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) we are want to find the relation:

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

Considering the case of $k = 2$, we will use the fact that $\mathbf{v}_i^\top \mathbf{v}_j$ is 1 if $i = j$ and 0 otherwise and that $z_{ij} = \mathbf{x}_i^\top \mathbf{v}_j$ to define $\left\|\mathbf{x}_i - \sum_{j=1}^k z_{ij}\mathbf{v}_j\right\|^2$:

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

Therefore, we have proven the equation as desired.

(b) We start by the given definition:

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

we can rearrange the equation as follows:

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

since we know that $\mathbf{v}_j^\top \boldsymbol{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j^\top\mathbf{v}_j = \lambda_j$, we can substitute as follows:

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

Therefore, we have proven the equation as desired.

(c) We know that $J_d = 0$, therefor we can get the following relation from part b:

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

since we know that $\sum_{j=1}^{d} \lambda_j$ into $\sum_{j=1}^{k} \lambda_j$ and $\sum_{j=k+1}^{d} \lambda_j$, we can do the following substitutions:

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

Therefore we find the error as desired.

■

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

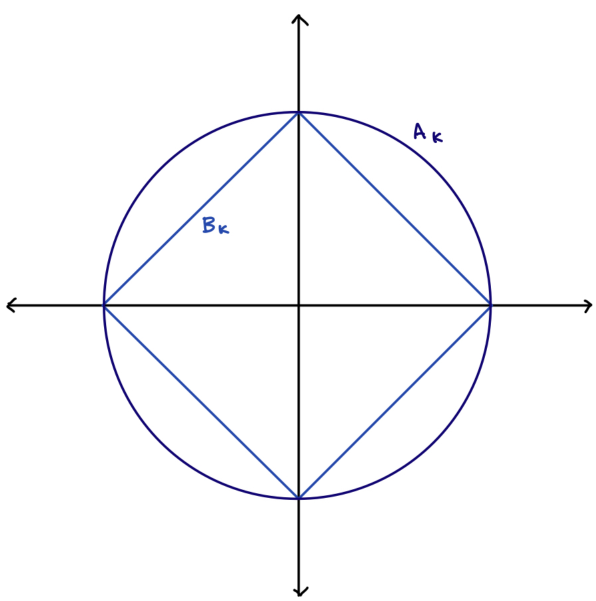
$$\begin{aligned} \text{minimize: } & f(\mathbf{x}) \\ \text{subj. to: } & \|\mathbf{x}\|_p \leq k \end{aligned}$$

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 norm-ball B_k and Euclidean norm-ball A_k:



Given the optimization problem:

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

Using Lagrangian, we know that:

$$
\begin{aligned}
P* &= \inf_{x \in X} \sup_{\lambda \geq 0} L(x, \lambda) \\
&= \inf_{x \in X} \sup_{\lambda \geq 0} f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k) \\
&= \inf_{x \in X} \sup_{\lambda \geq 0} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p - \lambda k
\end{aligned}
$$

When minimizing $\inf_{x \in X} \sup_{\lambda \geq 0} f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p - \lambda k$ we can drop λk since it does not rely on x, thus we know that to optimize for x we need to solve:

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

Therefore, we find the relationship as desired.
We can intuitively understand why l_1 regularization will give sparser solutions than l_2 by visualizing the ball's 3D representations. A solution is given by the "surface" of the ball but not by an "edge"; since l_1 is a octohedral and l_2 is a sphere, l_1 will have a greater chance of picking an edge and therefore the penalty will cause more weights to be zero.

■

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).

■