Variational Bounds for Discriminant Analysis

Yung-Kyun Noh, Jihun Hamm, and Daniel D. Lee*
Dept. of Electrical and Systems Engineering
University of Pennsylvania, Philadelphia, PA 19104

Fisher discriminant analysis (FDA) is the canonical technique to project high-dimensional labeled data onto a lower dimensional feature subspace [1]. This is accomplished by maximizing the Fisher criterion over a projection vector $\vec{w}$:

$$J_F(\vec{w}) = \frac{\vec{w}^T S_b \vec{w}}{\vec{w}^T S_w \vec{w}}, \tag{1}$$

where S_b is the between-class scatter matrix, and S_w is the average within-class scatter matrix. By maximizing this quotient, FDA seeks to separate the different labeled classes as much as possible while limiting the average spread within each class. FDA, as well as its kernelized generalization, has previously been shown to be very useful in extracting features for classification [2].

However, the Fisher criterion is not directly related to the actual Bayes classification error. In particular, with heteroscedastic data where the within-class covariances of the data are quite different, maximizing the Fisher criterion may be not optimal for separating the different classes [3]. Instead, given two projected class distributions, $p_1(x)$ and $p_2(x)$, we should really optimize the projection to minimize the classification error:

$$E = \frac{1}{2} \int \min[p_1(x), p_2(x)]\, dx. \tag{2}$$

Unfortunately, this error integral is analytically intractable to compute. One alternative is to use the following upper bound on the error:

$$E \leq \frac{1}{2} \int p_1(x)^{\alpha} p_2(x)^{1-\alpha}\, dx, \tag{3}$$

Variationally optimizing this bound over the range $0 \leq \alpha \leq 1$ results in a quantity known as the Chernoff distance between two probability distributions [4]. We demonstrate how this distance can be used for discriminant analysis, and show its improvement over conventional FDA.

*Presenting author

We also derive a complementary family of variational bounds on the error. These bounds arise by considering the following inequality:

$$1 - E = F = \frac{1}{2}\int \max[p_1(x), p_2(x)]\, dx \tag{4}$$

$$\geq \frac{1}{2}\int p_1(x)^{\frac{p_1(x)}{p_1(x)+p_2(x)}} p_2(x)^{\frac{p_2(x)}{p_1(x)+p_2(x)}}\, dx. \tag{5}$$

The advantage of considering these lower bounds on the *max* function is that the bounds are quite tight and can easily be generalized to more than two classes. Then by using Jensen's inequality, this bound can be extended to the variational form:

$$F \geq \frac{1}{2}\left[\frac{1}{2}\int p_1(x)\left(\frac{p_1(x)}{p_m(x)}\right)^{\beta} dx + \frac{1}{2}\int p_2(x)\left(\frac{p_2(x)}{p_m(x)}\right)^{\beta} dx\right]^{1/\beta}, \tag{6}$$

where $p_m(x) = \frac{1}{2}(p_1(x) + p_2(x))$ and is valid for all $0 \leq \beta \leq 1$. In the limit that $\beta \to 0$, we then obtain a bound on classification error based upon mutual information. We then show how these new variational bounds can be used for novel forms of discriminant analysis. Finally, we discuss how these new algorithms may be generalized using kernels and how they can be used with regularization techniques.

References

[1] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, 2001.

[2] S. Mika, G. Rätsch, J. Weston, B. Schölkopf, A. Smola, K. Müller. Constructing descriptive and discriminative nonlinear features: Rayleigh coefficients in kernel feature spaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25:623–627, 2003.

[3] M. Loog, R. P. W. Duin, R. Haeb-Umbach. Multiclass linear dimension reduction by weighted pairwise Fisher criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:762–766, 2001.

[4] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, USA, 1991.

Topic: learning algorithms
Preference: oral