Math189R SP19
Homework 5
Monday, Mar 11, 2019

Feel free to work with other students, but make sure you write up the homework and code on your own (no copying homework *or* code; no pair programming). Feel free to ask students or instructors for help debugging code or whatever else, though.

1 (Murphy 12.5 - Deriving the Residual Error for PCA) It may be helpful to reference section 12.2.2 of Murphy.

(a) Prove that

$$\left\|\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j\right\|^2 = \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j.$$

Hint: first consider the case when $k = 2$. Use the fact that $\mathbf{v}_i^\top \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. Recall that $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$.

(b) Now show that

$$J_k = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j\right) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j.$$

Hint: recall that $\mathbf{v}_j^\top \mathbf{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

(c) If $k = d$ there is no truncation, so $J_d = 0$. Use this to show that the error from only using $k < d$ terms is given by

$$J_k = \sum_{j=k+1}^{d} \lambda_j.$$

Hint: partition the sum $\sum_{j=1}^{d} \lambda_j$ into $\sum_{j=1}^{k} \lambda_j$ and $\sum_{j=k+1}^{d} \lambda_j$.

(a)

$$
\begin{aligned}
\left\| x_i - \sum_{j=1}^{k} z_{ij} v_j \right\|^2 &= \left(x_i - \sum_{j=1}^{k} z_{ij} v_j \right)^\top \left(x_i - \sum_{j=1}^{k} z_{ij} v_j \right) \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \left(\sum_{j=1}^{k} z_{ij} v_j \right)^\top \left(\sum_{j=1}^{k} z_{ij} v_j \right) \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \sum_{j=1}^{k} \sum_{l=1}^{k} z_{ij} v_j^\top z_{il} v_l \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j \\
&= x_i^\top x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^\top x_i + \sum_{j=1}^{k} v_j^\top x_i x_i^\top v_j \quad (\text{since } v_j^\top v_i = 1 \text{ if } i = j) \\
&= x_i^\top x_i - \sum_{j=1}^{k} z_{ij} v_j^\top x_i v_j^\top,
\end{aligned}
$$

(b) By definition

$$
\begin{aligned}
J_k &= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^\top x_i - \sum_{j=1}^{k} z_{ij} v_j^\top x_i x_i^\top v_j \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^\top x_i - \sum_{j=1}^{k} v_j^\top \frac{1}{n} \left(\sum_{i=1}^{n} x_i x_i^\top \right) v_j \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{k} v_j^\top \Sigma v_j \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{k} \lambda_j,
\end{aligned}
$$

(c) Since $J_d = 0$, $\sum_{j=1}^{d} \lambda_j = \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i$. Then

$$
\begin{aligned}
J_k &= \frac{1}{n} \sum_{i=1}^{n} x_i^\top x_i - \sum_{j=1}^{d} \lambda_j + \sum_{j=k+1}^{d} \lambda_j \\
&= \sum_{j=k+1}^{d} \lambda_j.
\end{aligned}
$$

■

2 (ℓ_1-Regularization) Consider the ℓ_1 norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|.$$

Draw the norm-ball $B_k = \{\mathbf{x} : \|\mathbf{x}\|_1 \leq k\}$ for $k = 1$. On the same graph, draw the Euclidean norm-ball $A_k = \{\mathbf{x} : \|\mathbf{x}\|_2 \leq k\}$ for $k = 1$ behind the first plot. (Do not need to write any code, draw the graph by hand).

Show that the optimization problem

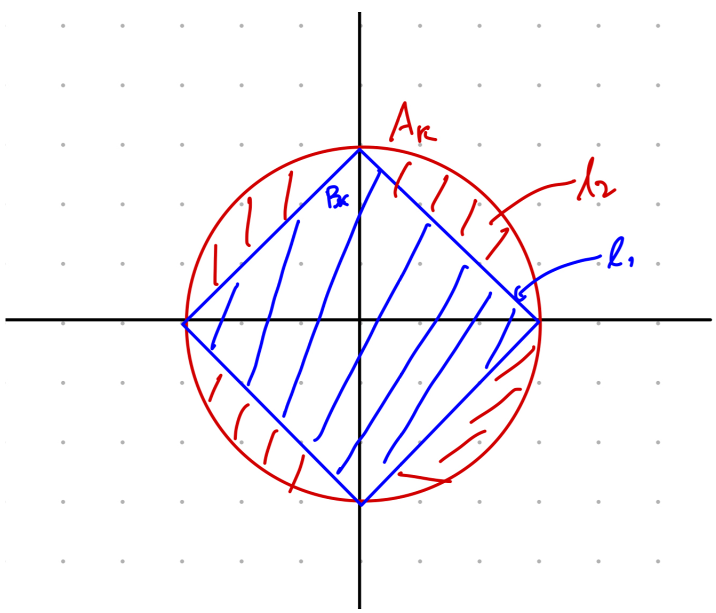
$$\begin{aligned} \text{minimize: } & f(\mathbf{x}) \\ \text{subj. to: } & \|\mathbf{x}\|_p \leq k \end{aligned}$$

is equivalent to

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p$$

(hint: create the Lagrangian). With this knowledge, and the plots given above, argue why using ℓ_1 regularization (adding a $\lambda\|\mathbf{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2 regularization for suitably large λ.

Drawing of the balls B_k and A_k:



We approach the optimization problem with the goal to minimize $f(x)$ subject to the constraint $\|x\|_p \leq k$. This is equivalent to the problem of finding the infimum over x and the supremum over $\lambda \geq 0$ of the Lagrangian $L(x, \lambda) = f(x) + \lambda(\|x\|_p - k)$.

The dual form allows us to exchange the infimum and supremum, expressed as:

$$\sup_{\lambda \geq 0} \inf_x \{f(x) + \lambda(\|x\|_p - k)\} = \sup_{\lambda \geq 0} g(\lambda)$$

The value of x that minimizes $f(x) + \lambda(\|x\|_p - k)$ will also be the minimizer for $f(x) + \lambda\|x\|_p$ since the term $-\lambda k$ is independent of x. Therefore, the optimization can be simplified to:

$$\text{minimize}\{f(x) + \lambda\|x\|_p\}$$

for an appropriate $\lambda \geq 0$.

Considering this in the context of ℓ_1 regularization, we interpret it as projecting the true optimal solution of the problem onto an ℓ_1 norm ball. The geometry of the ℓ_1 norm ball, characterized by its sharper vertices, increases the likelihood of the solution having elements that are exactly zero, unlike the ℓ_2 norm ball which is rotationally invariant. In higher dimensions, the ℓ_1 penalty thus favors solutions with more zero weights in comparison to the ℓ_2 penalty, achieving the desired sparsity.

$\blacksquare$

Extra Credit (Lasso) Show that placing an equal zero-mean Laplace prior on each element of the weights $\boldsymbol{\theta}$ of a model is equivelent to ℓ_1 regularization in the Maximum-a-Posteriori estimate

$$\text{maximize: } \mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) = \frac{\mathbb{P}(\mathcal{D}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})}.$$

Note the form of the Laplace distribution is

$$\text{Lap}(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

where μ is the location parameter and $b > 0$ controls the variance. Draw (by hand) and compare the density $\text{Lap}(x|0,1)$ and the standard normal $\mathcal{N}(x|0,1)$ and suggest why this would lead to sparser solutions than a Gaussian prior on each elements of the weights (which correspond to ℓ_2 regularization).

■