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. We know that:

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

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T\mathbf{x}_i + \sum_{j=1}^{k} z_{ij}^2\mathbf{v}_j^T\mathbf{v}_j$$

Since all $\mathbf{v}_j$ vectors are orthonormal, $\mathbf{v}_j^T\mathbf{v}_j = 1$, and $z_{ij} = \mathbf{v}_j^T\mathbf{x}_i$, the above results in:

$$\mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}^2 + \sum_{j=1}^{k} z_{ij}^2 = \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}^2$$

$$= \mathbf{x}_i^T \mathbf{x}_i - \sum_{j=1}^{k} (\mathbf{x}_i^T \mathbf{v}_j)^T (\mathbf{x}_i^T \mathbf{v}_j) = \mathbf{x}_i^T \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}_j$$

b. We know that $\boldsymbol{\Sigma}$, the correlation matrix of $\mathbf{X}$, can be written as $\frac{1}{n}\sum_{j=1}^{k} \mathbf{x}_i \mathbf{x}_i^T$. Therefore, $\frac{1}{n}\sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{x}_i \mathbf{x}_i^T \mathbf{v}_j$ can be written as $\frac{1}{n}\sum_{j=1}^{k} \mathbf{v}_j^T \boldsymbol{\Sigma} \mathbf{v}_j$. Therefore,

$$J_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \boldsymbol{\Sigma} \mathbf{v}_j$$

. However, since $\mathbf{v}_j^T \boldsymbol{\Sigma} \mathbf{v}_j = \lambda_j \mathbf{v}_j^T \mathbf{v}_j$, and $\mathbf{v}_j^T \mathbf{v}_j = 1$, the above can be expressed as:

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

c. From above, we know that $J_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j$, and that $J_d = 0 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{d} \lambda_j$. Therefore,

$$J_k - J_d = \sum_{j=1}^{d} \lambda_j - \sum_{j=1}^{k} \lambda_j$$

Note that $J_d = 0$ and $\sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{k} \lambda_j + \sum_{j=k+1}^{d} \lambda_j$. The above therefore yields:

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

■

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

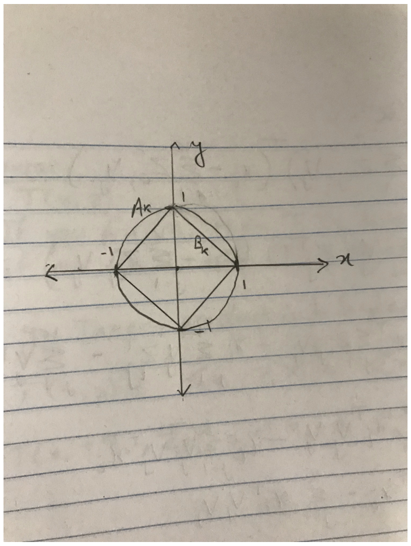
$$\begin{aligned} \text{minimize: } & f(\mathbf{x}) \\ \text{subj. to: } & \|\mathbf{x}\|_p \leq k \end{aligned}$$

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.

The graphs are drawn below:



Minimising $f(\mathbf{x})$ with constraints $||\mathbf{x}||_p \leq k$ can be solved with Lagrangian Multipliers. Using Lagrangian Multipliers leaves us with the optimisation problem $L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \lambda(||\mathbf{x}||_p - k)$. In order to create the Lagrangian, we can take the partial derivatives with respect to $\mathbf{x}$ and λ. We therefore get,

$$\frac{\partial}{\partial \mathbf{x}} L(\mathbf{x}, \lambda) = f'(\mathbf{x}) + \lambda$$

And

$$\frac{\partial}{\partial \lambda} L(\mathbf{x}, \lambda) = ||\mathbf{x}||_p - k$$

We can see that the terms of the Lagrangian governing the optimal values of $\mathbf{x}$, do not depend on k. This optimatisation problem is therefore equivalent to

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

Based on what was discussed in class, the optimal solution solution occurs at the point where a contour line of $f(\mathbf{x})$ meets the constraint. With the ℓ_1 norm, we have 'sharper' constraint, which makes it more likely for a contour line of $f(\mathbf{x})$ to intersect at one of the corners of the norm-ball. With ℓ_2 regularisation, a contour of $f(\mathbf{x})$ is less likely to intersect with the norm-ball along the axes. In fact, there are no 'corners' for the contour to intersect with. Since intersection on the axes, i.e., the corners of the ℓ_1 norm ball correspond to sparser solutions (on the axes, at least one dimension is nil), the ℓ_1 norm usually yields sparser solutions. Of course, the λ value needs to be suitably large in order to push the parametres values close to 0.

$\blacksquare$

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).

■