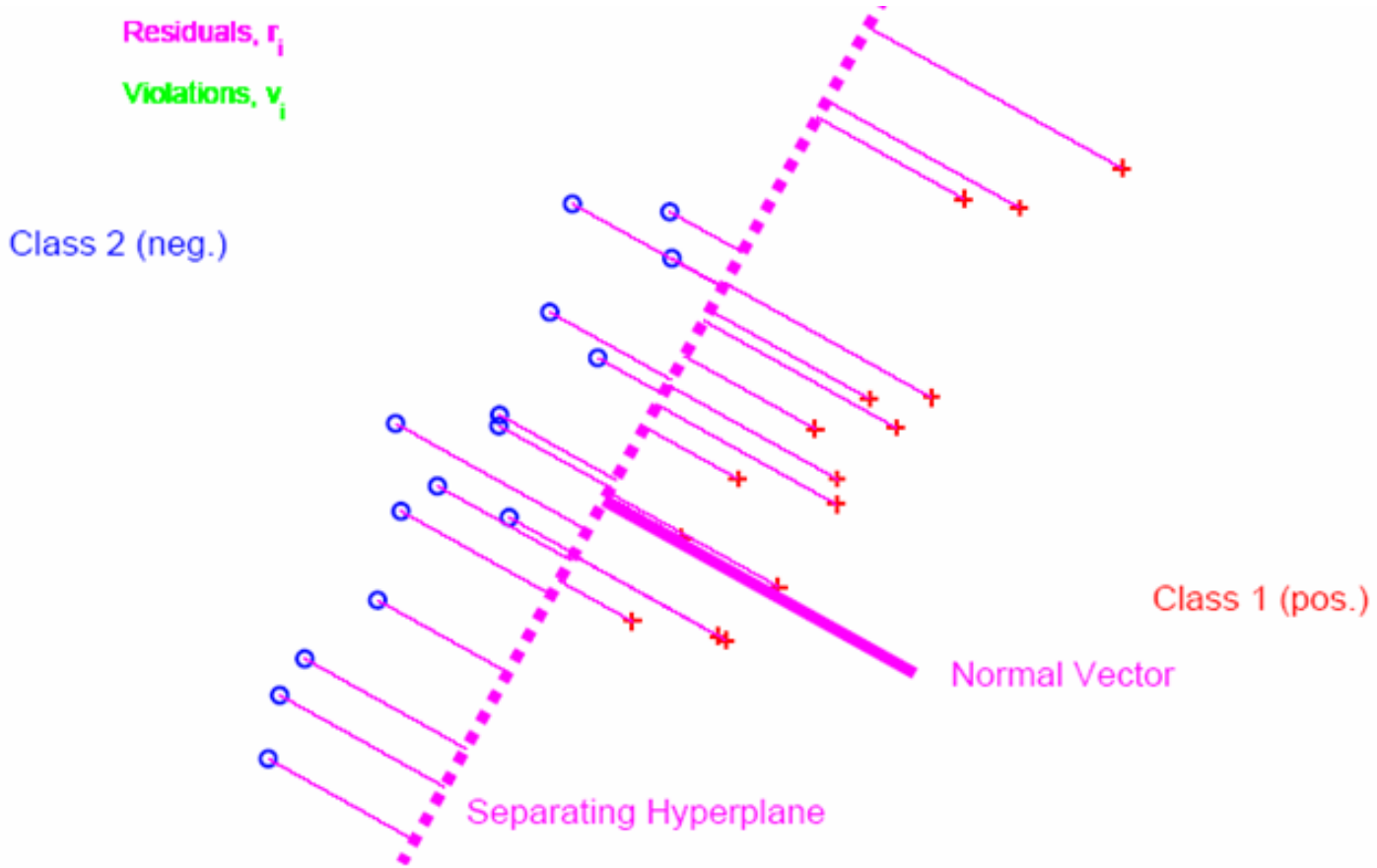
$$v_0^T x + b_0 = 0,$$

  is the unique separating hyperplane with a maximal margin.
- v_0 is a linear function only of the support vectors.

Distance Weighted Discrimination (DWD)

- In the HDLSS situation Marron, Todd and Ahn (2007) observe that the projection of the data onto the normal vector of the SVM separating hyperplane produces substantial data piling.
- They show that data piling may affect the generalization performance.
- They propose the DWD method, which avoids the data piling problem and improves generalizability.
- All the training data have a role in finding the DWD hyperplane, but data closer to the hyperplane have a bigger impact than data that are farther away.

