10708 Graphical Models: Homework 4

Due Monday, April 22, beginning of class

April 3, 2013

Instructions: There are four questions on this assignment. There is a problem involves coding. You can program in whatever language you like, although we suggest MATLAB. Do *not* attach your code to the writeup. Instead, put your code in a directory called "andrewid-HW4" and tar it into a tgz named "andrewid-HW4". For example, epxing-HW4.tgz. Email your tgz file ONLY to gunhee@cs.cmu.edu, seunghak@cs.cmu.edu and kpuniyan@cs.cmu.edu. Refer to the web page for the policies regarding collaboration, due dates, extensions, and late days.

1 Hilbert Space Embeddings [30 points]

We discussed in class that Hilbert Space Embeddings are attractive because certain probability "rules" also hold for the analogous RKHS operators. In class we discussed the RKHS version of the sum rule. It is highly recommended you fully understand the proof of the RKHS sum rule (in Lecture 20) before doing this question.

Let A, B, and C be random variables. In this question you will prove that if $C \perp A | B$ then

$$\boldsymbol{\mathcal{C}}_{CA} = \boldsymbol{\mathcal{C}}_{C|B}\boldsymbol{\mathcal{C}}_{B|A}\boldsymbol{\mathcal{C}}_{AA}$$

This is the RKHS analog to $\mathbb{P}[C, A] = \sum_B \mathbb{P}[C|B]\mathbb{P}[B|A]\mathbb{P}[A]$ if $C \perp A|B$.

We will assume that all three random variables A, B, C are embedded in RKHS $\mathcal{F}$. The corresponding feature functions for $\mathcal{F}$ will be indexed by ϕ. Thus, just to clarify notations/definitions:

$$\begin{aligned}
\boldsymbol{\mathcal{C}}_{CA} &= \mathbb{E}_{CA}[\phi_C \otimes \phi_A] && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[C,A] \\
\boldsymbol{\mathcal{C}}_{CB} &= \mathbb{E}_{CB}[\phi_C \otimes \phi_B] && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[C,B] \\
\boldsymbol{\mathcal{C}}_{BA} &= \mathbb{E}_{BA}[\phi_B \otimes \phi_A] && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[B,A] \\
\boldsymbol{\mathcal{C}}_{AA} &= \mathbb{E}_{AA}[\phi_A \otimes \phi_A] && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[A,A] \\
\boldsymbol{\mathcal{C}}_{BB} &= \mathbb{E}_{BB}[\phi_B \otimes \phi_B] && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[B,B] \\
\boldsymbol{\mathcal{C}}_{C|B} &= \boldsymbol{\mathcal{C}}_{CB}\boldsymbol{\mathcal{C}}_{BB}^{-1} && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[C|B] \\
\boldsymbol{\mathcal{C}}_{C|A} &= \boldsymbol{\mathcal{C}}_{CA}\boldsymbol{\mathcal{C}}_{AA}^{-1} && \text{RKHS analog of } \boldsymbol{\mathcal{P}}[C|A]
\end{aligned}$$

1. First you will prove that $\boldsymbol{\mathcal{C}}_{BA} = \boldsymbol{\mathcal{C}}_{B|A}\boldsymbol{\mathcal{C}}_{AA}$ (the RKHS analog of $\mathbb{P}[B,A] = \mathbb{P}[B|A]\mathbb{P}[A]$.
   (a) Write the rule $P[B,A] = \mathbb{P}[B|A]\mathbb{P}[A]$ in matrix form.
   (b) Prove that $\mathbb{P}[B,A] = \mathbb{P}[B|A]\mathbb{P}[A]$ using expectations and δ indicator vectors (as done on Lecture 20 Slide 17 for the sum rule).
   (c) Now prove the RKHS version: $\mathcal{C}_{BA} = \mathcal{C}_{B|A}\mathcal{C}_{AA}$
2. Now you will prove that if $C \perp A|B$, then $\mathcal{C}_{CA} = \mathcal{C}_{C|B}\mathcal{C}_{BA}$ (the RKHS analog of $\mathbb{P}[C,A] = \sum_B \mathbb{P}[C|B]\mathbb{P}[B,A]$.
   (a) Write the rule $P[C,A] = \sum_B \mathbb{P}[C|B]\mathbb{P}[B,A]$ in matrix form.
   (b) Prove that $P[C,A] = \sum_B \mathbb{P}[C|B]\mathbb{P}[B,A]$ using expectations and δ indicator vectors (as done on Lecture 20 Slide 17 for the sum rule). Indicate in your proof in what step you have used the conditional independence assumption.
   (c) Now prove the RKHS version: $\boldsymbol{\mathcal{C}}_{CA} = \boldsymbol{\mathcal{C}}_{C|B}\boldsymbol{\mathcal{C}}_{BA}$. Indicate in your proof in what step you have used the conditional independence assumption.

