Haocheng Ye
Math189R SU18
Homework 5
Monday, May 21, 2018

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

1.

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

2.

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

3. Since $J_d = 0$, we have

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

And thus

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

■

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.

We can first write the optimization problem in Lagrange multiplier form

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

which is

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

Since λk does not depend on $\mathbf{x}$, then the optimization problem becomes

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

The reason that $l1$ regularization gives sparser solution than using $l2$ regularization is that the probability that the solution lands on one of the vertices is much higher in $l1$ regularization that $l2$. Since the solution is essentially is the tangent intersection between the level curve of the solution and the level curve of regularization term, $l1$ regularization has sharper edges, which is more likely to intersect with the solution level curve when one of the solution is value zero. On the other hand, $l2$ is more likely to intersect on any point on circle, which makes the solution not as sparse as $l1$. Even though that we can only see the different effect of these two regularization methods in 2-dimension, this effect can be generalized in higher dimensions, in which $l1$ regularization will always have a higher probability to make more zero coefficients than $l2$ does. ∎

Extra Credit (Lasso) Show that placing an equal zero-mean Laplace prior on each element of the weights $\boldsymbol{\theta}$ of a model is equivalent 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).

■