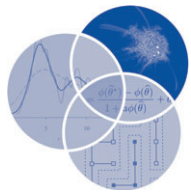
Focus Article



Distance-weighted discrimination

J. S. Marron*

Distance-weighted discrimination is a classification (discrimination) method. Like the popular support vector machine, it is rooted in optimization; however, the underlying optimization problem is modified to give better generalizability, particularly in high dimensions. The two key ideas are that distance-weighted discrimination directly targets the data piling problem and also correctly handles unknown, unbalanced subclasses in the data. A useful property of distance-weighted discrimination, beyond just good classification performance, is that it provides a direction vector in high-dimensional data space with several purposes, including indication of driving phenomena behind class differences, data visualization, and batch adjustment tasks.

How to cite this article:
WIREs Comput Stat 2015, 7:109–114. doi: 10.1002/wics.1345

Keywords: classification; discrimination; machine-learning; distance-weighted

INTRODUCTION

Distance-weighted discrimination (DWD) was invented by Marron, Todd, and Ahn[1]. Historical perspective and motivation are given in *Motivation* section. The need for DWD was realized from a high-dimensional analysis of the support vector machine (SVM), of the type shown in Figure 1. In particular, projections onto the normal vector of the separating plane can be subject to data piling, leading to serious loss of generalizability. *Computation* section gives references to numerical details, and also to available software. Applications of DWD to the important tasks of data visualization and bias adjustment are discussed in *Visualization* and *Bias Adjustment* sections. Papers studying the mathematical underpinnings of DWD are reviewed in *Asymptotic Theory* section. A number of useful variations are explored in *Variations* section.

MOTIVATION

The simplest version of the statistical task of classification (here this is a synonym for discrimination) starts with two classes, e.g., Healthy versus Ill people, and the goal is to develop a rule for assigning people to a class, based on some measurements. The starting point is a training data set of measurements made on people whose health status is known. Many papers have been written on approaches to this problem, starting perhaps with the linear discriminant analysis (LDA) of Fisher[2] (see Ref 3 for good access to the literature). There are many useful perspectives from which this large literature can be viewed.

One dichotomy of existing methods is into classical statistical approaches versus machine-learning type approaches. Statistical methods are based on fitting probability distributions to the data in each class, and then using these distributions to construct the classifier, perhaps using a likelihood ratio rule. In contrast, the basis of machine learning methods is an optimization problem. Both methodologies have a long history of relative strengths and weaknesses, and there is no uniformly dominant methodology. However, useful insights come from the simultaneous consideration of both sets of methodologies.

Support Vector Machine

An idea that is central to research in machine learning is the SVM, invented by Vapnik[4,5] (see Refs 6 and 7 for good introductions to this set of ideas). An important point is that the generic SVM does not involve probability distributions in any way. Basic motivation comes

*Correspondence to: marron@unc.edu

Department of Statistics, University of North Carolina, Chapel Hill, NC, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

