Thitaree Tanprasert (Mint)
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 $\mathbf{v}_j^\top \boldsymbol{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

(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) Based on the hint, we will consider $k = 2$, which is

$$\begin{aligned}
\left\| \mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j \right\|_2^2 &= (\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^\top (\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j) \\
&= \mathbf{x}_i^\top \mathbf{x}_i - \mathbf{x}_i^\top \sum_{j=1}^{k} z_{ij}\mathbf{v}_j - (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^\top \mathbf{x}_i + (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j)^\top (\sum_{j=1}^{k} z_{ij}\mathbf{v}_j) \\
&= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^\top \mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^\top z_{ij}^\top z_{ij}\mathbf{v}_j
\end{aligned}$$

Next, we know from the hint that $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$, so we can replace z_{ij} in the equation.

$$\begin{aligned}\left\| \mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j \right\|_2^2 &= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} (\mathbf{x}_i^\top \mathbf{v}_j)\mathbf{v}_j^\top \mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^\top (\mathbf{x}_i^\top \mathbf{v}_j)^\top (\mathbf{x}_i^\top \mathbf{v}_j)\mathbf{v}_j \\ &= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j + \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j \mathbf{v}_j\end{aligned}$$

Then, we know that $\mathbf{v}_i^\top \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise. So, we can simplify the equation as follow:

$$\begin{aligned}\left\| \mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j \right\|_2^2 &= \mathbf{x}_i^\top \mathbf{x}_i - 2\sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j + \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j \\ &= \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j\end{aligned}$$

(b) Based on the given definition, we can expand and rearrange the terms as follow:

$$\begin{aligned}J_k &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} \mathbf{v}_j^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{v}_j \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}[\mathbf{v}_j^\top \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i \mathbf{x}_i^\top)\mathbf{v}_j]\end{aligned}$$

By definition, we know that $\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i \mathbf{x}_i^\top) = \mathbf{\Sigma}$. Therefore, we get

$$\begin{aligned}J_k &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}(\mathbf{v}_j^\top \mathbf{\Sigma}\mathbf{v}_j) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k}\lambda_j\end{aligned}$$

(c) Based on the hint, we will partition the summation in the result from previous part.

$$J_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - (\sum_{j=1}^{d}\lambda_j - \sum_{j=k+1}^{d})\lambda_j$$

Since $J_d = 0$, we know that

$$J_d = 0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{d} \lambda_j$$

$$\sum_{j=1}^{d} \lambda_j = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i$$

Plugging this into the equation for J_k above, we get

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

■

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

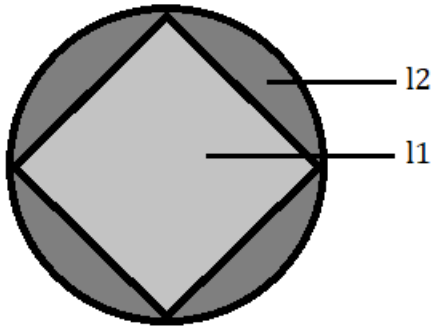
$$\begin{aligned} &\text{minimize: } f(\mathbf{x}) \\ &\quad \text{subj. to: } \|\mathbf{x}\|_p \leq k \end{aligned}$$

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.

In the figure below, the tilted square corresponds to B_k. The length of each side is $k\sqrt{2}$. The circle corresponds to A_k, and its radius = k.



We would like to minimize $f(\mathbf{x})$ subject to $\|\mathbf{x}\|_p \leq k$, which is the same as finding the $\inf_{\mathbf{x}} \sup_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda)$. Then, we can replace $\mathcal{L}(\mathbf{x}, \lambda)$ with its definition $f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p - \lambda k$. The last term does not depend on $\mathbf{x}$, so we can remove it. Also, we know that sup and inf can switch place in this case, because they are for different variables, $\mathbf{x}$ and λ. So, sup is now in front and turns this problem into a minimization problem. Therefore, if we minimize $\inf_x f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p - \lambda k$, we also minimize $f(\mathbf{x})$ subject to $\|\mathbf{x}\|_p \leq k$ as desired.

We know that the optimization we seek is the point where the error plot "meets" the norm-ball. It is clear that there are many spots on the edge of the norm-ball with ℓ_2 regularization. But for ℓ_1 regularization, the error can only meets the norm-ball close to one of the 4 peaks (in 2D case). Because the ℓ_1 norm-ball cannot be rotated, its peaks are

always at the point where at least one feature is zero. At higher dimensionality, we can turn the features that are very near zero to zero and cause the solution to become sparse.

■

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).

■