Math189R SU17
Homework 5
Monday, June 5, 2017

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 have $\left\|\vec{x}_i - \sum_{j=1}^{k} z_{ij}\vec{v}_j\right\|^2$

$= \left(\vec{x}_i - \sum z_{ij}\vec{v}_j\right)^T\left(\vec{x}_i - \sum z_{ij}\vec{v}_j\right)$

$= \left(\vec{x}_i^T - \sum z_{ij}\vec{v}_j^T\right)\left(\vec{x}_i - \sum z_{ij}\vec{v}_j\right)$ Multiply out to get:

$= \vec{x}_i^T\vec{x}_i - \vec{x}_i^T\sum z_{ij}\vec{v}_j - \vec{x}_i\sum z_{ij}\vec{v}_j^T + \sum z_{ij}\vec{v}_j^T\sum z_{ij}\vec{v}_j$

$= \vec{x}_i^T x_i - \underbrace{\sum z_{ij}^2 - \sum z_{ij}^2}_{\text{Because } z_{ij} = \vec{x}_i^T\vec{v}_j} + \sum z_{ij}\vec{v}_j^T \sum z_{ij}\vec{v}_j$

$= \vec{x}_i^T\vec{x}_i - 2\sum z_{ij}^2 + \sum z_{ij}^2$ Because $v_i^T v_j$ is 1 if $i=j$ & 0 otherwise

$= \vec{x}_i^T x_i - \sum z_{ij}^2 = \boxed{\vec{x}_i^T x_i - \sum_{j=1}^{k}\vec{v}_j^T x_i \vec{x}_i^T \vec{v}_j}$ as desired

b) We have $J_k = \frac{1}{n}\sum_{i=1}^{n}\left[\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\left(\vec{v}_j^T \vec{x}_i x_i^T \vec{v}_j\right)\right]$

$$= \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\left[\vec{v}_j^T\left(\frac{1}{n}\sum_{i=1}^{n}\Sigma_i\right)\vec{v}_j\right]$$

avg of covariance matrices

$$= \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\vec{v}_j^T\Sigma\vec{v}_j$$

Because $\vec{v}_j^T\Sigma\vec{v}_j = \lambda_i\vec{v}_j^T v_j = \lambda_i$

$$\Rightarrow \boxed{\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\lambda_i}$$ as desired.

c) From part b, we know that

$$J_d = \frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{d}\lambda_i = 0$$

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}\vec{x}_i^T\vec{x}_i - \sum_{j=1}^{k}\lambda_i}_{=J_k} - \sum_{j=k+1}^{d}\lambda_i = 0$$

$$\boxed{J_k = \sum_{j=k+1}^{d}\lambda_i}$$ as desired

**2 ($\ell_1$-Regularization)** Consider the $\ell_1$ norm of a vector $\mathbf{x} \in \mathbb{R}^n$:

$$\|\mathbf{x}\|_1 = \sum_i |\mathbf{x}_i|.$$

Draw the norm-ball $B_k = \{\mathbf{x} : \|\mathbf{x}\|_1 \le k\}$ for $k = 1$. On the same graph, draw the Euclidean norm-ball $A_k = \{\mathbf{x} : \|\mathbf{x}\|_2 \le k\}$ for $k = 1$ behind the first plot. (Do not need to write any code, draw the graph by hand).

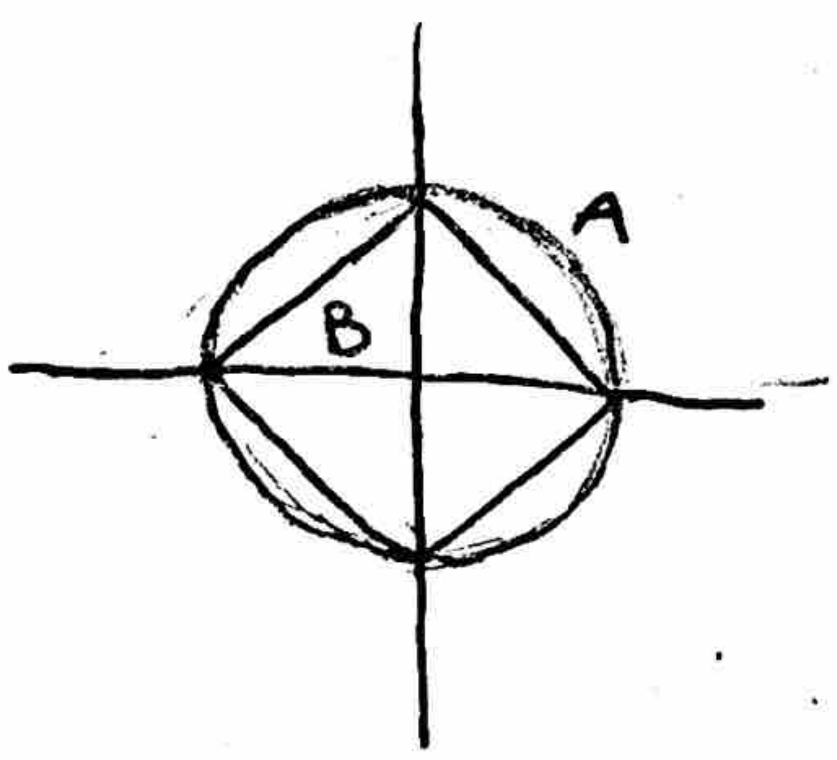
Show that the optimization problem

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

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 $\ell_1$ regularization (adding a $\lambda\|\mathbf{x}\|_1$ term to the objective) will give sparser solutions than using $\ell_2$ regularization for suitably large $\lambda$.




We have the Lagrangian:

$$L(x, \lambda) = f(x) + \lambda(\|x\|_p - k)$$

$$= f(x) + \lambda\|x\|_p - \cancel{\lambda k}$$ does not depend on $x$ so doesn't matter

So minimizing $f(\vec{x})$ subject to $\|\vec{x}\|_p \le k$ is equivalent to minimizing $f(\vec{x}) + \lambda\|\vec{x}\|_p$

Checked solution to confirm intuition:

Because $\ell_1$ regularization has edges and corners, when the solution is projected onto the surface, it is much more likely (than a sphere) (infinitely more likely in fact) to land on a corner where one of the parameters will be sent to zero, yielding sparser solutions.

■