Math189R SP20
Homework 5
Monday, March 2, 2020

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. Magnitude of vector can be rewritten as following:

$$\|x_i - \sum_{j=1}^{k} z_{ij} v_j\|^2 = (x_i - \sum_{j=1}^{k} z_{ij} v_j)^T (x_i - \sum_{j=1}^{k} z_{ij} v_j)$$

$$= \left[x_i^T - \left(\sum_{j=1}^{k} z_{ij} v_j\right)^T\right](x_i - \sum_{j=1}^{k} z_{ij} v_j) \qquad \text{distribute the terms}$$

$$= x_i^T x_i - \underbrace{x_i^T \sum_{j=1}^{k} z_{ij} v_j - \left(\sum_{j=1}^{k} z_{ij} v_j\right)^T x_i}_{\text{identical} \Rightarrow \text{combine}} + \left(\sum_{j=1}^{k} z_{ij} v_j\right)^T \left(\sum_{j=1}^{k} z_{ij} v_j\right)$$

$$= x_i^T x_i - 2 \sum_{j=1}^{k} z_{ij} v_j^T x_i + \left(\sum_{j=1}^{k} z_{ij} v_j\right)^T \left(\sum_{j=1}^{k} z_{ij} v_j\right)$$

$$= x_i^T x_i - 2 \sum_{j=1}^{k} v_j^T x_i z_{ij} + \sum_{j=1}^{k} z_{ij}^T v_j^T v_j z_{ij} \qquad \text{Substitute in } z_{ij} = x_i^T v_j$$

$$= x_i^T x_i - \underbrace{2 \left(\sum_{j=1}^{k} v_j^T x_i x_i^T v_j\right) + \sum_{j=1}^{k} (v_j^T x_i) \cancel{v_j^T v_j} (x_i^T v_j)} \qquad \text{Since } v_j^T v_j = 0$$

$$= x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j$$

b. $\frac{1}{n}\sum_{i=1}^{n}\left(x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j\right)$

$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} v_j^T x_i x_i^T v_j$ distribute the summation

$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i - \frac{1}{n}\sum_{j=1}^{k} v_j^T\left(\sum_{i=1}^{n} x_i x_i^T\right) v_j$ since the covariance matrix $\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{k} v_j^T \Sigma v_j$ Substitute in $v_j^T \Sigma v_j = \lambda_j$

$= \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i - \sum_{j=1}^{k} \lambda_j$

c. As shown in b, $J_d = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T - \sum_{j=1}^{d}\lambda_j$, when $J_d = 0$,

$\frac{1}{n}\sum_{n=1}^{n} x_i x_i^T - \sum_{j=1}^{d}\lambda_j$.

If we write J_k by partitioning the sum,

$$J_k = \cancel{\frac{1}{n}\sum_{n=1}^{n} x_i x_i^T - \sum_{j=1}^{k}\lambda_j} + \sum_{j=k+1}^{d}\lambda_j$$

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

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).

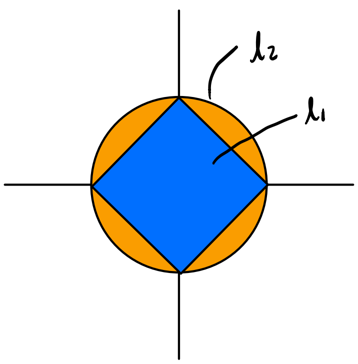
Show that the optimization problem

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

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.



min $f(x)$: $\|x\|_p \leq k$ can be re-written into the following:

$\inf_x \sup_{\lambda \geq 0} \mathcal{L}(x, \lambda)$

$= \inf_x \left(\sup_{\lambda \geq 0} f(x) + \lambda(\|x\|_p - k)\right)$

Since k is a constant, when we minimize over x, it does not affect the result. Thus, we can re-write as $\sup_{\lambda \geq 0} (\inf f(x) + \lambda\|x\|_p)$, which is essentially minimize: $f(x) + \lambda\|x\|_p$ for some $\lambda \geq 0$

l_1 regularization is more sparse because the chance of touching its edge instead of face is infinitely larger than the l_2 ball. Thus, l_1 results in more weight becoming 0, compared to l_2.

■

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).

■