10-708: Probabilistic Graphical Models, Spring 2015

22 : Optimization and GMs (aka. "LDA, Sparse Coding, Matrix Factorization, and All That")

Lecturer: Yaoliang Yu *Scribes: Yu-Xiang Wang, Su Zhou*

1 Introduction

Graphical models are cool but hard to explain in layman's words. Matrix factorizations are simple and you can even tell your grandmother about it. They are often intimately related and the interface of the two subcommunities can borrow ideas from one field and apply to another. In additional most results can be thought of as removing certain restrictions and adding additional flexibility to the the naive matrix factorization.

2 Topic models and LSI

Long before Latent Dirichlet Allocation (LDA), Latent Semantic Indexing has been very popular. It is essentially running SVD on the document matrix with bag of words models $X = U\Lambda V^T$, where the inner dimension k is the number of topics. The associated optimization problem can be written as

$$\min_{U,V} \|X - UV^T\|^2$$

where U and V are not entirely identifiable (since we can take any invertible transformation of the rows of U and V without affecting the objective function). However, it is not hard to see that SVD recovers an optimal solution.

The model is very simple but has some cool applications including:

- cross-language retrieval (by concatenate features of different language in X and find co-shared U). TOEFL/GRE synonym in the same way.
- Matching paper submission to reviewers.

3 Exponential family PCA

LSI can be generalized into Generalized linear losses via a probabilistic model.

$$X \sim Pr(\cdot|\Theta)\Theta = UV^T$$

when $Pr(\cdot|\Theta)$ is Gaussian, the reduces to standard matrix factorization. Here the difference is to consider exponential family.

The MLE can also be represented in a matrix factorization form

$$\min_{U,V} -\langle X, UV^T\rangle + G(UV^T)$$

where G is the log-partition function. It is strictly convex and has analytic form. Also note that the maximization of the log-likelihood is simply the Fenchel conjugate. The whole thing is bi-strictly convex and alternating minimization is guaranteed to converge to a stationary point.

For example, in Poisson model the optimization is an unconstrained optimization which looks like

$$\min_{U,V} \sum_{w,d} -X_{w,d}[UV^T]_{w,d} + \exp([UV^T]_{w,d})$$

4 Nonnegative matrix factorization

Another widely accepted method is Nonnegative Matrix Factorization (NMF). It gains popularity because it simply makes a lot of sense to assume the factors are non-negative. Also it has an interesting multiplicative update algorithm with the property that all 0 will remain at 0. This is a form of gradient descent with an adaptively chosen stepsize (refer to the notes for the update equations).

$$\min_{U\geq 0, V\geq 0} -\langle X, UV^T\rangle + G(UV^T)$$

NMF is mainly used for feature learning on image data. Due to the nonnegativity constraint, the solutions are likely be sparse and localized and that corresponds very well to parts in an image or arguably regions of the brain. There is also a KL-divergence version of the NMF, which solves

$$\min_{U\geq 0, V\geq 0} \sum_{w,d} -X_{w,d} \log[UV^T]_{w,d} + [UV^T]_{w,d}$$

If you compare the above KL-divergence version of the NMF and the Poisson PCA, you will see that the forms are very similar. There are some differences though as one cannot be simply reparameterized as another. KL-divergence NMF's low-rank matrix is linear in the exponentiated space, while in Poisson PCA, the low-rank factorized matrix is linear in the original logarithmic space.

NMF leads to sparse solutions mainly because they have a nonnegativity constraint. The solution will be 0 when certain constraints are active.

Since the algorithm will shrink some coordinates to zero and they stay at zero, one should always randomly initialize dense factor U and V to avoid trivial solutions. In particular, $U = 0$ and $V = 0$ is a trivial stationary point.

5 Polysemy, Synonymy, pLSI and LDA

LSI is good ay synonym but less so at Polysemy. So people introduce the notion of a topic where different topic can share the same words but they may mean different things in different topics. This leads to probabilistic LSI (Hoffman ML'01). LDA is essentially a Bayesian version of pLSI with an additional Dirichlet prior. It is claimed that this seemingly minor change actually avoids probabilistic LSI from overfitting to data.