- Define the residual of the i-th data vector as the distance of this vector to the separating hyperplane.



The DWD hyperplane

$$v_1^\top x + b_1 = 0,$$

solves the optimization problem

$$\begin{aligned} &\textit{minimize} \quad \sum_{i=1}^{N} \frac{1}{r_i} \\ &\textit{subject to} \quad \| v \| = 1, \; r_i \geq 0, \; i = 1, 2, \ldots N. \end{aligned} \tag{2}$$

Mean Difference (MD)

- Also called the nearest centroid method (Scholkopf and Smola (2002)).
- The separating hyperplane of this method is the one that orthogonally bisects the segment joining the centroids or means of the two classes.
- That is, if the means of the classes C_+ and C_- are given by

$$\overline{x}_+ = \frac{1}{m} \sum_{x_i \in C_+} x_i \quad \text{and} \quad \overline{x}_- = \frac{1}{n} \sum_{x_i \in C_-} x_i, \tag{3}$$

respectively, then the MD hyperplane has normal vector

$$u = \overline{x}_+ - \overline{x}_- \tag{4}$$

and bisects the segment joining the means $\overline{x}_+$ and $\overline{x}_-$.

Maximal Data Piling (MDP)

- This method was proposed by Ahn and Marron (2010).
- It was specially designed for the HDLSS context and we need to assume $d \geq N - 1$ and that the subspace generated by the data set has dimension $N - 1$.
- There exist direction vectors onto which the projection of the training data are piled completely at two distinct points, one for each class.

- For the MDP hyperplane

$$v_2^\top x + b_2 = 0,$$

the unit normal vector v_2 is the direction vector for which the projections of the two class means have maximal distance, and the projection of each data point onto the vector is the same as its class mean.

- For the MDP hyperplane

$$v_2^\top x + b_2 = 0,$$

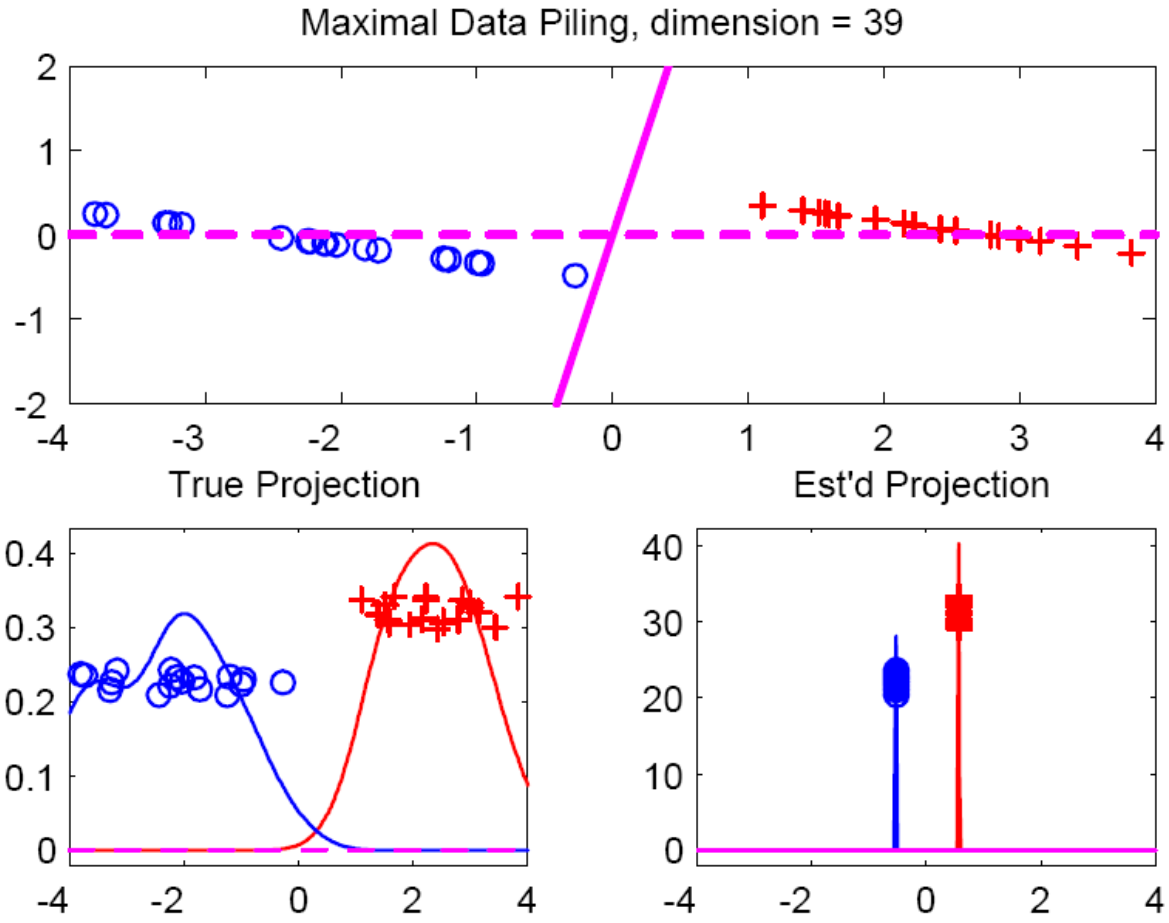
the unit normal vector v_2 is the direction vector for which the projections of the two class means have maximal distance, and the projection of each data point onto the vector is the same as its class mean.

