EPFL | MGT-418 : Convex Optimization | Project 1

Questions – Fall 2021

Sparse Graphical Models for Binary Variables

Description

We consider the problem of estimating an undirected graphical model for multivariate binary data. More precisely, we aim to estimate the probability mass function $p(x)$ of a binary random vector $X = (X_1, \ldots, X_n)^\top \in \{-1, 1\}^n$ based on m training samples $\hat{x}^k \in \{-1, 1\}^n$, $k = 1, \ldots, m$, drawn independently from $p(x)$.

As an example, we might want to learn semantic connections between different words indexed by $i = 1, \ldots, n$. In this context, the random variable X_i indicates whether word i appears in a random document (*i.e.*, $X_i = 1$ if word i appears and $X_i = -1$ otherwise) and the training samples represent m randomly chosen documents (*i.e.*, $\hat{x}_i^k = 1$ if word i appears in document k and $\hat{x}_i^k = -1$ otherwise).

An Ising model for the probability mass function $p(x)$ is given by

$$p_\theta(x) = \exp\Bigg(\sum_{i=1}^n \theta_{ii} x_i + \sum_{\substack{i,j=1 \\ j \neq i}}^n x_i \theta_{ij} x_j - A(\theta) \Bigg),$$

where $\theta \in \mathbb{S}^n$ and

$$A(\theta) = \log \Bigg(\sum_{x \in \{-1,1\}^n} \exp\Bigg(\sum_{i=1}^n \theta_{ii} x_i + \sum_{\substack{i,j=1 \\ j \neq i}}^n x_i \theta_{ij} x_j \Bigg) \Bigg)$$

is a normalization constant. The parameters $\theta_{ij} = \theta_{ji}$, $i \neq j$, capture the dependence between the random variables X_i and X_j. In fact, one can show that if $\theta_{ij} = 0$, then X_i and X_j are conditionally independent given all other random variables X_k, $k \in \{1, \ldots, n\} \setminus \{i, j\}$.

The Ising model can conveniently be represented as a conditional dependence graph, where the nodes correspond to the different random variables X_i, $i = 1, \ldots, n$. We connect two random variables X_i and X_j with an edge whenever $\theta_{ij} \neq 0$, and θ_{ij} can be viewed as the weight of the edge.

The likelihood to observe the training samples $\hat{x}^k$, $k = 1, \ldots, m$, is given by $l(\theta) = \prod_{k=1}^m p_\theta(\hat{x}^k)$, while the log-likelihood function can be represented as

$$L(\theta) = \log\big(l(\theta)\big) = \sum_{k=1}^m \log\big(p_\theta(\hat{x}^k)\big) = \sum_{k=1}^m \Bigg(\sum_{i=1}^n \theta_{ii} \hat{x}_i^k + \sum_{\substack{i,j=1 \\ j \neq i}}^n \hat{x}_i^k \theta_{ij} \hat{x}_j^k \Bigg) - mA(\theta).$$

In this exercise, we are interested in estimating a sparse graphical model, that is, we wish to identify a matrix $\theta \in \mathbb{S}^n$ that has a high empirical (log-) likelihood and only few non-zero off-diagonal entries, *i.e.*, the corresponding conditional dependence graph should have few edges. To this end, we will minimize the sum of the negative empirical log-likelihood and a sparsity-inducing ℓ_1-norm penalty term $\rho \, \|\theta - \mathrm{diag}(\theta_{11}, \ldots, \theta_{nn})\|_1$, where $\rho \geq 0$ is the penalty weight, $\|Q\|_1 := \sum_{i=1}^n \sum_{j=1}^n |Q_{ij}|$ for an arbitrary matrix $Q \in \mathbb{S}^n$, and $\mathrm{diag}(\theta_{11}, \ldots, \theta_{nn})$ represents the diagonal matrix with diagonal entries $\theta_{11}, \ldots, \theta_{nn}$. Thus, we aim to solve the convex optimization problem