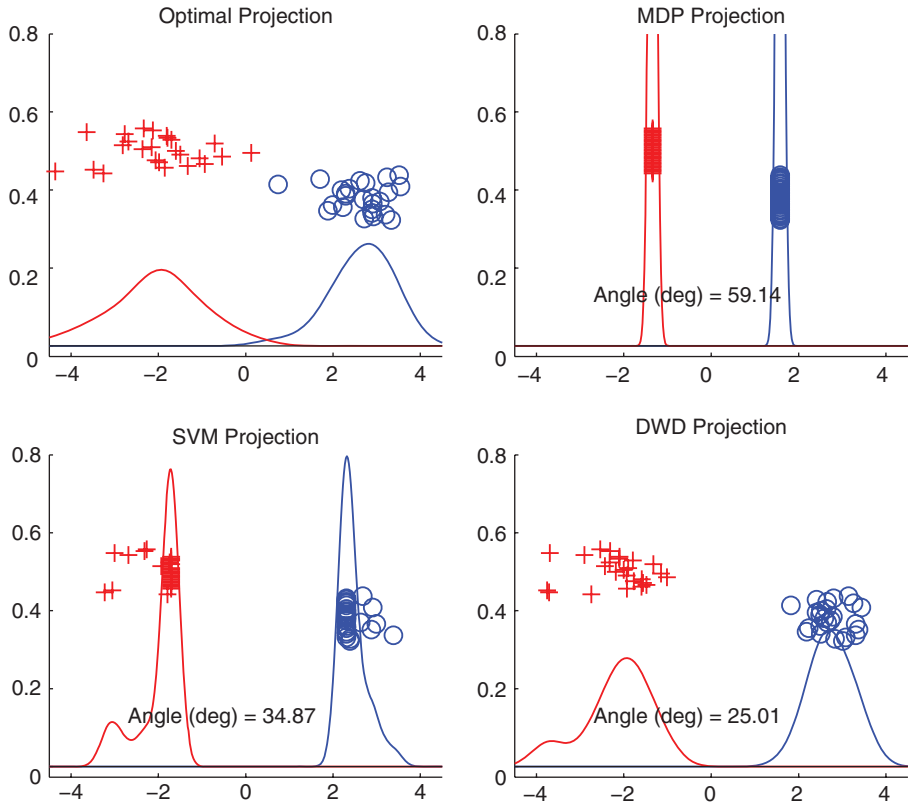


FIGURE 1 | Simulated example illustrating data piling in high dimensions, and how it is mitigated by DWD. Each plot shows projection scores on a particular direction in the space (with common axes for direct comparison, as indicated by the titles). Angles to the optimal in degrees are shown for each other direction, indicating superior generalizability for DWD.

from considering a simple two-dimensional bull's eye data set, and discrimination methods based on a separating line, such as the linear discriminant where the measurements consist of vectors $\underline{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^2$, where one class is distributed in a cloud near the origin, and the other class is distributed in a roughly circular fashion (plus noise) all around the origin. For a number of popular methods such as LDA, these data are very challenging. In particular, there is nothing close to a separating line for such data. However, Aizerman et al.[8] found a clever way to use purely linear methods in an effective way, even for this very challenging toy example. Their idea is to embed the data into an appropriate higher dimensional space. In this case, a very appropriate embedding is $\underline{x} \to \begin{bmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{bmatrix}$. At first glance, this mapping appears useless as there is no new information. However, this four-dimensional version of the data now exhibits a far different class relationship, where the two classes are now completely separable in the subspace of the last two coordinates. This elegant idea has been extended in many interesting ways in the machine-learning literature and is frequently implemented using a Gaussian kernel embedding.

At this point, it is very relevant to compare the statistical and machine-learning approaches to classification. While it is clear that the embedded four-dimensional data is far easier to work with, note that it is very hard to see how it can be usefully modeled with conventional probability distributions because the data lie along an unfamiliar two-dimensional manifold in four-dimensional space. With substantial effort, some distributions could be developed in this case; however, this becomes much harder in higher dimensions with unknown data distributions, and would at the least require many different case wise solutions in general. This motivates taking a machine-learning approach, by considering optimization approaches that attempt to effectively separate the classes, e.g., by a separating plane. The standard SVM optimization does this by considering projected distances from each data point to the separating plane, and seeking to maximize the smallest of these. This approach gives excellent

results in the above two-dimensional bull's eye toy example.

The High-Dimensional Challenge

However, in higher dimensions, the standard SVM formulation typically suffers substantially in terms of generalizability, the ability to classify new data effectively, because of a tendency to be too strongly affected by small scale noise artifacts. This point is illustrated in Figure 1, which is based on a simulated Gaussian data set in $d = 60$ dimensions. Each of the two classes have 25 data points generated from isotropic (i.e., identity covariance matrix) Gaussian distributions with differing means. Each of the panels of Figure 1 represent different one dimensional projections of the data, onto various direction vectors in $\mathbb{R}^{60}$. In each case, the one-dimensional distributions of data from the two classes are shown as jitter plots, where the horizontal coordinate is the projection score, and the vertical component is random and provides visual separation. The upper left panel shows projections onto the direction determined by the underlying distributional means. This shows that the underlying distributions are essentially separable. However, there is a very substantial challenge to finding this direction, using just the data, because of the high dimensionality of the space.

