Darrin P. Lewis, Ph.D.
205 East 78th Street, Apt. #12B, New York, New York 10075
dlewis.nyc@gmail.com
212-734-0721

Statement of Research Interest

My research is centered on creating machine learning methods and applying them to problems in computational biology. Computational biology differs from the traditional practice of biology by the exploitation of high throughput experiments for data gathering and algorithmic manipulation and analysis of the results.

In my thesis research [8], I introduced a novel nonstationary kernel combination (NSKC) technique [9] and used it to exploit heterogeneous data sets to predict protein function. This work extended the maximum entropy discrimination (MED) research of Tony Jebara [4, 3]. NSKC uses a mixture of gaussian distributions, with interpoint distances for each component expressed by a different kernel function. Classification is accomplished using the log ratio of two such models with their parameters estimated to maximize margin. Using NSKC, we achieved top results for several benchmark data sets and a protein function annotation experiment. We also use a toy problem to demonstrate that NSKC is fundamentally more powerful than linear combination techniques, such as SDP [5, 6]. This work has sparked further interest in nonstationary kernel combinations [2].

I explored the addition of SVM-like regularization to the nearest neighbor technique, creating regularized nearest neighbor (RNN) [7]. Experiments with a toy problem were promising. Generalization performance on noisy data was substantially better with RNN than with 1-NN and 3-NN. Nearest neighbor is natively multi-class, as opposed to the SVM which is natively two-class. This makes RNN particularly suitable for computational biology, which has many multi-class discrimination problems.

I created a technique that I called a pocket perceptron mean machine (P2M2) that used an ensemble of perceptrons for classification. The perceptrons are trained using the "pocket" heuristic [1] to handle non-separable data. Each perceptron is trained on a different permutation of the training data, resulting in a different solution due to instability of perceptron training. These perceptrons are used as an ensemble to classify unlabeled data. P2M2 is fast and is easily parallelized and scaled for huge data sets as are typical in computational proteomics.

I derived an analytical null model for k-nearest neighbor classification. Using this null model, I was able to compute p-values and surprise scores associated with particular class labelings of a data set. I used these p-values to validate clustering and other classification algorithms [11, 10].

In the course of my research, I have worked with non-parametric models, kernel methods, support vector machines, Bayesian statistics, large-margin probabilistic techniques, clustering, principal components analysis, nearest-neighbor techniques, and hidden Markov models. I have worked with microarray data sets, alignment scores, protein-protein interaction scores, and structure data sets, among others.

References

[1] S. Gallant. Perceptron-based learning algorithms. *IEEE Transactions on Neural Networks*, 1990.

[2] M. Gonen and E. Alpaydin. Localized multiple kernel learning. In *25th International Conference on Machine Learning (ICML)*, pages 352–359, 2008.

[3] T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems*, volume 12, December 1999.

[4] T. Jebara. *Machine Learning: Discriminative and Generative*. Kluwer Academic, Boston, MA, 2004.

[5] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. I. Jordan. Learning the kernel matrix with semi-definite programming. In C. Sammut and A. Hoffman, editors, *Proceedings of the 19th International Conference on Machine Learning*, Sydney, Australia, 2002. Morgan Kauffman.

[6] G. R. G. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble. Kernel-based data fusion and its application to protein function prediction in yeast. In R. B. Altman, A. K. Dunker, L. Hunter, T. A. Jung, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 300–311. World Scientific, 2004.

[7] D. Lewis and W.S. Noble. Regularized nearest neighbor. *Neural Information Processing Systems 16*, 2003 (submitted).

[8] D. P. Lewis. Combining kernels for classification. *Ph.D. dissertation*, 2006.

[9] D. P. Lewis, T. Jebara, and W. S. Noble. Nonstationary kernel combination. In *Proceedings of the International Conference on Machine Learning*, New York, NY, 2006. ACM Press.

[10] P. Pavlidis, D. P. Lewis, and W. S. Noble. Exploring gene expression data with class scores. In R. B. Altman, A. K. Dunker, L. Hunter, K. Lauderdale, and T. E. Klein, editors, *Proceedings of the Pacific Symposium on Biocomputing*, pages 474–485, New Jersey, 2002. World Scientific.

[11] J. Qin, D. P. Lewis, and W. S. Noble. Kernel hierarchical clustering of gene expression data. *Bioinformatics*, 19(16):2097–2104, 2003.