$$\underset{\theta \in \mathbb{S}^n}{\text{minimize}} \quad A(\theta) - \Bigg(\sum_{i=1}^n \theta_{ii} \hat{\mu}_i + \sum_{\substack{i,j=1 \\ j \neq i}}^n \theta_{ij} \hat{s}_{ij} \Bigg) + \rho \, \|\theta - \mathrm{diag}(\theta_{11}, \ldots, \theta_{nn})\|_1 \,, \tag{1}$$

where we denote by $\hat{\mu} = \frac{1}{m} \sum_{k=1}^m \hat{x}^k$ the sample mean and by $\hat{s} = \frac{1}{m} \sum_{k=1}^m \hat{x}^k (\hat{x}^k)^\top$ the sample second-order moment matrix.

Even though problem (1) is convex, it is hard to solve because the normalization constant $A(\theta)$ involves a sum of 2^n terms. In [1, Lemma 5] it has been shown, however, that $A(\theta)$ admits a more tractable convex upper bound of the form

$$\bar{A}(\theta) = \frac{n}{2}\log(\frac{e\pi}{2}) - \frac{1}{2}(n+1) - \frac{1}{2}\left(\max_{v\in\mathbb{R}^{n+1}} v^\top r + \log\det\big(-(R(\theta)+\mathrm{diag}(v))\big)\right),$$

where e denotes the base of the natural logarithm, $r = (1, \frac{4}{3}, \ldots, \frac{4}{3}) \in \mathbb{R}^{n+1}$ and

$$R(\theta) = \begin{bmatrix} 0 & \theta_{11} & \theta_{22} & \ldots & \theta_{nn} \\ \theta_{11} & 0 & 2\theta_{12} & \ldots & 2\theta_{1n} \\ \theta_{22} & 2\theta_{21} & 0 & \ldots & 2\theta_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta_{nn} & 2\theta_{n1} & 2\theta_{n2} & \ldots & 0 \end{bmatrix}.$$

This result motivates us to solve the approximate estimation problem

$$\underset{\theta\in\mathbb{S}^n}{\text{minimize}} \quad \bar{A}(\theta) - \left(\sum_{i=1}^n \theta_{ii}\hat{\mu}_i + \sum_{\substack{i,j=1 \\ j\neq i}}^n \theta_{ij}\hat{s}_{ij}\right) + \rho\,\|\theta - \mathrm{diag}(\theta_{11},\ldots,\theta_{nn})\|_1\,. \tag{2}$$

Questions

1. Verify that problems (1) and (2) are convex.

2. The data file `20news_w60.mat` available from Moodle contains 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. Each newsgroup document consists of 60 binary entries that correspond to 60 distinct words. Solve problem (2) for this data set with $\rho = 10^{-3}, 5\times10^{-3}, 10^{-2}$ using the solver `sdpt3`, which is available from Moodle (see hint below). A skeleton of the code you will have to implement is provided in the Matlab file `p1q2.m`. The Matlab function `adj2gephilab.m` plots the resulting undirected graph and visualizes the sparsity of the optimal solution $\theta^\star$. It also saves two `csv` files, which can be used later to plot the graph with `gephi`, that is, a visualization and exploration software for graphs and networks.

   Hint: Download the file *sdpt3.zip* from Moodle, place it in your favorite directory and unzip it. Then, add the paths to this unzipped folder and its subfolders to Matlab.

Bonus Question (not related to optimization and therefore not examinable)

Prove that if $\theta_{ij} = \theta_{ji} = 0$, the random variables X_i and X_j are indeed conditionally independent given X_k, $k \in \{1,\ldots,n\}\setminus\{i,j\}$. Specifically, show that if $\theta_{ij} = \theta_{ji} = 0$, then

$$p_\theta(x_i, x_j | x_{-(i,j)}) = p_\theta(x_i|x_{-(i,j)})p_\theta(x_j|x_{-(i,j)}),$$

where $x_{-(i,j)}$ denotes the vector x without its i-th and j-th components.

References

[1] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.