A perhaps surprising direction is shown in the upper right, where all of the data in one class project to a single point, and the data in the other class project to a different point. To understand this, recall that the 25 points in the first class determine a 25-dimensional subspace. For every direction vector in the 35-dimensional orthogonal component, the entire class will project to 0. Similarly there is also a 35-dimensional orthogonal subspace for the other class, and in the 10-dimensional intersection, both classes project to 0. In this context, the existence of directions with projections such as those shown in the upper right panel is not surprising. As shown by Ahn and Marron[9], in fact, such directions exist with probability 1 when the underlying distributions are absolutely continuous with respect to Lebesgue measure. The direction that maximizes the separation between projection points, shown here, is called the maximal data piling (MDP) direction. An interesting aside is that the formula for the MDP direction is quite similar to the formula for LDA, where the pooled within class covariance matrix estimate is replaced by the overall sample covariance matrix. The MDP projections highlight a pivotal challenge to high-dimensional classification: overfitting. If one generates a new set of data from the probability distributions used here, there will also be an MDP direction; however, it will be completely different. One way to study this is to look at the angle between this direction, and the optimum (whose projections are shown in the top left panel) which is almost 60°. This casts a large amount of doubt on how well the class label of a new data point from either class can be determined using the MDP direction, the root cause of which is that MDP is strongly driven by small-scale noise artifacts in the data, which are irrelevant to the actual classification. A related aside is a misconception, stated on page 119, equation (69) of Ref 3, that LDA can be characterized by finding the direction to maximize the difference between class means, while minimizing the projected variances. In fact that criterion results in the MDP direction, which turns out to be the same as LDA only when the dimension is smaller than the sample size, but is different in important ways otherwise.

Projections onto the SVM direction are shown in the lower left panel. Because this direction maximizes the minimum distance to the separating hyperplane, it is not surprising that these points show more separation between classes than is seen in the top right panel (note the common horizontal axes), or even in the top left true direction. This SVM direction has improved generalizability properties over the MDP in this case, as shown by the decreased angle of only about 35° to the optimal. However, the lower left panel also indicates a troubling property of SVM, there is a tendency for a large number of data points to pile up. Those are the points that exactly achieve the minimum distance to the separating plane, which are called support vectors. This piling of the support vectors in this projection negatively impacts the generalizability of the SVM, again because small-scale noise artifacts have too strong an effect. While this impact is less strong than for MDP, it is clearly quite substantial in Figure 1, which provides the motivation for the improved method given in the next section.

Distance-Weighted Discrimination

DWD, introduced by Marron, Todd, and Ahn,[1] directly targets the poor high-dimensional properties of SVM, by modifying the underlying optimization problem, as detailed in the next section. The main idea is based on the observation that the SVM direction ultimately is a function of only the support vectors (in a sense similar to the way in which the median feels only the central observation(s) in a data set). A natural way to avoid the piling of the support vectors is to allow all data points to have at least some influence on the DWD direction. Many of the

properties of the SVM can be maintained by giving less influence to points farther from the separating plane. DWD achieves this by minimizing the sum of inverse distances from each data point to the separating hyperplane. Thus, each data point can be viewed as a pole, which pushes the separating plane away from it. The beneficial effects of solving this improved optimization problem can be seen in the lower right panel of Figure 1. Note that there is no longer any data piling. The improved generalizability of DWD over SVM is demonstrated by the angle to the optimal now going down to around 25°. This intuitive idea, concerning the negative impact of support vector data piling on the SVM direction, has been verified in the simulation study of Marron, Todd, and Ahn.[1] The original DWD method was modified to handle unequal class sizes, and also general prior probabilities, by Qiao et al.[10] That paper also studies solutions to the data piling problem of SVM and how that is solved by DWD, based on loss function ideas. Additional results in a similar direction can be found in Ref 11.

COMPUTATION

The DWD optimization problem is formally developed, in section 2 of Ref 1. Also discussed there is the choice of tuning parameter, and the fact that there is a simple default choice, which usually works very well, in contrast to tuning of the SVM.

While a number of approaches to the numerical calculation of the DWD direction are possible, second-order cone programming, as detailed in section 2 of Ref 1 is recommended. An R implementation is available in Ref 12, and Matlab software is found in Ref 13.

