Math 189R problem set 4

Adam Guo 2020-02-17

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

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

   Hint: first consider the case when $k = 2$. Use the fact that $\boldsymbol{v}_i^T \boldsymbol{v}_j$ is 1 if $i = j$ and 0 otherwise. Recall that $z_{ij} = \boldsymbol{x}_i^T \boldsymbol{v}_j$.

   Solution:

$$\begin{aligned}
\left\| \boldsymbol{x}_i - \sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j \right\|^2 &= (\boldsymbol{x}_i - \sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j)^T (\boldsymbol{x}_i - \sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j) \\
&= \boldsymbol{x}_i^T \boldsymbol{x}_i - \boldsymbol{x}_i \sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j - \left(\sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j^T \right) \boldsymbol{x}_i + \sum_{j,l=1}^{k} \boldsymbol{v}_j^T z_{ij}^T z_{il} \boldsymbol{v}_l \\
&= \boldsymbol{x}_i^T \boldsymbol{x}_i - 2 \sum_{j=1}^{k} z_{ij} \boldsymbol{v}_j^T \boldsymbol{x}_i + \sum_{j=1}^{k} \boldsymbol{v}_j^T z_{ij}^T z_{ij} \boldsymbol{v}_j \\
&= \boldsymbol{x}_i^T \boldsymbol{x}_i - 2 \sum_{j=1}^{k} \boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}_j + \sum_{j=1}^{k} \boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}_j \\
&= \boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{j=1}^{k} \boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}_j
\end{aligned}$$

   (b) Now show that

$$J_k = \frac{1}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{j=1}^{k} \boldsymbol{v}_j^T \boldsymbol{x}_i \boldsymbol{x}_i^T \boldsymbol{v}_j \right) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{x}_i - \sum_{j=1}^{k} \lambda_j.$$

   Hint: recall that $\boldsymbol{v}_j^T \boldsymbol{\Sigma} \boldsymbol{v}_j = \lambda_j \boldsymbol{v}_j^T \boldsymbol{v}_j = \lambda_j$.

   Solution:

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

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

Solution:

$$
\begin{aligned}
J_k &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{j=1}^{k}\lambda_j \\
&= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{j=1}^{d}\lambda_j + \sum_{j=k+1}^{d}\lambda_j
\end{aligned}
$$

Since $J_d = 0$,

$$
0 = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{i=1}^{d}\lambda_j
$$

Thus,

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

2. **(ℓ_1-Regularization)** Consider the ℓ_1 norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$:

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

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

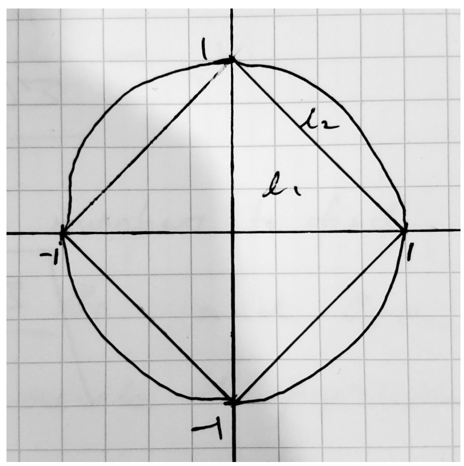
$$\begin{aligned} \text{minimize: } & f(\boldsymbol{x}) \\ \text{subj. to: } & \|\boldsymbol{x}\|_p \leq k \end{aligned}$$

is equivalent to

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

(hint: create the Lagrangian). With this knowledge, and the plots given above, argue why using ℓ_1 regularization (adding a $\lambda\|\boldsymbol{x}\|_1$ term to the objective) will give sparser solutions than using ℓ_2 regularization for suitably large λ.