- The bias b_2 can be calculated as

$$b_2 = -v_2^\top (m\overline{x}_+ + n\overline{x}_-)/N. \tag{5}$$

Example

$d = 39$, $n = m = 20$. The data of C_- and C_+ were drawn from multivariate normal distributions with identity covariance matrix and means $\mu_- = (-2.2, 0, \ldots, 0)^\top$ and $\mu_+ = (+2.2, 0, \ldots, 0)^\top$, respectively.



- The MDP method is equivalent to the Fisher Linear Discrimination (FLD) in the non-HDLSS situation.
- Bickel and Levina (2004) have demonstrated that FLD has very poor HDLSS properties.
- Ahn and Marron (2010) showed that, although data piling may not be desirable, the MDP method can work very well and better than FLD in some settings in the HDLSS context.

Naive Bayes (NB)

- It uses the Bayes Rule to obtain the normal vector of the separating hyperplane (Bickel and Levina (2004)).
- This method assumes that the common covariance matrix Σ of the two classes is diagonal, i.e. the entries of the random vectors are uncorrelated.

- Suppose that the classes have normal distributions $N_d(\mu_0, \Sigma)$ and $N_d(\mu_1, \Sigma)$.
- Let u be the MD normal vector and let $D = \text{diag}(\widehat{\Sigma})$ be the diagonal matrix whose entries are the diagonal elements of the pooled covariance matrix

$$\widehat{\Sigma} = \frac{1}{m+n-2}\left[\sum_{x_i \in C_+}(x_i - \overline{x}_+)^2 + \sum_{x_i \in C_-}(x_i - \overline{x}_+)^2\right]. \quad (6)$$

- By Bickel and Levina (2004) the normal vector of the NB hyperplane is given by

$$v_3 = D^{-1}u. \quad (7)$$

The bias of the NB hyperplane can be calculated as

$$b_3 = -v_3^\top \widehat{\mu}, \tag{8}$$

where $\widehat{\mu} = (\overline{x}_+ + \overline{x}_-)/2$.

Assumptions

The random vectors in C_+ and C_- are independent with d-multivariate normal distributions $N_d(v_d, I_d)$ and $N_d(0, I_d)$, respectively.

- The difference between these classes is determined by the mean vector v_d.
- The length of v_d, i.e. $\| v_d \|$, is crucial for classification performance.