The maximum-likelihood problem for pLSI turns out to be

$$\begin{aligned} \min_{U,V} & \sum_{w,d} -X_{w,d} \log[UV^T]_{w,d} \\ s.t. \quad & U, V \geq 0, U^T 1 = 1, V^T 1 = 1 \end{aligned}$$

Comparing the previous KL-Divergence version of NMF, the only difference is that now it has an additional constraint that each row of U and V must be in a probability simplex. This adds a semantic meaning to the factors.

6 Revisiting PCA

(Deterministic) Principal Component Analysis (PCA) enjoys a set of salient properties: Orthogonal basis, implicit Gaussian assumption, 2nd order decoupling (decouping of PCA) and in some sense, an assumption that the data is "noiseless" (we will see what it means later).

"These are not necessarily the best way of doing dimension reduction.", some smart folks thought. Then they come up with extensions of PCA which features the following:

1. Sparse coding with overcomplete basis.
2. Exponential family PCA/Kernel PCA with non-linear basis.
3. Independent Component Analysis that allows for high order de-correlation.
4. Probabilistic PCA that addresses additive noise to the model.

The remaining parts of Yaoliang's lecture focused on pPCA and Sparse Coding. The second and third item is left as supplementary readings.

7 Probabilistic PCA

'probablistic PCA" Assume a generative model.

$$X_i|V \sim \mathcal{N}(Uv_i, \sigma^2 I), \quad v_i \sim \mathcal{N}(0, \sigma^2 I)$$

for $i = 1, ..., n$. The maximum likelihood optimization problem is

$$\min_{U,\sigma^2} \log\det(UU^T + \sigma^2 I) + \langle S, (UU^T + \sigma^2 I)^{-1} \rangle$$

where S is the sample covariance matrix.

This is a highly non-convex problem, since $\log\det$ is a concave function and UU^T is essentially a rank-constraint. Quite remarkably, we can solve the above maximum likelihood problem in closed form via SVD of $S = U^* \mathrm{diag}([s^k]_{k=1}^W) V^*$.

The optimal solution

$$U = U^* \mathrm{diag}([\sqrt{(s_k^2 - \sigma^2)}]_{k=1}^K)$$

where $\sigma^2 = \frac{1}{W-K} \sum_{k=K+1}^W s_k^2$.

This is proven by the Von Neuman's trace inequality on the trace of the product of two matrices (see http://en.wikipedia.org/wiki/Von_Neumann%27s_trace_inequality) that says

To see this, first note that we can replace UU^T by a X with row space and column space U^*., since adding any component orthogonal to U^* will only increase the objective function. Now the problem reduces to vector problems on the diagonals, which solves

$$\min_{\sigma_k^2, \sigma^2} \sum \log(\sigma_k^2 + \sigma^2) + \frac{s_k^2}{\sigma^2 + \sigma^2} + \sum_{k=K+1}^W \log\sigma^2 + \frac{s_k^2}{\sigma^2}.$$

The same reduction can also be seen from Von Neuman's trace inequality

$$\sum_i \sigma_{w-i}(A)\sigma_i(B) \leq < A, B > \leq \sum_i \sigma_i(A)\sigma_i(B).$$

where we replace the second term in the maximum likelihood objective with its lower bound. The lower bound can be attained when we take UU^T to have the range that matches the learning K-dimensional eigenspace of S. Then the solution to the spectrum can be obtained by simply setting derivative to 0.

Note that the under the assumption that the data is contaminated by noise, the solution is no longer rank-projection, but rather the solution to

$$\min_X \|S - X\|_2^2 + \lambda \|X\|_*$$

with an adaptively chosen regularization weight λ using the remaining described spectrum of S.

It is claimed in the slides that the pPCA recovers PCA when $\sigma = 0$. This is true but it's pretty meaningless because if the true parameter $\sigma^2 = 0$, S is always rank K and it doesn't make sense to do dimension reduction anyway.

Since the probabilistic model comes out, there are many extensions to this. We refer readers to Slide 21 for the references.

One notable connection is that if one applies pPCA to count data (Multinomial rather than normal) with a Dirichlet prior, one gets back LDA with Z marginalized out. However, we can no longer marginalize out V as in the Gaussian case.

8 Sparsity and Choosing K

The remaining issues involve getting exact sparsity and choosing number of topics K from data. LDA's Dirichlet prior, and other Bayesian models (even with shrinkage) cannot get exact sparsity.

