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

1. Observe that

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

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

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T\mathbf{x}_i + \sum_{j=1}^{k}\sum_{l=1}^{k} z_{ij}^T z_{il}\mathbf{v}_j^T\mathbf{v}_l \tag{3}$$

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T\mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^T z_{ij}^T z_{ij}\mathbf{v}_j \tag{4}$$

$$= \mathbf{x}_i^T\mathbf{x}_i - 2\sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T\mathbf{x}_i + \sum_{j=1}^{k} \mathbf{v}_j^T \mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_j \tag{5}$$

$$= \mathbf{x}_i^T\mathbf{x}_i - \sum_{j=1}^{k} z_{ij}\mathbf{v}_j^T\mathbf{x}_i \tag{6}$$

using the orthonormality of $\mathbf{v}_j$ vectors.

2. Recall that

$$\mathbf{\Sigma} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^T. \tag{7}$$

Hence, (using the fact that $\mathbf{v}_j^\top\mathbf{\Sigma}\mathbf{v}_j = \lambda_j\mathbf{v}_j^\top\mathbf{v}_j = \lambda_j$),

$$J_k = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top\mathbf{x}_i\mathbf{x}_i^\top\mathbf{v}_j \right) \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top \left(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top \right) \mathbf{v}_j \tag{9}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k} \mathbf{v}_j^\top\mathbf{\Sigma}\mathbf{v}_j \tag{10}$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top\mathbf{x}_i - \sum_{j=1}^{k} \lambda_j \tag{11}$$

as desired.

3. From the results of part (b), see that

$$J_d = 0 \implies \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^T\mathbf{x}_i = \sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{k} \lambda_j + \sum_{k+1}^{d} \tag{12}$$

since $k < d$. Therefore

$$J_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j \tag{13}$$

implies that

$$\begin{aligned} J_k &= \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i - \sum_{j=1}^{k} \lambda_j & (14) \\ &= \sum_{j=1}^{k} \lambda_j + \sum_{k+1}^{d} \lambda_j - \sum_{j=1}^{k} \lambda_j & (15) \\ &= \sum_{j=k+1}^{d} \lambda_j, & (16) \end{aligned}$$

concluding the proof.

∎

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 first know that the optimization problem above is equivalent to the Lagrangian formulation:

$$\inf_{\mathbf{x}} \sup_{\lambda \geq 0} \mathcal{L}(\mathbf{x}, \lambda) = \inf_{\mathbf{x}} \sup_{\lambda \geq 0} [f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k)]. \tag{17}$$

(took hint from solution here) We can then interchange the inf and sup in the dual space to get

$$\sup_{\lambda \geq 0} g(\lambda) = \sup_{\lambda \geq 0} \inf_{\mathbf{x}} [f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k)] \tag{18}$$

where

$$g(\lambda) = \inf_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda). \tag{19}$$

Finally, we notice that

$$\sup_{\lambda \geq 0} \inf_{\mathbf{x}} [f(\mathbf{x}) + \lambda(\|\mathbf{x}\|_p - k)] = \sup_{\lambda \geq 0} \inf_{\mathbf{x}} [f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p] \tag{20}$$

since λk is independent of $\mathbf{x}$. We have reduced the optimization problem above to the problem

$$\text{minimize: } f(\mathbf{x}) + \lambda\|\mathbf{x}\|_p, \tag{21}$$

concluding the proof.

For sparsity, we wish to choose a regularization that is more likely to create 0 weights. From our pictures, we notice that when penalizing a model with the l_2 norm, it will be less likely that its solutions will be sparse compared to the likelihood of penalizing the model with the l_1 norm (this is due to the probability of our solution being able to touch a corner of the l_1 ball). The curved edges of the l_2 ball decentivizes sparse solutions as solutions will intersect the face of the ball (where it will be nonzero).

See drawings on image. ∎

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).

Again, we rewrite the problem. The maximum-a-posteriori estimate problem above is equivalent to its log formulation:

$$\begin{aligned}
\text{maximize: } \log\mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) &= \log\frac{\mathbb{P}(\mathcal{D}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})} && (22)\\
&= \log\mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) + \log\mathbb{P}(\boldsymbol{\theta}) - \log\mathbb{P}(\mathcal{D}) && (23)\\
&= \log\mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) + \log\mathbb{P}(\boldsymbol{\theta}) && (24)
\end{aligned}$$

where, like in the previous problem, we drop the term independent of $\boldsymbol{\theta}$. It suffices now to show that

$$-\log\mathbb{P}(\boldsymbol{\theta}) = \lambda\|\boldsymbol{\theta}\|_1 + C \tag{25}$$

where C is some constant. Given that $\boldsymbol{\theta}_i$ is of $\text{Lap}(\boldsymbol{\theta}_i \mid 0, b)$ distribution, see that

$$\begin{aligned}
-\log\mathbb{P}(\boldsymbol{\theta}) &= -\log\prod_{i=1}^{N}\exp\big(-\frac{\|\boldsymbol{\theta}_i\|}{b}\big) + \log(2b) && (26)\\
&= \frac{1}{b}\|\theta_i\|_1 + \log(2b) && (27)
\end{aligned}$$

which shows that the problem above is equivalent to

$$\text{minimize: } -\log\mathbb{P}(\boldsymbol{\theta}|\mathcal{D}) = -\log\mathbb{P}(\mathcal{D}|\boldsymbol{\theta}) + \frac{1}{b}\|\theta_i\|_1, \tag{28}$$

as desired. See drawings on image. The $\text{Lap}(x|0,1)$ encourages sparsity as it is more ample near $x = 0$ than the Gaussian prior. ∎