VISUALIZATION

A very common tool for data visualization, especially in high dimensions is principal component analysis, see Ref 14 for a good introduction. In particular, scatterplots of pairs of scores (i.e., coefficients of the projections of the data points onto eigenvectors) often reveals important relationships among the data points. When there are known subgroups in the data, the PCA view may not reveal all aspects of interest, because it is driven completely by overall variation, and ignores the subgroup information. In this situation, better directions, which focus more on group differences, can be very useful. DWD has distinguished itself for this purpose. Because of its explicit goal of separation, it tends to give a much better view than the direction separating group means. Furthermore, because of its tendency to avoid data piling, DWD also generally gives much less distracting views than are available from SVM. Often it is useful to combine DWD scores with PCA scores into a scatterplot matrix. As the DWD direction may not be orthogonal to some of the PCA directions, interpretability can be enhanced by using *orthogonal* PCA scores, by restricting the PCA to the hyperplane orthogonal to the DWD direction.

Statistical Verification

While DWD is very useful for providing visual separation between subgroups in data, there is a trade-off which needs to be kept in mind: it can be too good. In particular, as shown in Ref 15, for very high-dimensional data, when group labels do not correspond to any actually different groups, DWD can find projections that give apparently clear visual differences. It is important to guard against such spurious visual impressions by combining statistical inference with the visualization. The DiProPerm test proposed by Wei et al.[15] is such a methodology, that is intimately linked with the visualization process.

The DiProPerm approach starts with a *DI*rection vector, which could be DWD or another direction of choice. The data are then *PRO*jected onto that vector, and the difference between groups is summarized using a univariate statistic computed on the projections, such as the two sample t statistic. Because the DWD direction is aimed at separating the class projections, care needs to be taken in assessing the statistical significance of the univariate summary statistic. DiProPerm does this using a standard *PERM*utation method. The class labels are randomly permuted, the DWD direction is recomputed for each permutation, the permuted data are then projected, and again summarized, to provide a null distribution, whose tail probability then gives the DiProPerm p-value. A number of properties of this test are analyzed in Ref 15.

BIAS ADJUSTMENT

Another important application of DWD is the removal of various types of biases in a number of different types of chip-based genetic measurements, with microarrays being the best known of these. A common term for such data distortions is batch effects, referring to systematic variations in the manufacture process, although that can include many other things as well, such as different operators, dates of use, and laboratory conditions. An important point is that there are various contexts in which batch adjustment is performed. A particularly challenging one is exploratory

data analysis, such as clustering, done without predetermined knowledge of the effect of interest. Benito et al.[16] proposed DWD as an adjustment method for such situations. While this method has performed very well in this context, the reason was not widely understood, until Liu et al.[17] In that paper, it was revealed that this widely observed good performance of DWD is due to hidden subclasses, of varying proportions among the data. Methods based on means are strongly impacted by these hidden subgroups. DWD is much more robust, as it is driven by distances between subgroups and not their proportions.

ASYMPTOTIC THEORY

Also of interest is the mathematical statistics of DWD. Here there is a lot of scope for future work. Qiao et al.[10] have taken a first look at this, e.g., establishing the Fisher Consistency of DWD under suitable assumptions. In view of the important high-dimensional properties of DWD a particularly relevant type of mathematical statistics is high-dimension low-sample size asymptotics, as outlined in Ref 18. A first paper in this direction is Ref 19.

VARIATIONS

A number of useful variations of DWD should also be noted.

Multi-Class DWD

The above discussion has been in terms of binary DWD, where there are only two classes. Also important is the case of multiple classes. There are various ways that pairwise DWD can be extended to this case, but most appealing is to formulate a proper joint optimization problem. This has been performed in Ref 20, where the resulting solution results in a direction vector for each class. The resulting classification is done by choosing the class with maximal projection score.

