11-745 Advanced Statistical Learning Seminar Final Paper

Shashank Singh
Statistics & Machine Learning Departments
Carnegie Mellon University
Pittsburgh, PA 15213
sss1@andrew.cmu.edu

Abstract

We review some advances over the past decade in four topics in large scale machine learning: Unsupervised Clustering, Classification with Structured Learning, Semi-Supervised Clustering/Metric Learning, and Non-negative Matrix Factorization.

1 Large-Scale Unsupervised Clustering

Papers: [8, 1, 4] Background: [3]

These papers study large-scale unsupervised learning, primarily in the context of organizing a large document corpus without supervision (topic modeling), especially in a hierarchical manner that facilitates user exploration of large datasets.

Early work ([4]) proposed a generative model (Latent Dirichlet Allocation) describing the following generation procedure for each document:

1. A topic mixture θ is sampled from a k-dimensional Dirichlet prior.
2. For each word w_n, a topic z_n is drawn according to the weights of θ.
3. Each word w_n is then sampled from a multinomial distribution depending on z_n according to a prior distribution β_{z_n}.

To learn the prior parameters α and β, [4] uses a variational inference procedure which maximizes a lower bound on the likelihood (or, equivalently, minimizes the KL-divergence between the estimated posterior and an approximation of the true posterior).

It has been fairly well-established that normalizing documents to lie on a unit sphere via Term frequency-Inverse Document frequency (Tf-Idf) normalization significantly improves the performance of many text-mining algorithms. The two-parameter von Mises-Fisher (vMF) distribution plays the role of a Gaussian distribution (a smooth, unimodal, symmetric distribution decaying exponentially away from its mean) on the unit sphere, and hence serves as a natural basis for generative clustering models of documents. [8] studies three increasingly expressive vMF clustering models: a basic Bayesian vMF mixture model, a hierarchical augmentation of this, and a temporal vMF model for time-varying cluster centers. All three models are learned via variational inference. Two schemes are proposed for estimating the posterior distribution of the vMF concentration parameters: an MCMC sampling approach and a bounding approach, which uses an asymptotic normality approximation (empirically, the MCMC approach performs better). Empirical results also show that the variational inference approach consistently outperforms a similar EM approach for likelihood maximization.

General conclusions/observations:

- Variational Bayesian inference provides a powerful alternative to EM in many settings.
- von Mises-Fisher distributions are useful for modeling data distributed on the unit sphere, and hence for modelling documents in combination with Tf-Idf normalization.

2 Web-Scale Classification with Structured Learning

Papers: [9, 6]   Background: [3]

These papers study classification in the context that some classes arr subsets or supersets of other classes; i.e., the classes form a heirarchy. This is natural for web-scale classification, where, for example, topic or image labels mined from text lie at different levels of the hierarchy.

[9] models hierarchies by using a Gaussian distribution centered at the parameters of a class as a prior distribution for the parameters of the of child nodes of that class. This encourages nearby classes in the hierarchy to have similar parameters, and also naturally solves the general problem of hyperparameter selection inherent in Bayesian models (only the parameters for the root node must be set manually). Within each level of the hierarchy, [9] propose to use (multi-class) logistic regression and call the resulting model Hierarchical Bayesian Logistic Regression. Empirically, HBLR significantly outperforms both non-Bayesian hierarchical models, as well as flat (non-hierarchical) methods. As with the hierarchical clustering models above, the model can be learned efficiently via variational inference, and can also be parallelized by alternating iterations between even and odd layers of the hierarchy.