2 Expectation of Dirichlet distribution [15 points]

The Dirichlet distribution is a continuous distribution on the K-simplex, $\{\theta = (\theta_1, \ldots, \theta_K)$, such that $\theta_i \geq 0$ for $i = 1, \ldots, K$, and $\sum_{i=1}^K \theta_i = 1\}$:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^K \alpha_i)}{\prod_{i=1}^K \Gamma(\alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}$$

where $\alpha_i > 0$ are parameters. Compute the followings with full derivation steps. (Giving only final answers will get no point).

1. (5 pts) $E[\theta_k]$.
2. (5 pts) $Cov[\theta_j \theta_k]$.

3. (5 pts) $E[\log \theta_k]$. (*Hint*: First show that the Dirichlet distribution is in the exponential family, and take a first derivative of the cumulant function. You can use the digamma function $\psi(\theta_k) = \frac{\partial}{\partial \theta_k} \log \Gamma(\theta_k)$).

3 Posterior of Dirichlet Process [15 points]

1. (5pt) A multinomial distribution with a vector $\theta = (\theta_1, \ldots, \theta_K)$, where $\theta_i \geq 0$ for $i = 1, \ldots, K$, and $\sum_{i=1}^{K} \theta_i = 1$, is represented as follows.

$$p(x) = \frac{n!}{\prod_{i=1}^{K} x_i!} \prod_{i=1}^{K} \theta_i^{x_i}.$$

Show that Dirichlet distribution $\theta \sim Dir(\alpha)$ (given in problem 1) is a conjugate prior of $p(x|\theta) \sim Multi(\theta)$.

2. (10pt) Let H be a distribution over Θ and α is a positive real number. For any finite measurable partition $A_1, \ldots, A_r$ of Θ, G is called a Dirichlet process with base distribution H and concentration parameter α, denoted by $G \sim DP(\alpha, H)$, if

$$(G(A_1), \ldots, G(A_r)) \sim Dir(\alpha H(A_1), \ldots, \alpha H(A_r)).$$

Suppose we have observed values $X_1, \ldots, X_n$ from G, and let $n_s = \#\{i : X_i \in A_s\}$ be the number of observed values in A_s. If we have $DP(\alpha, H)$ as prior, derive the posterior $G|X_1, \ldots, X_n$.

4 Structured Sparsity [40 points]

In this problem, we will implement sparse group lasso which optimizes the following:

$$\min_{\beta_1,\ldots,\beta_P} \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{P} x_i^j \beta_j \right)^2 + \lambda_1 \sum_{j=1}^{P} |\beta_j| + \lambda_2 \sum_{l=1}^{L} \sqrt{\sum_{j \in \boldsymbol{g}_l} \beta_j^2}. \tag{1}$$

Here input data is $\boldsymbol{X}$ which is N by P matrix, where N is the number of samples, and P is the number of features, and output data is $\boldsymbol{y} = [y_1, \ldots, y_N]^T$ which is N by 1 vector. We denote row index by subscript and column index by superscript. For example, x_i^j represents the element of $\boldsymbol{X}$ at the i-th row and j-th column. Similarly, y_i represents the i-th sample in $\boldsymbol{y}$.