In general, the posterior mean estimators is rarely sparse but the maximum a posteriori (MAP) estimators can be sparse. The notes suggest that MAP estimator is not Bayesian hence unsatisfactory. Arguably, however, both MAP and posterior mean estimators can be Bayes estimator for the same prior from decision-theoretic point of view. They just correspond to different loss functions. MAP corresponds to L1 loss, while posterior mean corresponds to the square loss.

The actual solution to this issue in Bayesian modelling often requires the not-so-pretty use of slab-and-spike type of priors that assigns point mass at sparse solutions.

Somehow, the lecture only uses sparsity to motivate the discussion of sparse coding, where sparsity is often induces via an L1 penalty. This is what the next section is about.

As for the model selection problem of choosing K, usually people use Bayesian Non-parametric, but that is beyond the scope of this lecture. We will sort of address this problem by showing that some norm-regularized matrix factorization actually corresponds to infinite sparse coding where we do not need to specify K ahead of time. This is what the next next section is about.

9 Sparse coding

Assume factorization model $X = UV^T$. Wavelets, fourier, random projection can be thought of having a fixed dictionary U and the problem is finding V. In sparse coding, we solve for U and V at the same time.

Specifically, a commonly accepted regime is when U is overcomplete and V is sparse. "Overcomplete" means that U has more columns than rows and one can think of U as a dictionary of parts and every data point is a sparse linear combination of the parts. The more overcomplete U is the sparser V should be.

Variants of this models for exponential family, to structure sparsity and so forth have been proposed too and this is also stacked into deep neural networks. A general formulation can be written as

$$\min_{U,V} \ell(X, UV^T) + \lambda \sum_{d,k} g(V_{d,k}) \tag{1}$$

for some sparse inducing norms. Often U needs to have bounded norm otherwise V can be set to arbitrarily near 0 and the solution is ill-posed.

Solution to such problem is often done by alternating minimization, e.g., k-SVD. Provable guarantees are still an active area of research (search on "provable dictionary learning", "provable overcomplete dictionary learning", "provable sparse coding" and etc).

10 Infinite sparse coding

An interesting result for sparse coding occurs when we take K to infinity. Consider the following special case of (1).

$$\min_{\Theta} \ell(X, \Theta) + \lambda \min_{\Theta = UV^T, |U_{:,k}| \leq 1 \sum_{k=1}^{K} \|V_{:,k}\|} \tag{2}$$

For finite K this is a non-convex problem in Θ but as $K \to \infty$,

$$|||\Theta||| := \min_{\Theta = UV^T, |U_{:,k}| \leq 1 \sum_{k=1}^{K} \|V_{:,k}\|}$$

is a norm. The dual norm

$$|||\Theta|||^o = \max\{u^T \Theta v : |u| \leq 1, \|v\| \leq 1\}$$

Specifically, when $|u| = \|u\|_2, \|v\| = \|v\|_2$, the $|||\Theta|||^o = \|\Theta\|_2$ the spectral/operator norm. Thus, $|||\Theta||| = \|\Theta\|_*$ gets back the nuclear norm, or the trace norm. The essential rank will be obtained adaptively in some sense. But there is still one parameter λ to tune.

11 Extensions to these ideas

Like we discussed previously, similar ideas have been exploited a lot in both matrix factorization and graphical modelling. Zhu & Xing proposed Sparse Topic Modelling which is discriminative rather than generative. Zhu and Xing paper has many more parameters than observations. There fore they add strong regularizers with l1 norms and Θ is actuallly clusters of S. The tweak apparently worked pretty well (better than LDA) for topic modelling. At the same time, because it is not sampling or computing posterior mean, the $L1$ norm will give exact sparsity.

The ideas of using a latent low-dimensional subspace to represent high-dimensional data got really popular in the past decades because they actually work very well in computer vision data.

Any combinations of the keywords: "Supervised", "predictive", "sparse", "large-margin" and etc, might have generated a paper. It is like these methods are occupying the world.

Nowadays, people often stop modelling probabilistically and directly model the optimization objective to enjoy the structural implication of the different regularizer. However, this approach is less principled and does not often lead to an easy construction of confidence/credibility intervals, which makes inference using these models often challenging (if not intractable). For predictive tasks, it does not matter much since one can always evaluate on the hold-out data, but for scientific discoveries, probabilistic inferences are of great importance. This is probably the reason why courses like probabilistic graphical model still exists and are this popular.