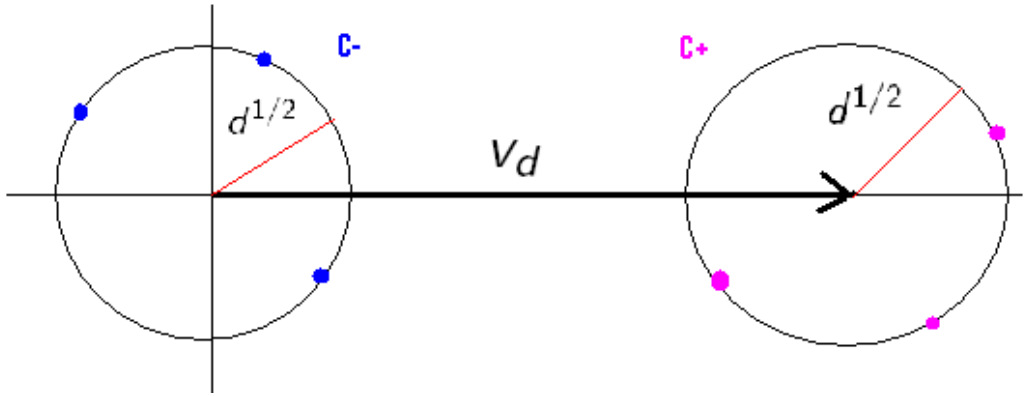


Figure: Case $\| v_d \| \gg d^{1/2}$

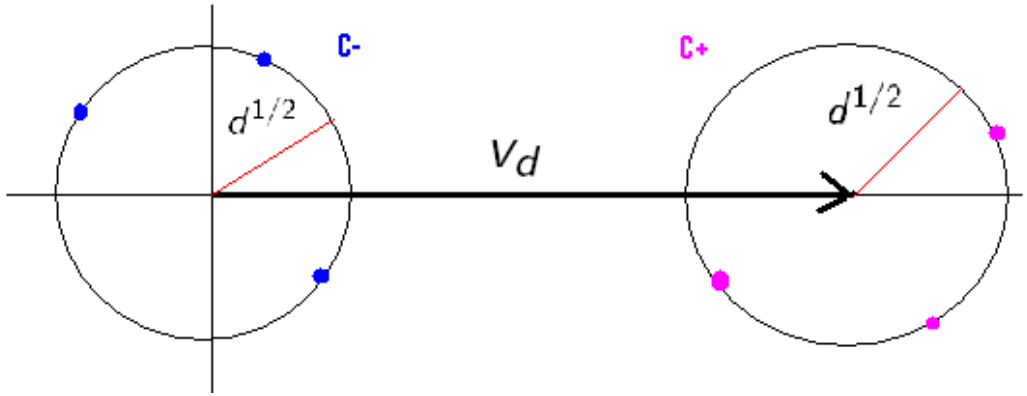


Figure: Case $\| v_d \| \gg d^{1/2}$

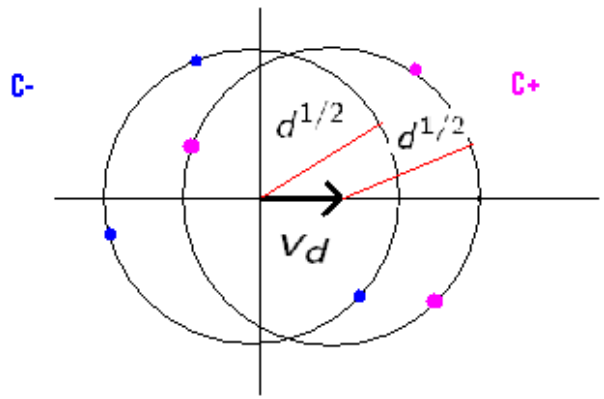


Figure: Case $\| v_d \| \ll d^{1/2}$

- Here, $\| v_d \| \approx d^{1/2}$ is a critical boundary for classification performance.
- Because the separating hyperplane of the methods are determined by their normal vector, the behavior of classification is studied considering the direction of this vector.
- Under our assumptions, v_d is the optimal direction for the normal vector of the separating hyperplane.
- Thus, a discrimination method has good properties if the angle between its normal vector and v_d is close to zero.

Asymptotic behavior

Theorem 4.1

If v represents the normal vector of MD, SVM, DWD or MDP hyperplane we have that

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} 0, & \text{if } \| v_d \| d^{-1/2} \to \infty; \\ \frac{\pi}{2}, & \text{if } \| v_d \| d^{-1/2} \to 0; \\ \arccos\left(\frac{c}{(\gamma + c^2)^{1/2}} \right), & \text{if } \| v_d \| d^{-1/2} \to c > 0; \end{cases}$$

as $d \to \infty$, where $\gamma = \frac{1}{m} + \frac{1}{n}$.