Solution:



The Lagrangian is given by

$$\mathcal{L}(\boldsymbol{x}, k) = f(x) + \lambda(\|\boldsymbol{x}\|_p - k) = f(x) + \lambda\|\boldsymbol{x}\|_p - \lambda k$$

We want to minimise the Lagrangian to find the optimal solution for $\boldsymbol{x}$. Since λk does not depend on $\boldsymbol{x}$, this optimisation problem is equivalent to minimising $f(x) + \lambda\|\boldsymbol{x}\|_p$. We see that the Euclidean distance from the origin to the "corners" of the l_1 ball than the distance to the edges. Hence, the optimal solution to the problem is more likely to first intersect with a corner of the l_1 ball than an edge. Compare this to the l_2 ball which is equidistant in Euclidean space to the origin, and therefore equally likely to first intersect the optimal solution at any point. Hence, using the l_1 ball is more likely to yield optimal solutions that lie on the x or y axes, and are therefore more sparse.

3. **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

$$\text{Lap}(x|\mu, b) = \frac{1}{2b}\exp\left(-\frac{|x-\mu|}{b}\right)$$

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).

Solution:

Maximising $\mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D})$ is equivalent to maximising the log likelihood,

$$\ln \mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D}) = \ln \frac{\mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\mathcal{D})} = \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \ln \mathbb{P}(\boldsymbol{\theta}) - \mathbb{P}(\mathcal{D})$$

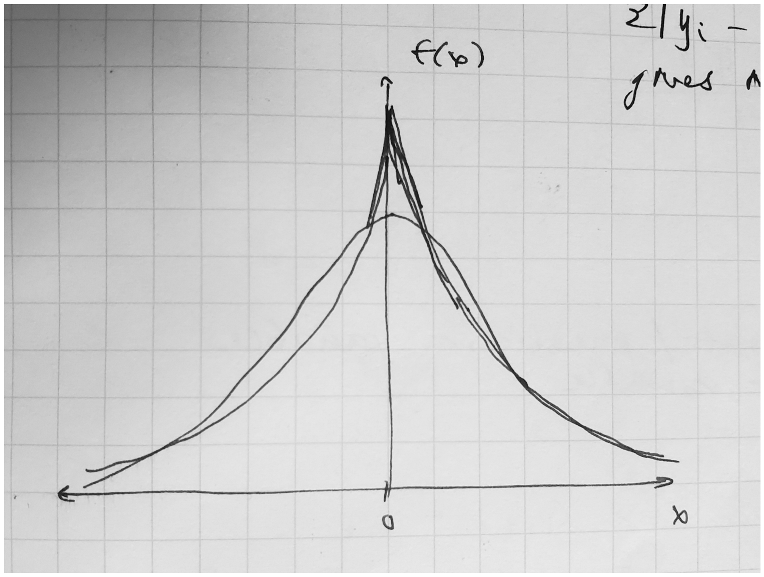
Since $\mathbb{P}(\mathcal{D})$ does not depend on $\boldsymbol{\theta}$, this is equivalent to

$$\begin{aligned}
\text{maximise:} \quad & \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \ln \mathbb{P}(\boldsymbol{\theta}) \\
&= \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \ln\left(\exp\left(-\frac{|\boldsymbol{\theta}_i|}{b}\right)\right) \\
&= \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \ln\left(\prod_i \exp\left(-\frac{|\boldsymbol{\theta}_i|}{b}\right)\right) \\
&= \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \sum_i \left(-\frac{|\boldsymbol{\theta}_i|}{b}\right) \\
&= \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) - \frac{1}{b}\sum_i |\boldsymbol{\theta}_i| \\
&= \ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) - \lambda\|\boldsymbol{\theta}\|_1, \quad \lambda = \frac{1}{b}
\end{aligned}$$

This is subsequently equivalent to

$$\text{minimise: } -\ln \mathbb{P}(\mathcal{D} \mid \boldsymbol{\theta}) + \lambda\|\boldsymbol{\theta}\|_1$$

which is of the same form as l_1 regularisation.



We see that the Laplace distribution has a sharp peak at $x = 0$ that gives it a higher probability of being exactly 0 than the normal distribution. Hence, the weights are more likely to be exactly 0, and therefore more likely to be sparse.