Statistical learning: homework 1

October 2, 2014

Let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be an undirected graph with a set $\mathcal{V}$ of p vertices and a set $\mathcal{E}$ of edges. The **adjacency** matrix A of $\mathcal{G}$ is a $p \times p$ matrix with elements $a_{ij} = \mathbf{1}_{(v_i, v_j) \in \mathcal{E}}$, *i.e.*, elements 1 at coordinates corresponding to vertices connected by an edge, 0 otherwise. The **degree** matrix D of $\mathcal{G}$ is defined by $D = \mathrm{Diag}(A\mathbf{1}_p)$ where $\mathbf{1}_p$ is the all one vector in dimension p. In other words, D is a diagonal matrix with d_{ii} corresponding to the degree of vertex i. The **Laplacian** matrix $\mathcal{L}$ is defined by

$$\mathcal{L} = D - A.$$

The objective of this homework is to study some of its theoretical properties, and the empirical behavior of a penalized estimator based on a regularity measure defined using $\mathcal{L}$.

1 Analysis of the graph Laplacian penalty

1.1 A property of quadratic forms

Let $M \in \mathbb{R}^{p \times p}$ be a symmetric, positive semidefinite matrix ($M \succeq 0$), *i.e.*, such that $v^\top M v \geq 0 \quad \forall v \in \mathbb{R}^p$. Denote $M = U^\top \Lambda U$ its spectral decomposition: the columns of U are the eigenvectors of M and Λ is a diagonal matrix with the corresponding eigenvalues $\lambda_1 \geq \ldots \geq \lambda_p \geq 0$ on its diagonal.

We denote $\|v\|^2 = v^\top v = \sum_{j=1}^p v_j^2$ the squared Euclidean norm of $v \in \mathbb{R}^p$.

1.1.1 First eigenvector

Prove that

$$\left\{ \begin{array}{l} \max_{v \in \mathbb{R}^p} v^\top M v \\ \|v\|^2 = 1 \end{array} \right. = \lambda_1,$$

and that this value is reached for $v = u_1$.

You are advised **not** to use Lagrangian duality. Instead, you can use the following steps:

1. Prove that $v^\top M v = \alpha^\top \Lambda \alpha$ for some $\alpha \in \mathbb{R}^p$ with $\|\alpha\|^2 = 1$.

2. Prove that
$$\begin{cases} \max_{\alpha\in\mathbb{R}^p} \alpha^\top \Lambda \alpha \\ \|\alpha\|^2 = 1 \end{cases} = \lambda_1, \tag{P_1}$$
and deduce the optimal v.

1.1.2 Other eigenvectors

Prove that
$$\begin{cases} \max_{v\in\mathbb{R}^p} v^\top M v \\ \|v\|^2 = 1 \\ v \in \{v_1, \dots, v_{k-1}\}^\perp \end{cases} = \lambda_k, \tag{P_k}$$
where $\{v_1, \dots, v_{k-1}\}$ are argmax to problems $P_1, \dots, P_{k-1}$. and that this value is reached for $v = u_k$.

1.2 Quadratic forms with Laplacian matrices

We now consider the quadratic form obtained using the Laplacian matrix defined in the header of this homework.

1.2.1 Dirichlet's energy over $\mathcal{G}$

Prove that $v^\top \mathcal{L} v = \sum_{(v_i,v_j)\in\mathcal{E}} (v_i - v_j)^2$ for $v \in \mathbb{R}^p$.

$v^\top \mathcal{L} v$ is small if the values in v are smooth over the graph, *i.e.*, if connected nodes typically have similar values.

1.2.2 Using $\mathcal{L}$ for statistical inference

Assume we observe n samples $(x_i, y_i)_{i=1,\dots,n}$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. We denote $X \in \mathbb{R}^{n\times p}$ the matrix whose rows are x_i, and $Y \in \mathbb{R}^n$ the vector with values y_i. We consider the prediction function $f(x) = x^\top \hat{\beta}$, where $\hat{\beta}$ is an argmin of the following optimization problem:
$$\min_{\beta\in\mathbb{R}^p} L(X\beta, Y) + \mu_1 \beta^\top \mathcal{L} \beta + \mu_2 \|\beta\|^2, \tag{E}$$
for some $\mu_1, \mu_2 \in \mathbb{R}_*^+$.