[6] follows up on [9] with two main modeling differences. Firstly, rather than a Bayesian model with the hierarchical relationships used to define the priors of the parameters, they use the hierarchical dependencies to define regularization terms for the parameters in a manner that still encourages classes nearby in the hierarchy to share similar model parameters. Secondly, they generalize to the case where dependencies between class-labels can be an arbitrary graph rather than specifically a hierarchy. Within each layer of the hierarchy, [6] propose to use either an SVM (HR-SVM) or logistic regression (HR-LR). As opposed to previous hierarchical classification models, incorporating dependencies into the regularization term rather than constraints reduces the dependency between the parameters, which in turn allows produces a highly parallelizable optimization problem. For general graphs, the algorithm can be parallelized up to the chromatic number of the graph, which reduces to alternatively optimizing odd and even layers, in the case of hierarchial models. Empirically, HR-SVM consistently outperforms HR-LR, as well as several established benchmarks on some of the largest datasets that have been studied.

General conclusions/observations:

- Hierarchies are effective, efficient organizational structures for basic supervised and unsupervised machine learning problems, and can make even simple (e.g., linear) models quite powerful.
- Again, variational Bayesian inference provides a powerful alternative to EM.
- Regularization/priors are very important in large-scale clustering/classification tasks where many (or even the majority) of classes have only a few instances.

3 Semi-Supervised Clustering and Metric Learning

Papers: [12, 2, 7]

These papers study the problem of learning an appropriate metric (for K-means clustering or otherwise) using information about the similarity of a small number of given data points (i.e., limited supervision). They all operate on n data points in $\mathbb{R}^d$, and take in supervision in the form of two sets, a "must-link" set $\mathcal{M}$ of pairs of points that should lie in the same cluster and a "cannot-link" set $\mathcal{C}$ of pairs of points that should lie in different clusters.

Older work largely follows one of two approaches to using supervision to improve clustering: Probabilistic Constrained Clustering (PCC) and Distance Metric Learning (DML). The first two papers, [12] and [2], are of the latter (DML) approach, which attempts to learn a metric under which the distances are reduced between pairs in $\mathcal{M}$ and increased between pairs in $\mathcal{C}$. Specifically, [12] learns the correlation matrix A of a metric of the form $d_A(x, y) = \sqrt{(x-y)^T A(x-y)}$ by solving the optimization problem

$$\begin{aligned} \min_A \quad & \sum_{(x,y)\in\mathcal{M}} d_A^2(x,y) \\ \text{such that} \quad & \sum_{(x,y)\in\mathcal{C}} d_A^2(x,y) \geq 1 \\ \text{and} \quad & A \succeq 0. \end{aligned}$$

This step is performed as a preprocessing step, and K-means is then run on the data with the resulting metric. [2] modifies this by combining the metric learning and clustering steps. Specifically, in each iteration, for each pair $(x, y) \in \mathcal{M}$ that are placed in different clusters, the objective includes a penalty increasing with $d_A(x, y)$, and, similarly, for each pair $(x, y) \in \mathcal{C}$ that are placed in the same cluster, the objective includes a penalty which decreases with $d_A(x, y)$. Their algorithm then alternates between optimizing the objective over the clustering assignments and the covariance A of the metric.

The third paper, [7], argues that DML approaches lead to overfitting, because even well clustered data will be transformed so as to exaggerate certain distances, and also that, because the metric learning step is typically ignorant of the clustering objective ([2] notwithstanding), it need not learn a metric that is necessarily optimal for clustering. Their approach is to learn a transformation of the data that maps points in $\mathcal{M}$ and $\mathcal{C}$ in naturally well-separated clusters according to a specified clustering objective. Their framework is general, but is worked out for the case of Gaussian or vMF models.

General conclusions/observations:

- Metric learning is strongly dependent on the task at hand; e.g., it is difficult to define a "good" metric without a goal, such as clustering, in mind.
- Even more specifically, in the case of semi-supervised clustering, it helps to know the rough form of the clusters, or at least of the clustering objective or model (e.g., Gaussians Mixture, von Mises-Fisher, etc.) being used.

