A nomenclature for linear discriminant algorithms

Patrick Haffner
AT&T Labs-Research, Florham Park, NJ 07932
haffner@research.att.com

Statistical machine learning focuses on non-parametric models that learn the minimum necessary to perform the targeted discrimination. For instance, in a binary classification problem, SVMs will only use the examples located within a boundary margin between the two classes to compute the optimal separation. However, one can consider other discrimination settings that use the entirety of the data. We first review four different discrimination settings, supported by the example in Figure 1. The task is to separate the '+' class from the '-' class. Training examples for each classes live in separate horizontal lines. Discrimination always involves positive examples whose scores one wants to push up, and negative examples one wants to push down. This can be expressed in a geometric framework, when one maximizes a separation margin, or in a probabilistic framework, where all scores are normalized to sum to 1, and just pushing up the positive ones necessarily pushes down the negative ones.

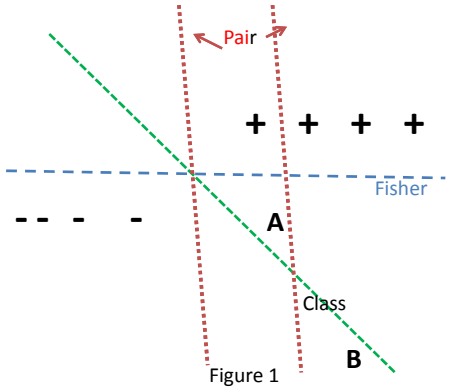

Figure 1

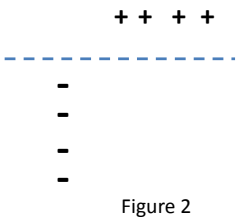

Figure 2

Class discrimination The traditional discriminant setting is, given an input, to return a class. The system is trained so that the score of the target class is larger than any other classes. In a probabilistic framework, this can be achieved by maximizing, for target class c, the posterior $P(c|\mathbf{x}) = \frac{f_c(\mathbf{x})}{\sum_d f_d(\mathbf{x})}$. Commonly used scores are exponential linear $f_c(\mathbf{x}) = \exp \mathbf{w}_c \cdot \mathbf{x}$, resulting in a logistic regression function. Similarly, in a geometric framework, one can maximize the width of the separation margin between classes. This is shown as the diagonal separation plane in Figure 1.

Fisher discrimination The key idea, usually expressed in terms of intra-class variance minimization, is to impose that the score, while providing proper separation, should have comparable values for all examples in a given class. A typical geometric interpretation is to perform a least square regression to the class labels, which is exactly equivalent to the Fisher discriminant [4]. In our example in Figure 1, as class variances are minimized by a projection on the vertical, one obtains an horizontal separation plane. Under the assumption that each class lives in a horizontal manifold, test example 'A' is properly classified as negative by the Fisher discrimination, while it is misclassified by the class discrimination.

Pairwise discrimination In a ranking setting, one would like to sort all the examples according to whether they are relevant to class c or not. For bipartite ranking, where examples can be split into two

sets, this can be achieved by maximizing the score difference for each positive-negative sample pair. The proportion of pairs which are properly ranked corresponds the the Area under the ROC Curve (AUC). A geometric interpretation leads a simple extension of SVMs (SVM Rank). In Figure 1, a pairwise discrimination approach will enforces a "horizontal" ranking of the data. Knowledge of the class priors (which is not required during training) makes it possible to draw a specific (nearly) vertical separation plane (we show here two examples). While it ignores data distribution assumptions, and would misclassify example 'A', it is good at maximizing the precision over top ranked examples, considering test example 'B' as positive.

A new contribution proposed here is to provide a probabilistic interpretation through a Maximum Entropy derivation that extends previous work on Boosting. Note both RankBoost and AdaBoost optimize pairwise discrimination [5]. However, the normalization used during AdaBoost training is tricky to interpret [3], as it requires the knowledge of the labels.

Density discrimination This fourth framework is little known, and leads to paradoxical remarks. For identifying the examples that are the most likely to belong to a class, one should maximize the ratio $P(i|c) = \frac{f_c(\mathbf{x}_i)}{\sum_j f_c(\mathbf{x}_j)}$. This can be interpreted as the probability to observe example i given class c. Noting that $P(\mathbf{x}|c) = \sum_{\mathbf{x}_i=\mathbf{x}} P(i|c)$, this scales as the likelihood ratio $\frac{P(\mathbf{x}|c)}{P(\mathbf{x})}$. An exponential linear form has been derived from the Maximum Entropy framework in [2].

At first, this approach seems to have the same qualities are pairwise discrimination: independence from class priors, and good reranking performance using $P(i|c)$. Surprisingly, the separation we obtained for Figure 1 is close to horizontal (as with Fisher discrimination). However, it only enforces similar scores for the positive class, which leads to the asymmetric separation observed in Figure 2.

We have reviewed four types of discrimination, using both geometric and probabilistic languages. Our first goal is to help identify the optimal algorithm for a given task, while corroborating previous observations [1]. To guarantee a good precision for the top-scoring examples returned by the classifier, one would prefer *pairwise* discrimination. *Fisher* discrimination should be considered when one wants to use more knowledge about the data than simply the examples that live on the boundary (assume for instance teach class lives in a separate manifold). Model-free estimates of the probability $P(X|c)$ are obtained through density discrimination (experiments suggest that traditional discriminant approaches are poor candidates). We also hypothesize that if two algorithms perform the same type of discrimination with the same regularization (L1 or L2), chances are that their performance will be very similar. For instance, for the same type of discrimination, it seems that there is little difference between the logistic or the hinge loss (SVM).

Finally, our new contribution is a general framework based on Maximum Entropy that covers all four types of discrimination. We will present in particular new connections between Pairwise, Fisher and Density discrimination, that help understand how they differ.

References

[1] Rich Caruana and Alexandru Niculescu-mizil. An empirical comparison of supervised learning algorithms. In *In Proc. 23 rd Intl. Conf. Machine learning (ICML06*, pages 161–168, 2006.

[2] Miroslav Dudik, Steven Phillips, and Robert E. Schapire. Performance Guarantees for Regularized Maximum Entropy Density Estimation. In *Proceedings of COLT'04*, Banff, Canada, 2004. Springer Verlag.

[3] G. Lebanon and J. Lafferty. Boosting and maximum likelihood for exponential models. In *Advances in Neural Information Processing Systems 14*, 2002.

[4] Sebastian Mika. *Kernel Fisher Discriminant, PhD Thesis*. 2002.

[5] Cynthia Rudin and Robert E. Schapire. Margin-based ranking and an equivalence between adaboost and rankboost. *Journal of Machine Learning Research*, 2009.