Asymptotic behavior

Theorem 4.1

If v represents the normal vector of MD, SVM, DWD or MDP hyperplane we have that

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} 0, & \text{if } \| v_d \| d^{-1/2} \to \infty; \\ \frac{\pi}{2}, & \text{if } \| v_d \| d^{-1/2} \to 0; \\ \arccos\left(\frac{c}{(\gamma + c^2)^{1/2}} \right), & \text{if } \| v_d \| d^{-1/2} \to c > 0; \end{cases}$$

as $d \to \infty$, where $\gamma = \frac{1}{m} + \frac{1}{n}$.

- $\| v_d \| \gg d^{1/2}$: consistent
- $\| v_d \| \ll d^{1/2}$: strongly inconsistent

Different asymptotic behavior of the NB method

Proposition 4.1

Let v be the NB normal vector and assume $m+n>6$. Suppose $v_d = \beta \mathbf{1}_d$ where $\beta = \beta_d \to c$ as $d \to \infty$, with $0 \le c \le \infty$, then

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} \arccos\left(\dfrac{\widetilde{\mu}}{(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}}\right), & \text{if } c=\infty; \\ \arccos\left(\dfrac{c\widetilde{\mu}}{(\gamma+c^2)^{1/2}(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}}\right), & \text{if } c \ge 0; \end{cases} \tag{9}$$

as $d \to \infty$, where $\gamma = \frac{1}{m}+\frac{1}{n}$,

$$\widetilde{\mu} = \frac{m+n-2}{m+n-4}, \qquad \widetilde{\sigma}^2 = \frac{2(m+n-2)^2}{(m+n-4)^2(m+n-6)}. \tag{10}$$

Different asymptotic behavior of the NB method

Proposition 4.1

Let v be the NB normal vector and assume $m+n>6$. Suppose $v_d = \beta \mathbf{1}_d$ where $\beta = \beta_d \to c$ as $d \to \infty$, with $0 \le c \le \infty$, then

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} \arccos\left(\dfrac{\widetilde{\mu}}{(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}}\right), & \text{if } c=\infty; \\ \arccos\left(\dfrac{c\widetilde{\mu}}{(\gamma+c^2)^{1/2}(\widetilde{\sigma}^2+\widetilde{\mu}^2)^{1/2}}\right), & \text{if } c \ge 0; \end{cases} \tag{9}$$

as $d \to \infty$, where $\gamma = \frac{1}{m}+\frac{1}{n}$,

$$\widetilde{\mu} = \frac{m+n-2}{m+n-4}, \qquad \widetilde{\sigma}^2 = \frac{2(m+n-2)^2}{(m+n-4)^2(m+n-6)}. \tag{10}$$

- The NB normal vector is always inconsistent.
- It is strongly inconsistent when $c=0$.

Similar asymptotic behavior of the NB method

Proposition 4.2

Let v be the NB normal vector and assume $m + n > 6$. Suppose $v_d = (d^\delta, 0, \ldots, 0)^\top$, with $\delta > 0$. Then

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} 0, & \text{if } \delta > 1/2; \\ \dfrac{\pi}{2}, & \text{if } \delta < 1/2; \end{cases} \tag{11}$$

as $d \to \infty$.

Similar asymptotic behavior of the NB method

Proposition 4.2

Let v be the NB normal vector and assume $m + n > 6$. Suppose $v_d = (d^\delta, 0, \ldots, 0)^\top$, with $\delta > 0$. Then

$$\text{Angle}(v, v_d) \xrightarrow{w} \begin{cases} 0, & \text{if } \delta > 1/2; \\ \dfrac{\pi}{2}, & \text{if } \delta < 1/2; \end{cases} \tag{11}$$

as $d \to \infty$.

- The NB has a similar behavior to the other four methods when $c = 0$ ($\delta < 1/2$) and $c = \infty$ ($\delta > 1/2$).

- **First case:** The mean vector must interact with all of the estimated marginal variances. This introduces a large amount of noise into the classification process.
- **Second case:** Only one estimated marginal variance has substantial influence on the classification, so its effect is asymptotically negligible.

Comparison of the methods

- Hall, Marron and Neeman (2005) only studied misclassification probabilities, where little contrast between methods was available.
- We provide better comparison between methods through deeper study of the asymptotic behavior of them in terms of the normal vectors of their separating hyperplanes.

