Bennett Mountain
Math189R SP19
Homework 5
Monday, Mar 4, 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 $\mathbf{v}_j^\top \Sigma \mathbf{v}_j = \lambda_j \mathbf{v}_j^\top \mathbf{v}_j = \lambda_j$.

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

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

$= x_i^T x_i - 2\sum_{j=1}^{k} z_{ij} v_j^T x_i + \sum_{j=1}^{k} z_{ij}^T z_{ij} v_j^T v$ by bringing in x_i^T into the summation, v_j becomes v_j^T

$= x_i^T x_i - 2\sum_{j=1}^{k} v_j^T x_i x_i^T v_j + \sum_{j=1}^{k} v_j^T x_i x_i^T v_j$ since $v_i^T v_j = 1$ if $i = j$ and 0 otherwise, plug in $z_{ij} = x_i^T v_j$

$\boxed{= x_i^T x_i - \sum_{j=1}^{k} v_j^T x_i x_i^T v_j}$ as we wished

■

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

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

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

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

as we wished

C) We can utilize the fact that $J_d = 0$ by $\sum_{j=1}^{d} \lambda_j = \frac{1}{n}\sum_{i=1}^{n} x_i^T x_i$

Thus,

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

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

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).

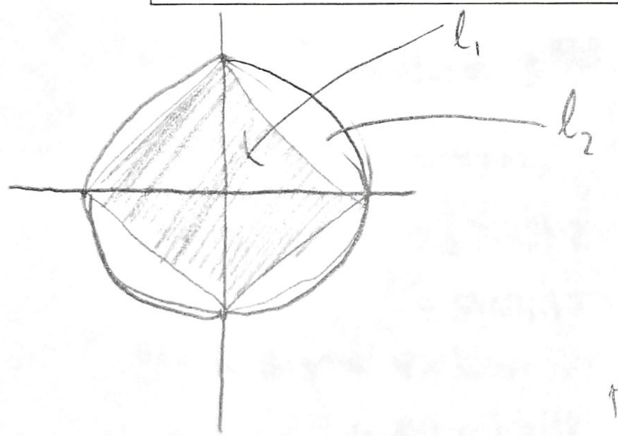
Show that the optimization problem

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

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.



minimize $f(x)$ subj. to $\|x\|_p \leq k$ is the same as

$\inf_x \sup_{\lambda \geq 0} L(x,\lambda) = \inf_x \sup_{\lambda \geq 0} f(x) + \lambda(\|x\|_p - k)$ flipping inf and sup, we find

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

we know that minimizing $f(x) + \lambda(\|x\|_p - k)$ for x is the same as min $f(x) + \lambda(\|x\|_p)$, as $-\lambda k$ doesn't depend on x, optimizing for the value of x is the same as minimize: $f(x) + \lambda\|x\|_p$ for $\lambda \geq 0$.

Considering the plot and our result, the l_1 regularization is essentially projecting the optimal solution onto an l_1 norm-ball. The l_1 norm ball compared to the l_2 norm ball has straight rather than curved edges, thus making the likelihood of landing on an edge and not the face much much larger than the l_2 ball's curved edges. This is because when we rotate the l_2 ball, nothing happens. This isn't the case for the l_1 ball. Hence, the l_1 penalty wants more weights to be 0, rather than the l_2 ball (which helps with simplification of our model, interpretation, and overfitting). ∎

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).

This problem is equivalent to maximizing $\log \mathbb{P}(\theta|D)$, if we have the monotonicity of $\log(x)$. Hence, we have

$$\text{maximize: } \log \mathbb{P}(\theta|D) = \log \mathbb{P}(D|\theta) + \log \mathbb{P}(\theta) - \cancel{\log \mathbb{P}(D)}.$$

$\mathbb{P}(D)$ doesn't depend on θ so we can drop it. Now we have

minimize $-\log \mathbb{P}(D|\theta) - \log \mathbb{P}(\theta)$. With the prior $\theta_i \sim \text{Lap}(0,b)$, we have

$$-\log \mathbb{P}(\theta) = -\log \prod_i e^{\left(-\frac{|\theta_i|}{b}\right)} + z \qquad z \text{ is a constant}$$

$$= \frac{1}{b}\sum_i |\theta_i| + z$$

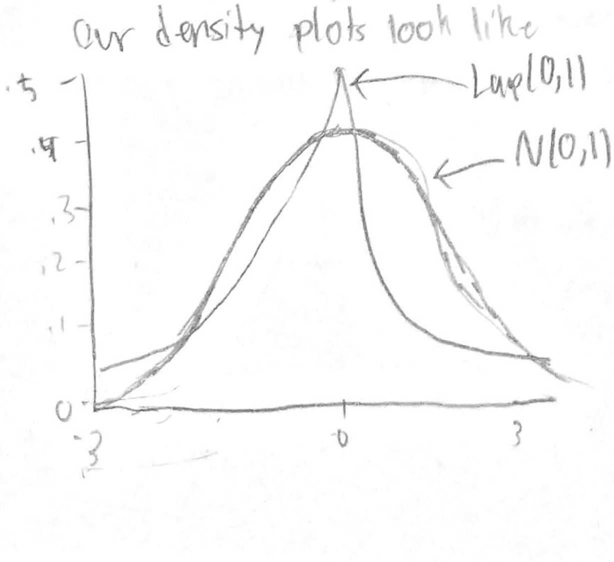
$$= \lambda \|\theta\|_1 + z \qquad \text{where } \lambda = 1/b$$

So, our original problem is the same as

$$\text{minimize: } -\log \mathbb{P}(D|\theta) + \lambda\|\theta\|_1$$

which is an ℓ_1 regularized maximum likelihood estimate, as we wished.

Our density plots look like



From this plot, we see Lap(0,1) has much more mass at $x=0$ than the Gaussian. Hence, when we use the Laplace prior instead of a Gaussian prior on the weights, the weights will be pushed to zero more, causing sparser solutions. ■