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 can prove this algebraically

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

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

$$= x_i^T x_i - 2\sum_{j=1}^{k} z_{ij} v_j^T x_i + \sum_{j=1}^{k}\sum_{i=1}^{k} z_{ij} v_j^T z_{ij} v_j \quad \text{(bringing } x_i^T \text{ into sum)}$$

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

$$= x_i^T x_i - 2\sum_{j=1}^{k} z_{ij} v_j^T x_i + \sum_{j=1}^{k} v_j^T x_i x_i^T v_j \quad \text{(since } v_i^T v_j = 1 \text{ iff } i = j\text{)}$$

$$= x_i^T x_i - 2\sum_{j=1}^{k} z_{ij} v_j^T x_i + \sum_{j=1}^{k} v_j^T x_i x_i^T v_j \quad \text{(since } z_{ij} \in \mathbb{R}\text{)}$$

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

(b) By definition

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

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

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

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

as desired.

(c) Since $J_d = 0$ we know $\sum_{j=1}^{d} \lambda_j = \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i$. Then

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

■

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

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

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.

Though not included in my solution we can think of the balls, respectively a square and circle on the xy plane for respective L1 and L2 norm. Using these We know the optimization problem

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

is equivalent to

$$\inf_x \sup_{\lambda \geq 0} L(x, \lambda) = \inf_x \sup_{\lambda \geq 0} f(x) + \lambda(\|x\|_p - k).$$

In its dual we can flip the inf and sup.

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

Since the minimizing value of $f(x) + \lambda(\|x\|_p - k)$ over x is equivalent to the minimizing value of $f(x) + \lambda\|x\|_p$ (-λk doesn't depend on x), we know the optimizing x will solve

minimize: $f(x) + \lambda \|x\|_p$

for some suitable value of $\lambda \geq 0$. Looking at the plot and this result, we can consider ℓ_1 regularization as projecting the actual optimal solution of your problem onto some suitably sized ℓ_1 norm ball. Since the ℓ_1 ball has sharper edges, the probability of landing on an edge and not on the face (where both elements of the vector are nonzero) is infinitely larger than the ℓ_2 ball. This is due to the rotation invariance of the ℓ_2 that certainly doesn't hold for the ℓ_1 ball. Generalizing to higher dimensions, we can see that the ℓ_1 penalty will encourage more weights to be zero compared to the ℓ_2 ball, as desired.

■

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).

■