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)

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

■

(b)

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

Where

$$
\mathbf{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^2
$$

From Murphy Eq. 12.37.

Substitute in λ_j

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

■

(c) Given that $J_d = 0$, we can write

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

Then, we can express J_k as such

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

■

■

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.

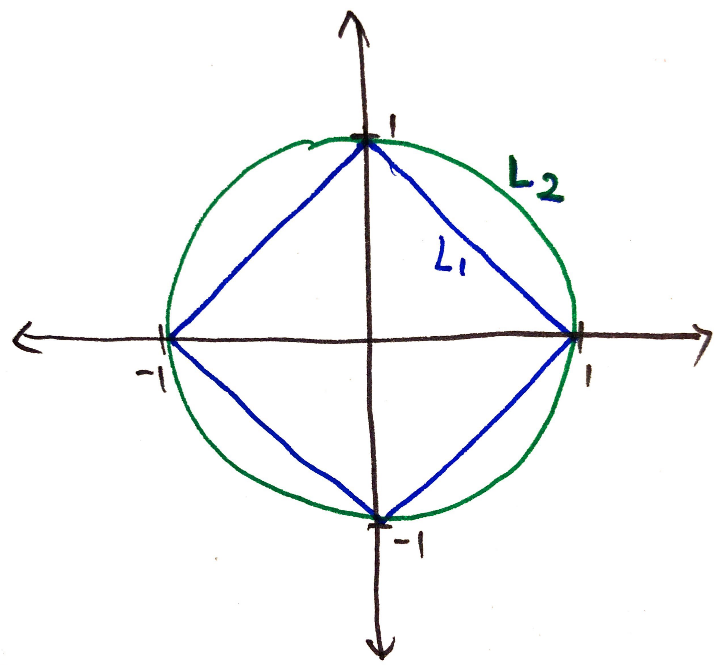

Figure 1: The norm balls for L1 and L2 metrics

For the optimization problem

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

We can construct the Lagrangian of $f(\mathbf{x})$, take the gradient to solve for the minimum.

$$\begin{aligned} L(\mathbf{x}, \lambda) =& f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k) \\ \nabla_{\mathbf{x}} L(\mathbf{x}, \lambda) =& \nabla_{\mathbf{x}} f(\mathbf{x}) + \lambda \nabla_{\mathbf{x}} \|\mathbf{x}\|_p = 0 \end{aligned}$$

For the optimization problem

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

We can simply take the gradient of this function, which we can call $g(\mathbf{x})$, to solve for the minimum

$$\nabla_{\mathbf{x}} g(\mathbf{x}) = \nabla_{\mathbf{x}} f(\mathbf{x}) + \lambda \nabla_{\mathbf{x}|} \|\mathbf{x}\|_p = 0$$

We see from the to equations above that the optimization problems are equivalent.

$\blacksquare$

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

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

where μ is the location parameter and $b > 0$ controls the variance. Draw (by hand) and compare the density $\mathrm{Lap}(x|0,1)$ 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).

■