4 Large-Scale Matrix Factorization

Papers: [10, 11, 5]

Given an data matrix $X \in \mathbb{R}^{n\times d}$, these papers seek factor matrices $W \in \mathbb{R}^{n\times k}$ and $H \in \mathbb{R}^{k\times d}$ such that $X \approx WH$. Typically, $k < d$, so that the factorization implicitly performs dimension reduction, and it is also desirable for W and H to be sparse.

One approach is non-negative matrix factorization (NMF), where X has non-negative entries, and we want W and H to also have non-negative entries. In principle, this gives an additive parts-based decomposition of components (e.g. decomposing portrait pictures into face and hair). Because it gives a parts-based decomposition, NMF tends naturally to give sparse solutions, but this sparsenss is intrinsic rather than being controlled by a predefined parameter. Hence, defining sparseness as in terms of the ℓ_1 and ℓ_2 norms,

$$\text{sp}(x) = \frac{\sqrt{n} - \|x\|_1/\|x\|_2}{\sqrt{n}-1},$$

[10] studies NMF with constraints on the sparseness sp(w_i) and sp(h_i) of the columns of W and the rows of H. They use a multiplicative optimization procedure studied previously for NMF and project in each iteration to satisfy the desired sparsity constraint.

A second approach is Structured Sparse Principal Component Analysis (SSPCA) [11]. PCA can be thought of as learning an orthogonal basis, or dictionary, which explains the data with low reconstruction error. As in LASSO and group LASSO, this dictionary can be sparsified by adding an appropriate regularization term (e.g., an ℓ_1 norm), and structured sparsity can be enforced with a grouped regularization term (e.g., an ℓ_1 norm of ℓ_2 norms of the appropriate groups). [11] propose an optimization scheme which cycles through optimizing over W, H, and the groups weights. While NMF finds a sparse dictionary, its elements are not necessarity structured. On the other hand, given a dataset of face images, SSPCA finds elements such as eyes and mouths; that is, in this context, SSPCA identifies local features.

Finally, [5] studies a general distributed scheme, distributed stochastic gradient descent (DSGD), for matrix factorization via stochastic gradient descent at very large scale. DSGD can be adapted for NMF as well as for other matrix factorization problems. DSGD is compatible with MapReduce, and provably converges under certain conditions. Furthermore, experiments show DSGD significantly outperforming other state-of-the-art algorithms for very large matrix factorization problems (e.g., Netflix dataset).

General conclusions/observations:

- Several basic problems for large datasets can be phrased in terms of matrix factorizations (e.g., variants of PCA, NFM, etc.).
- Matrix factorization algorithms can be effectively distributed and performed very efficiently at a large scale.

References

[1] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *JMLR*, 6:1345138, 2005.

[2] M. Bilenko, S. Basu, and R. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *International Conference on Machine Learning (ICML)*, 2004.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. In *Neural Information Processing Systems (NIPS)*, 2003.

[5] R. Gemulla, E. Nijkamp, P. Haas, and Y. Sismanis. Large-scale matrix factorization with distributed stochastic gradient descent. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 69–77, New York, NY, USA, 2011. ACM.

[6] S. Gopal and Y. Yang. Recursive regularization for large-scale classification with hierarchical and graphical dependencies. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 257–265, New York, NY, USA, 2013. ACM.

[7] S. Gopal and Y. Yang. Transformation-based probabilistic clustering with supervision. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

[8] S. Gopal and Y. Yang. Von mises-fisher clustering models. In *International Conference on Machine Learning (ICML)*, 2014.

[9] S. Gopal, Y. Yang, B. Bai, and A. Niculescu-Mizil. Bayesian models for large-scale hierarchical classification. In *Neural Information Processing Systems (NIPS)*, 2012.

[10] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:14571469, 2004.

[11] R. Jenatton, G. Obozinski, and F. Bach. Structured sparse principal component analysis. In *International Conference on AI and Statistics (AISTATS)*, volume 9 of *JMLR Workshop and Conference Proceedings*, 2010.

[12] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Neural Information Processing Systems (NIPS)*, 2002.