Discriminative Sparse Coding for Classification and Regression

Nishant A. Mehta Alexander G. Gray
College of Computing, Georgia Institute of Technology
266 Fifth St NW
Atlanta, GA 30332
niche@cc.gatech.edu agray@cc.gatech.edu
http://www.cc.gatech.edu/~niche

Sparse coding has seen a flurry of research activity in the machine learning community over the past decade. Nearly all of this work has focused on efficiently finding sparse codings that minimize reconstruction error. However, for learning problems the goodness of an embedding of data requires not only that the embedding be faithful to some properties of the original dataset (e.g. minimizing reconstruction error or preservering topology or local geometry), but also that the embedding perform well on a final learning task, such as classification or regression. There has been little work on incorporating discriminative information in the form of labels into the learning of good sparse codes. Although recently Mairal et al. [1] have made significant progress on discriminative learning of dictionaries used for sparse coding, their learning method focuses on learning one dictionary per class, with the discrimination power brought in by how well each dictionary approximates each point. Their method has been shown to yield state-of-the-art results in image classification tasks, which suggests this line of work deserves further pursuit. We note that a regression extension of their method does not seem possible, the method still has a reconstructive flavor, and an analysis of the generalization error of the technique has not yet been shown.

In this work, we propose a discriminative sparse coding method that seeks to learn dictionaries tuned to regression and classification tasks; for the case of classification tasks, this method uses a quite different, but intuitive, strategy than previous discriminative sparse coding efforts. Our objective was inspired by a recent algorithm, local coordinate coding, by Yu and Zhang [3]. The key is to directly incorporate labels into the objective to minimize, as follows

$$\min_{B,S,w} \|X - BS\|_F^2 + \lambda_w \sum_{i=1}^{n} \ell(w^T S_{(i)}, y_i) + \lambda_c \sum_{i=1}^{n} \sum_{j=1}^{m} |S_i^j| \|X_{(i)} - B_{(j)}\|_2^2,$$

for $\ell(\hat{y}, y)$ a convex loss function. Although this problem is not convex jointly in (B, S, w), for fixed S it is convex in (B, w) and for fixed (B, w) it is convex in S.

For simplicity, we describe here the case where $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$; the above then admits the reconstructive form (with respect to Z):

$$\min_{B,w,S} \|D(Z - \tilde{B}S)\|_F^2 + \lambda_c \sum_{i=1}^{n} \sum_{j=1}^{m} |S_i^j| \|X_{(i)} - B_{(j)}\|_2^2, \tag{1}$$

for $Z = [X^T, Y^T]^T$, $\tilde{B} = [B^T, w]^T$ and D a diagonal matrix with ones along the diagonal excepting $D_n^n = \sqrt{\lambda_w}$.

We first consider optimizing (1) via a simple alternating scheme: we switch between a *basis* step that optimizes $\tilde{B}$ (the dictionary B and the regression estimator w) and a *coding* step that optimizes the code S. Our purpose in this optimization procedure is to find a good discriminative dictionary (for a later function estimation step), rather than to find both a good dictionary and a good embedding S.

Once we have optimized B, we use the usual (non-discriminative) LCC objective to find a representation of each training point x in terms of the dictionary B, inducing the feature map $\phi_B(x) = \arg\min_s \|x - Bs\|_2^2 + \lambda_c \sum_{j=1}^{k} |s_j| \|x - B_{(j)}\|$. Regularized empirical risk minimization (e.g. using least squares loss for ridge regression) can then be applied on the feature mapped training data to estimate a hypothesis vector w.

Our main contribution is a learning theoretic analysis of such a discriminative sparse coding method. By framing the concept class of estimators induced by the above algorithm, we can provide bounds on the generalization error of the empirical risk minimization technique. The main difficulty not previously seen

in [3] is that the embedding is not independent of the labels $y_1, \ldots, y_n$ that later are used to estimate w. By drawing from recent work by Maurer and Pontil [2] on generalization bounds for coding schemes, which includes sparse coding as a special case, we can show generalization bounds for supervised learning with discriminative sparse coding methods.

References

[1] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[2] A. Maurer and M. Pontil. Generalization Bounds for K-Dimensional Coding Schemes in Hilbert Spaces. In *Algorithmic Learning Theory*, pages 79–91. Springer, 2010.

[3] K. Yu, T. Zhang, and Y. Gong. Nonlinear learning using local coordinate coding. *In Advances in Neural Information Processing Systems*, 22, 2009.