Closed-Form Supervised Dimensionality Reduction with GLMs

Irina Rish, Genady Grabarnik, Guillermo Cecchi
IBM T.J. Watson Research Center
Yorktown Heights, NY
{rish,genady,gcecchi}@us.ibm.com

The problem of *supervised dimensionality reduction* is to combine learning a good predictor with finding a *predictive structure*, such as a low-dimensional representation which captures the predictive ability of the features while ignoring the "noise". Indeed, performing dimensionality reduction *simultaneously* with learning a predictor often results into a better predictive performance than performing DR step separately from learning a predictor, as was demonstrated previously (e.g., see SVDM of [2], SDR-MM of [3], etc.), just as an embedded feature selection (e.g., sparse regression) often outperforms filter methods.

However, existing SDR approaches are typically limited to specific settings. For example, SVDM [2] effectively assumes a Gaussian-noise data model when minimizing sum-squared reconstruction loss, and is restricted only to classification problems when using SVM-like hinge loss as its prediction loss. SDR-MM method of [3] treats various data types (e.g., binary and real-valued) but is again limited only to multi-class classification problems. Recent work on distance metric learning [6, 5] is also limited to Gaussian data assumption, and discrete-label (typically binary) classification problem [6, 5]. Indeed, a majority of supervised dimensionality methods can be viewed as jointly learning a particular (often just linear) mapping from the feature space to a low-dimensional hidden-variable space, as well as a particular classifier that maps the hidden variables to the class label.

Our framework is more general as it treats both features and labels as exponential-family random variables, and allows to mix-and-match data- and label-appropriate *generalized linear models*, thus handling both classification and regression, with both discrete and real-valued data. It can be also viewed as a discriminative learning based on minimization of conditional probability of class given the hidden variables, while using as a regularizer the conditional probability of the features given the low-dimensional hidden-variable "predictive" representation.

The main advantage of our approach, besides generalization to a wider range of SDR problems, is that it uses simple, *closed-form* update rules when performing its alternate minimization procedure, and does not require performing optimization at every iteration of the procedure. This method yields a really short Matlab code, fast performance, and is always guaranteed to converge (to a local minimum, just like most of the existing hidden-variable model learning approaches). The convergence property, as well as closed form update rules, follow from the use of auxiliary functions bounding each part of the objective function (i.e., reconstruction and prediction losses). We exploits the additive property of auxiliary functions in order to "stack" together multiple objectives and perform, in a sense, a "multi-way" DR, i.e. joint dimensionality reduction from several datasets, such as feature vectors X and label Y.

More specifically, let X be an $N \times D$ data matrix with entries denoted X_{nd} where N is the number of i.i.d. samples, and n-th sample is a D-dimensional row vector denoted $\mathbf{x}_n$. Let Y be an N dimensional vector of class labels. We assume that our data points $\mathbf{x}_n$, $n = 1, ..., N$, are noisy versions of some "true points" $\theta_\mathbf{n}$ which live in a low-dimensional space, and that this low-dimensional representation is predictive about the class. It is assumed that noise is applied independently to each coordinate of $\mathbf{x}_n$ (i.e., that all dependencies among the dimensions are captured by low-dimensional representation), and that the noise follows exponential-family distributions with natural parameters $\theta_\mathbf{n}$, with possibly different members of the exponential family used for different dimensions. Namely, it is assumed that $N \times D$ parameter matrix Θ is a product of two low-rank matrices $\Theta_{nd} = \sum_l U_{nl} V_{ld}$ where the rows of the $L \times D$ matrix V correspond to the basis vectors, and the columns of the $N \times L$ matrix U correspond to the coordinates of the "true points" Θ_n, $n = 1, ...N$ in the L-dimensional space (for non-Gaussian noise, a *nonlinear* nonlinear surface in the original data space). We assuming exponential-family noise distribution for each X_{nd} with the corresponding natural parameter Θ_{nd}, i.e. $\log P(X_{nd}|\Theta_{nd}) = X_{nd}\Theta_{nd} - G(\Theta_{nd}) + \log P_0(X_{nd})$ where $G(\Theta_{nd})$ is the *cumulant*

or *log-partition* function which defines particular exponential family, e.g., Gaussian, multinomial, Poisson, etc. We can now view each row U_n as a new, low-dimensional representation of the corresponding data sample X_n. We also assume that the class label Y is a noisy function of this underlying low-dimensional representation, i.e. $Y_n = f(U_n)$ where f is some stochastic function. Generally, we can assume there are K prediction problems, so that Y is an $N \times K$ matrix. We will again assume noisy linear model with some exponential-family noise, $\log P(Y_n|\Theta_{Y_n}) = Y_n\Theta_{Y_n} - G_y(\Theta_{Y_n}) + \log P_0(Y_n)$. In general, we will use a *generalized linear model (GLM)* $E(X_d) = f_d(UV_d)$ for d-th feature (column d in X) with possibly different *link functions* f_d, and yet another GLM $E(Y) = f_y(UW)$ for the class label, where the logistic link function $f_y(\hat{\Theta})$ can be used for binary classification, identity link function $f_y(\hat{\Theta}) = \hat{\Theta}$, or any other appropriate link function for real-valued GLM can be used for regression.