Bidirectional Discrimination

A weakness of conventional DWD is that it is linear in the sense of only providing a separating hyperplane, which can be inadequate in some situations, which was the motivation for the high-dimensional embedding ideas discussed in *Support Vector Machine* section. While polynomial embeddings can be useful, as described there, they can also easily lead to over-fitting in high dimensions. Gaussian kernel embeddings can mitigate this somewhat, but both over-fitting and tuning challenges remain. An approach which has some of the improved flexibility of the kernel approaches, while keeping most of the superior generalizability properties of linear methods is the bi-directional discrimination proposed by Huang et al.[21] The main idea is to replace the single DWD direction by say a pair of directions that are combined into a single classifier using a product structure which essentially gives quadrants as classification regions. Calculation is performed based on a modified optimization problem, which can be solved in an iterative fashion using repeated DWD calculations.

Sparse DWD

In some high-dimensional data analytic contexts, there is reason to believe that most of the signal of interest is contained in a relatively few variables, and the remaining variables only contain noise. This is the motivation for sparse statistical methods, which have received a large amount of study. In some as yet unpublished work, Qiao and Zhang have proposed and studied a sparse variation of DWD for such contexts.

REFERENCES

1. Marron JS, Todd MJ, Ahn J. Distance-weighted discrimination. *J Am Stat Assoc* 2007, 102:1267–1271. doi:10.1198/016214507000001120.
2. Fisher RA. The use of multiple measurements in taxonomic problems. *Ann Hum Genet* 1936, 7:179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.
3. Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York: John Wiley & Sons; 2001.
4. Vapnik VN. *Estimation of Dependences Based on Empirical Data*. New York: Springer-Verlag; 1982.
5. Vapnik VN. *The Nature of Statistical Learning Theory*. Berlin: Springer; 2000.
6. Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge, UK: Cambridge University Press; 2000.
7. Schölkopf B, Smola AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA: MIT Press; 2001.

8. Aizerman A, Braverman EM, Rozoner LI. Theoretical foundations of the potential function method in pattern recognition learning. *Autom Remote Control* 1964, 25:821–837.
9. Ahn J, Marron JS. The maximal data piling direction for discrimination. *Biometrika* 2010, 97:254–259. doi:10.1093/biomet/asp084.
10. Qiao X, Zhang HH, Liu Y, Todd MJ, Marron JS. Weighted distance weighted discrimination and its asymptotic properties. *J Am Stat Assoc* 2010, 105:401–414. doi:10.1198/jasa.2010.tm08487.
11. Zhang L, Lin X. Some considerations of classification for high dimension low-sample size data. *Stat Methods Med Res* 2013, 22:536–549.
12. Huang H, Lu X, Liu Y, Haaland P, Marron JS. R/DWD: distance-weighted discrimination for classification, visualization and batch adjustment. *Bioinformatics* 2012, 28:1182–1183.
13. Marron JS. Smoothing, Functional Data Analysis, and Distance Weighted Discrimination Software, 2013. Available at: http://www.unc.edu/~marron/marron_software.html
14. Jolliffe I. *Principal Component Analysis*. New York: John Wiley & Sons; 2005.
15. Wei S, Lee C, Wichers L, Li G, Marron JS. Direction-projection-permutation for high dimensional hypothesis tests, 2013. Available at: http://arxiv.org/abs/1304.0796
16. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS. Adjustment of systematic microarray data biases. *Bioinformatics* 2004, 20:105–114. doi:10.1093/bioinformatics/btg385.
17. Liu X, Parker J, Fan C, Perou CM, Marron JS. *Visualization of Cross-Platform Microarray Normalization*. New York: John Wiley & Sons; 2009, 167–181.
18. Hall P, Marron JS, Neeman A. Geometric representation of high dimension, low sample size data. *J R Stat Soc Series B Stat Methodol*, 67: 427–444, 2005. 10.1111/j.1467-9868.2005.00510.x
19. Bolivar-Cime A, Marron JS. Comparison of binary discrimination methods for high dimension low sample size data. *J Multivar Anal*, 115: 108–121. 2013. 10.1016/j.jmva.2012.10.001
20. Huang H, Liu Y, Du Y, Perou CM, Hayes DN, Todd MJ, Marron JS. Multiclass distance-weighted discrimination. *J Comput Graph Stat* 2013, 22:953–969. doi:10.1080/10618600.2012.700878.
21. Huang H, Liu Y, Marron JS. Bidirectional discrimination with application to data visualization. *Biometrika* 2012, 99:851–864. doi:10.1093/biomet/ass029.

