Math189R SP19
Homework 5
Monday, Mar 11, 2019

CONSULTED SOLUTIONS FOR 1 and 2

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) Because z_{ij} is a real number - it follows that $z_{ij}^\top = z_{ij}$. Let us begin by simplifying the left hand side

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

as desired.

(b) Let us simplify the left hand side

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

as desired.

(c) If $J_d = 0$ then it follows that

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

since n is independent of k it then follows that

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

since $k < d$ it follows that all the terms up to k are cancelled, we then get

$$= \boxed{\sum_{j=k+1}^{d} \lambda_j,}$$

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.

For $k = 1$, we can solve for explicit equations for both ℓ_1 and ℓ_2 in two dimensions.

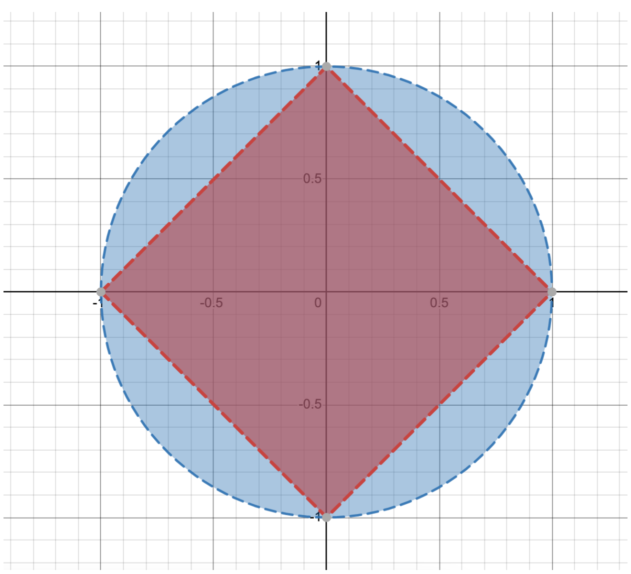
For ℓ_1 we get

$$-1 + |x| \leq y \leq 1 - |x|,$$

from $\{-1 \leq x \leq 1\}$ and for ℓ_2 we get

$$y^2 + x^2 \leq 1,$$

giving a graph of



∎

We know that the Lagrangian for minimizing $f(\mathbf{x})$ such that $||\mathbf{x}||_p \leq k$ is

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(||\mathbf{x}||_p - k),$$

since $\lambda k \in \mathbb{R}$, this value will not affect the minimization - and therefore it is equivalent to minimize

$$\mathcal{L}(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda||\mathbf{x}||_p,$$

showing that

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

is equivalent to

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

$\blacksquare$.

We can argue that the solutions for ℓ_1 will be more sparse than ℓ_2 because ℓ_1 solutions are solutions to a linear system, compared to the solutions of ℓ_2 which are for a spherical system. For small values of λ the radius is not as large - but for larger one λ - the gap we see in the graph from the first part is much more pronounced.

It then follows that when finding solutions for ℓ_1 we are more likely to land on a corner of the square which ℓ_1 covers. Since a corner requires one of the dimensions to be 0 - it is much more likely for ℓ_1 solutions to have more 0's than ℓ_2 solutions - making them sparser. ℓ_2 does not have any corners and it is equally likely for a solution to land any where on the sphere.

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).

■