When $\| v_d \| \gg d^{1/2}$

Theorem 5.1 (Case $m = 1$, $n = 2$)

Suppose $\| v_d \| = d^{\delta}$ *with* $1/2 < \delta < 1$. *Let* v_{SVM} *and* v_{MD} *be the normal vectors of the SVM and MD methods, respectively. Let* $A_{SVM} = \text{Angle}(v_{SVM}, v_d)$ *and* $A_{MD} = \text{Angle}(v_{MD}, v_d)$, *then*

$$A_{SVM}^2 - A_{MD}^2 = 2\frac{\mathcal{X}_1^2}{d} + O_p(d^{-(1+\epsilon)}) \qquad \text{as } d \to \infty,$$

for some $\epsilon > 0$ *and where* $\mathcal{X}_1^2$ *is a r.v. with the chi-square distribution with one degree of freedom.*

When $\| v_d \| \gg d^{1/2}$

Theorem 5.2 (Case $m = 1$, $n = 2$)

Suppose $\| v_d \| = d^{\delta}$ *with* $\delta > 1$. *Let* A_{SVM} *and* A_{MD} *be as in Theorem 5.1, then*

$$P(A_{SVM} > A_{MD}) \longrightarrow 1 \qquad \text{as } d \to \infty. \tag{12}$$

When $\| v_d \| \ll d^{1/2}$

Theorem 5.3 (Case $m = 1$, $n = 2$)

Suppose $v_d = (d^\delta, 0, \ldots, 0)^\top$ *with* $0 < \delta < 1/2$. *Let* A_{SVM} *and* A_{MD} *be as in Theorem 5.1, then*

$$\begin{aligned} A_{SVM} - A_{MD} &= \frac{N_0}{d} + \frac{\mathcal{X}_1^2}{\gamma^{1/2} d^{3/2-\delta}} + O(d^{-(3/2-\delta+\epsilon')}) \\ &= \frac{N_0}{d} + O_p(d^{-(1+\epsilon)}) \end{aligned}$$

as $d \to \infty$ *for some* $\epsilon', \epsilon > 0$, *where* N_0 *converges in distribution to the product of two independent standard normal random variables as* $d \to \infty$ *and* $\mathcal{X}_1^2$ *is a r.v. with the chi-square distribution with one degree of freedom.*

Conclusions

- The MD, SVM, DWD and MDP methods have the same asymptotic behavior when $d \to \infty$.
- However, depending on the asymptotic behavior of $\| v_d \|$ these methods may be asymptotically consistent or inconsistent.
- The NB sometimes behaves worse than the other four methods.
- Generally, MD is better than the SVM method when d is large but fixed.


J. Ahn and J. S. Marron.
The Maximal Data Piling Direction for Discrimination.
Biometrika, 97(1):254–259, 2010.


A. Bolivar-Cime and J. S. Marron.
Comparison of Binary Discrimination Methods for High Dimension Low Sample Size Data.
June, 2011.
Manuscript.


P. J. Bickel and E. Levina.
Some Theory for Fisher's Linear Discriminant Function, "Naive Bayes" and some alternatives where there are many more variables than observations.
Bernoulli, 10:989–1010, 2004.


C. J. C. Burges.
A Tutorial on Support Vector Machines for Pattern Recognition.
Data Mining and Knowledge Discovery, 2:121–167, 1998.


N. Cristianini and J. Shawe-Taylor.
An Introduction to Support Vector Machines and other kernel-based learning methods.
Cambridge University Press, Cambridge, U.K., 2000.


C. Cortes and V. N. Vapnik.
Support-Vector Networks.
Machine Learning, 20:273–297, 1995.


P. Hall, J.S. Marron, and A. Neeman.
Geometric Representation of High Dimension, Low Sample Size Data.
Journal of the Royal Statistical Society. Series B. Statistical Methodology., 67(3):427–444, 2005.


A. J. Izenman.
Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning.
Springer Texts in Statistics. Springer, New York, 2008.


J. S. Marron, M. J. Todd, and J. Ahn.

Distance-Weighted Discrimination.
Journal of the American Statistical Association, 102(480):1267–1271, 2007.


B. Scholkopf and A. J. Smola.
Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond.
The MIT press, Cambridge, Massachusetts, 2002.


V. N. Vapnik.
Estimation of Dependences Based on Empirical Data.
Springer Series in Statistics. Springer-Verlag, Berlin, 1982.


V. N. Vapnik.
The Nature of Statistical Learning Theory.
Springer-Verlag, Berlin, 1995.