SDR problem is formulated as joint optimization of two loss functions corresponding to the reconstruction loss L_r as a negative log-likelihood of the data $L_r = -\mathcal{L}_X(\Theta_X)$ and prediction loss L_p as the negative log-likelihood of the class labels $L_p = -\mathcal{L}_Y(\Theta_Y)$, where $\mathcal{L}_X(\Theta_X) = \sum_{nd} \log P(X_{nd}|\Theta_{Y_{nd}})$, $\mathcal{L}_Y(\Theta_Y) = \sum_{nk} \log P(Y_{nk}|\Theta_{Y_{nk}})$ and where $\Theta_X = UV$, $\Theta_Y = UW$, and the likelihoods above correspond to particular members of exponential family The optimization problem can be written as $\min_{U,V,W} L_p + \alpha L_r$ where α is the trade-off constant, or Lagrange multiplier.

While it is hard to come up with a globally optimal solution for the above (nonconvex) problem, we can employ the auxiliary function approach commonly used in EM-style algorithms in order to derive a set of closed-form iterative update rules that are guaranteed to converge to a local minimum. It is easy to show that an auxiliary function for the SDR objective can be derived for an arbitrary pair of L_r and L_p provided that we know their corresponding auxiliary functions, and using an additive property of of auxiliary functions. Namely, if $Q_1(\hat{\theta}, \theta)$ and $Q_2(\hat{\theta}, \theta)$ are auxiliary functions for $F_1(\theta)$ and $F_2(\theta)$, then it is easy to show that $Q(\hat{\theta}, \theta) = \alpha_1 Q_1(\hat{\theta}, \theta) + \alpha_2 Q_2(\hat{\theta}, \theta)$ is an auxiliary function for $F(\theta) = \alpha_1 F_1(\theta) + \alpha_2 F_2(\theta)$, where $\alpha_i > 0$ for $i = 1, 2$. Also, it is obvious that a function is an auxiliary for itself, i.e. $Q(\hat{\theta}, \theta) = F(\hat{\theta})$ is an auxiliary function for $F(\theta)$. This observations allows us to combine various dimensionality reduction approaches with appropriate predictive loss functions, given appropriate auxiliary functions for both. For a Bernoulli variables we use the variational bound on log-likelihood $\mathcal{L}(\theta) = \log P(s|\theta)$ that was originally proposed by [1] and subsequently used in logistic PCA algorithm of [4] (there is also a recent generalization of this bound to multinomial logistic regression that we plan to incorporate in our algorithm), while for Gaussian variables we just use the loglikelihood (sum-squared loss) itself. As a result, we obtain closed-form update rules for an alternating minimization that solves our SDR problem for a variety of data and label types.

We perform a variety of experiments, both on simulated and real-life problems. Results on simulated datasets convincingly demonstrate that our SDR approaches are capable of discovering underlying low-dimensional structure in even highly-dimensional noisy data, while outperforming SVM and SVDM, often by far, and practically always beating the unsupervised DR followed by learning a predictor. On real-life datasets, SDR approaches continue to beat the unsupervised DR by far, while often matching or somewhat outperforming SVM and SVDM.

References

[1] T. Jaakkola and M. Jordan. A variational approach to bayesian logistic regression problems and their extensions. In *Sixth International Workshop on Artificial Intelligence and Statistics*, 1997.

[2] F. Pereira and G. Gordon. The Support Vector Decomposition Machine. In *ICML2006*, 2006.

[3] Sajama and Alon Orlitsky. Supervised dimensionality reduction using mixture models. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 768–775, New York, NY, USA, 2005. ACM.

[4] A. Schein, L. Saul, and L. Ungar. A generalized linear model for principal component analysis of binary data. In *Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[5] K. Weinberger, J. Blitzer, and L. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In *NIPS2005*, 2005.

[6] E.P. Xing, A.Y. Ng, M.I. Jordan, and S. Russel. Distance Metric Learning, with Applications to Clustering with Side-information. In *NIPS2002*, pages 521–528, 2002.

Topic: learning algorithms
Preference: oral presentation