1. What is the expected behavior of the estimator obtained by minimizing the empirical risk penalized by $v^\top \mathcal{L} v$?

2. Prove that (E) is equivalent to
$$\min_{\beta\in\mathbb{R}^p} L(\tilde{X}\beta, Y) + \|\beta\|^2, \tag{E'}$$
where $\tilde{X} = XB$, $B = (\mu_1 \mathcal{L} + \mu_2 I_p)^{-\frac{1}{2}}$.

3. Using the results of 1.1, what can you say about the energy $\Omega(x) = \mu_1 \beta^\top \mathcal{L} \beta + \mu_2 \|\beta\|^2$ of the eigenvectors of B?

4. Prove that the transformation B decreases the relative energy Ω in the following sense:

$$\Omega(\tilde{x})/\|\tilde{x}\|^2 \leq \Omega(x)/\|x\|^2,$$

where $\tilde{x} = Bx$.

2 Simulations

The following exercise can be done in your favorite programming language. If you want to use R (available for free at `http://cran.r-project.org/`), we provide a few useful primitives in the `hw1-help.R` file on the website `http://lear.inrialpes.fr/people/mairal/teaching/2014-2015/M2ENS/`.

1. Download the `hw1-adj.txt` file from the website `http://lear.inrialpes.fr/people/mairal/teaching/2014-2015/M2ENS/`. It contains the 50×50 adjacency matrix of a graph with 50 vertices. Load the matrix A and compute the Laplacian matrix of the associated graph and its spectral decomposition $\mathcal{L} = U \Lambda U^\top$.

2. Generate $n = 100$ (x_i, y_i) pairs under the model

$$y_i = x_i^\top \beta + \varepsilon_i,$$

where ε_i are independent indentically distributed from a normal distribution with 0 mean and variance 1 and x_i are real vectors in dimension $p = 50$. Do so for two different choices of β:

- $\beta_{low} = u_p$,
- $\beta_{high} = u_1$,

where u_k is the eigenvector associated with the k-th largest eigenvalue of $\mathcal{L}$.

3. Compute the ridge regression estimator

$$\hat{\beta}_{\text{ridge}}(X, Y) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(X\beta, Y) + \lambda \|\beta\|^2 = \left(X^\top X + \lambda I_p\right)^{-1} X^\top Y$$

over both training sets, for $\lambda = 100$ (to save your time, you are not asked to play with λ, this value is the best choice for this problem). Using the closed form for the ridge regression estimator and the equivalence between (E) and (E'), compute

$$\hat{\beta}_{\mathcal{L}}(X, Y) = \operatorname*{arg\,min}_{\beta \in \mathbb{R}^p} L(X\beta, Y) + \mu_1 \beta^\top \mathcal{L} \beta + \mu_2 \|\beta\|^2,$$

for $\mu_1 = \lambda$ and $\mu_2 = 0.1$, over both training sets.

4. Generate 10,000 new independent points[1] (x_i, y_i) under each of the two settings $(\beta_{low}, \beta_{high})$. For both estimators $(\hat{\beta}_{\text{ridge}}, \hat{\beta}_{\mathcal{L}})$ over both settings, compute the relative risk:

$$R(\hat{\beta}, \beta) = \|X_{\text{test}}\hat{\beta} - Y_{\text{test}}\|^2 / \|X_{\text{test}}\beta - Y_{\text{test}}\|^2.$$

   For comparison, also compute $R(0, \beta)$, the relative risk when predicting $y = 0$ for all x.

5. Discuss the four estimated $R(\hat{\beta}, \beta)$. When is $\hat{\beta}_{\mathcal{L}}$ better than $\hat{\beta}_{\text{ridge}}$, when is it worse, and why?

[1] You can reduce this number if the resulting computation is too heavy for your computer.