In problem (1), to enforce group structured sparsity, we have $\lambda_2 \sum_{l=1}^{L} \sqrt{\sum_{j \in \boldsymbol{g}_l} \beta_j^2}$, where L is the number of feature groups and $\boldsymbol{g}_l$ represents the l-th feature group. For example, given 9

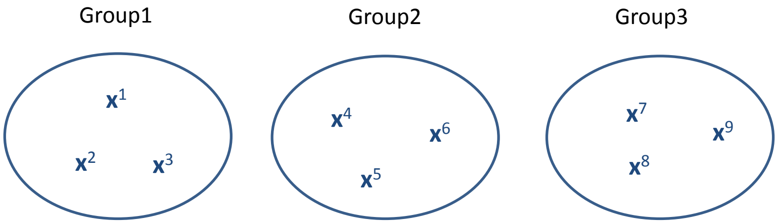

Figure 1: A example of groups of features denoted by $\boldsymbol{g}_1 = \{1, 2, 3\}$, $\boldsymbol{g}_2 = \{4, 5, 6\}$, and $\boldsymbol{g}_3 = \{7, 8, 9\}$.

input variables, we may have three groups of features $\boldsymbol{G} = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3\}$, where $\boldsymbol{g}_1 = \{1, 2, 3\}$, $\boldsymbol{g}_2 = \{4, 5, 6\}$, and $\boldsymbol{g}_3 = \{7, 8, 9\}$ (Note: $\boldsymbol{g}_l$ includes feature indices for the l-th group; and there is no overlap between different groups. See Figure 1).

Let us derive a block coordinate descent algorithm to optimize the problem 1, and then implement it.

1. (5pt) First, derive an optimality condition when $\beta_j = 0\ \forall j \in \boldsymbol{g}_l$ fixing all the other coefficients. [Hint: Use subgradient of the objective in (1) with respect to $\boldsymbol{\beta}_{\boldsymbol{g}_l}$.]
2. (5pt) If the optimality condition you just derived in (1) is satisfied, we can set $\beta_j = 0\ \forall j \in \boldsymbol{g}_l$. Now, suppose that the optimality condition you derived in (1) is NOT satisfied. Derive an optimality condition when $\beta_j = 0$ (i.e., j-th element in $\boldsymbol{g}_l$ is zero). [Hint: Use subgradient of the objective in (1) with respect to β_j.]
3. (5pt) If the optimality condition you just derived in (2) is satisfied, we can set $\beta_j = 0$. Now, suppose that the optimality condition you derived in (2) is NOT satisfied. Show your solution to find the value of β_j when $\beta_j \neq 0$ (i.e., j-th element in $\boldsymbol{g}_l$ is non-zero).

Now you are ready to implement block coordinate descent algorithm to solve problem (1). Implement sparse group lasso using block coordinate descent algorithm. Then run your code on the data downloaded from the class website. (It includes $\boldsymbol{X}$ for input, $\boldsymbol{y}$ for output, $\boldsymbol{G}$ for the groups of feature indices, and readme.txt.)

4. (10pt) Run your code with $\lambda_1 = \lambda_2 = [0.1, 0.01, 0.001, 0.0001]$, and then draw a plot for the number of non-zero coefficients against λ_1 parameters. (In your plot, X-axis represents λ_1 parameters, and Y-axis represents the number of non-zero coefficients, that is $|\{j : \beta_j \neq 0\ \forall j\}|$.
5. (10pt) Report objective function values in (1) you achieved with $\lambda_1 = \lambda_2 = [0.1, 0.01, 0.001, 0.0001]$.
6. (5pt) Report the number of groups having non-zero coefficients when $\lambda_1 = \lambda_2 = 0.001$. Discuss about structured sparsity which are induced by your sparse group lasso implementation.

[References]

- N. Simon, et al. A Sparse-group Lasso. Journal of Computational and Graphical Statistics, 2012.
- J. Friedman et al. A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736, 2010.