Introduction to Machine Learning
New York University
Summer 2020

Homework 1

Venkata Naga Sai Kiran Challa (Net ID: vc2118)

1. Let $\{x_1, x_2, ..., x_n\}$ be a set of points in d-dimensional space. Suppose we wish to produce a single point estimate $\mu \in R^d$ that minimizes the *mean squared-error*:

$$\frac{1}{n}(\|x_1 - \mu\|_2^2 + \|x_2 - \mu\|_2^2 + ... + \|x_n - \mu\|_2^2)$$

Find a closed form expression for μ and prove that your answer is correct.

Answer:

First, lets try out with all the possible values that we can come up with on an example and then try to prove it.

Let's consider an example vector like, $[2, 2, 3, 4, 5]$. So, $x_1 = 2,\ x_2 = 2,\ x_3 = 3$ and so on.

Right out of the bat we can think that 2 which occurred the most can minimise it. So, let see.

$$meansquared - error = \frac{1}{5}((2-2)^2+(2-2)^2+(3-2)^2+(4-2)^2+(5-2)^2) = 2.8$$

Now, it seems pretty close to minimum value 2 , but we will try rest of the values. Let's try 5, the maximum value.

$$meansquared - error = \frac{1}{5}((2-5)^2+(2-5)^2+(3-5)^2+(4-5)^2+(5-5)^2) = 4.6$$

It's a large value compare to the previous. As we are at it let's check the rest of the values as well. We checked the minimum value (which is also the *mode* of the vector) and maximum values. Let's check with the mean and median of the set as well.

We can easily find out the mean and median of the set as,

$Mean = \frac{1}{5}(2+2+3+4+5) = 3.2$

$Median = Middle\ value = 3$
Therefore,

$$meansquared - error = \frac{1}{5}((2-3.2)^2+(2-3.2)^2+(3-3.2)^2+(4-3.2)^2+(5-3.2)^2) = 1.36$$

$$meansquared - error = \frac{1}{5}((2-3)^2+(2-3)^2+(3-3)^2+(4-3)^2+(5-3)^2) = 1.4$$

Out of all the possible values we see that *mean* value minimises the most. Is it just for this case of is a general property? Let's find out.
We know that when we take differentiation *once* and equate to zero, we get the point that maximises/minimises the expression and if the *second* differentiation is grater than zero, that point gives us the minimum and vise versa.

Let $\overline{\mathrm{X}}$ be the vector of values $\{x_1, x_2, ..., x_n\}$.

$\Rightarrow$ L2 Norm of $\overline{\mathrm{X}} = \|X\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = x.x^T$

Thus, we can convert the *mean squared-error* (**MSE**) as,

$$\begin{aligned}
\mathbf{MSE} &= \frac{1}{n}[(\overline{\mathrm{X}} - \mu).(\overline{\mathrm{X}} - \mu)^T] \\
&= \frac{1}{n}\sum_{i=1}^{n}[(x_i - \mu).(x_i - \mu)^T] \\
&= \frac{1}{n}\sum_{i=1}^{n}[x_i.x_i^T - x_i\mu^T - \mu x_i^T + \mu^2]
\end{aligned} \tag{1}$$

We knows from the properties of vector spaces that,$x_i\mu^T = \mu x_i^T$

$$\begin{aligned}
\mathbf{MSE} &= \frac{1}{n}\sum_{i=1}^{n}[x_i.^2 - 2\mu x_i^T + \mu^2] \\
&= \frac{1}{n}[\sum_{i=1}^{n} x_i.^2 - 2\mu\sum_{i=1}^{n} x_i + \sum_{i=1}^{n}\mu^2] \\
&= \frac{1}{n}[\sum_{i=1}^{n} x_i.^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2]
\end{aligned} \tag{2}$$

This simplified equation can be *differentiated w.r.t* μ,

$$\Rightarrow \frac{1}{n}[-2\sum_{i=1}^{n} x_i + 2.n\mu] = 0 \tag{3}$$

$$\Rightarrow \boxed{\mu = \frac{\sum_{i=1}^{n} x_i}{n}} \tag{4}$$

We can note that the *second* differentiation of the above equation (3) is,

$$\begin{aligned} \mathbf{MSE}^{'} &= -2\sum_{i=1}^{n} x_i + 2.n\mu \\ \mathbf{MSE}^{''} &= 2.n > 0 \end{aligned} \tag{5}$$

Hence, we can justify that above equation (4) will minimises the *mean squared-error*.

2. Not all norms behave the same; for instance, the ℓ_1-norm of a vector can be dramatically different from the ℓ_2-norm, especially in high dimensions. Prove the following norm inequalities for d-dimensional vectors, starting from the definitions provided in class and lecture notes. (Use any algebraic technique/result you like, as long as you cite it.)

   a. $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty$
   b. $\|x\|_\infty \leq \|x\|_1 \leq d\|x\|_\infty$

Answer:

I will be taking generalised approach to this problem. First I am going to prove the first half in both a and b, and latter on proceed with proving the second half. We know,

$$\begin{aligned} \ell_1 - norm &= \|X\|_1 = \sum_{i=1}^{d} |x_i| \\ \ell_2 - norm &= \|X\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2} \\ \ell_p - norm &= \|X\|_p = (\sum_{i=1}^{d} x_i^p)^{\frac{1}{p}} \\ \ell_\infty - norm &= \|X\|_\infty = \lim_{x\to\infty} (\sum_{i=1}^{d} x_i^p)^{\frac{1}{p}} \\ &= max_i|x_i|, \; \forall \, i = 1, 2, ..d \end{aligned} \tag{6}$$

Let's look into *infinity norm*, $\ell_\infty - norm$. Why is it, $max_i|x_i|$? Well if you think about it, as $p \to \infty$, we can realise that most dominant term in x_i over takes the rest of the values. For more formal definition and proof, you can follow the Wikipedia's page, by clicking **here**.

For comparing different norm values, let's see them on a unit circle, taking $x_1 = [-1, 1]$ and $x_2 = [-1, 1]$ and restricting our space between $[0, 1], [1, 0], [0, -1] and [-1, 0]$. Thus, we will be finding x_1 and x_2 such that,

$$\begin{aligned} \ell_1 - norm &= \|X\|_1 = 1 \\ \ell_2 - norm &= \|X\|_2 = 1 \\ \ell_\infty - norm &= \|X\|_\infty = 1 \end{aligned} \tag{7}$$

$\ell_1 - norm$				
x_1	1	0.8	0.6	0.5
x_2	0	0.2	0.4	0.5
$\ell_2 - norm$				
x_1	1	0.8	0.4	0.5
x_2	0	0.6	0.91	0.87
$\ell_\infty - norm$				
x_1	1	0		
x_2	0	1		

Table 1: Change of $x_2\, w.r.t\, x_1$.

Thus,

$$\begin{aligned} \ell_1 = \|X\|_1 = \|(x_1, x_2)\| &\iff |x_1| + |x_2| = 1 \\ &\implies In\, Q1, x_2 = 1 - x_1 \\ &\implies In\, Q2, x_2 = 1 + x_1 \\ &\implies In\, Q3, x_2 = -x_1 - 1 \\ &\implies In\, Q4, x_2 = x_1 - 1 \end{aligned} \tag{8}$$

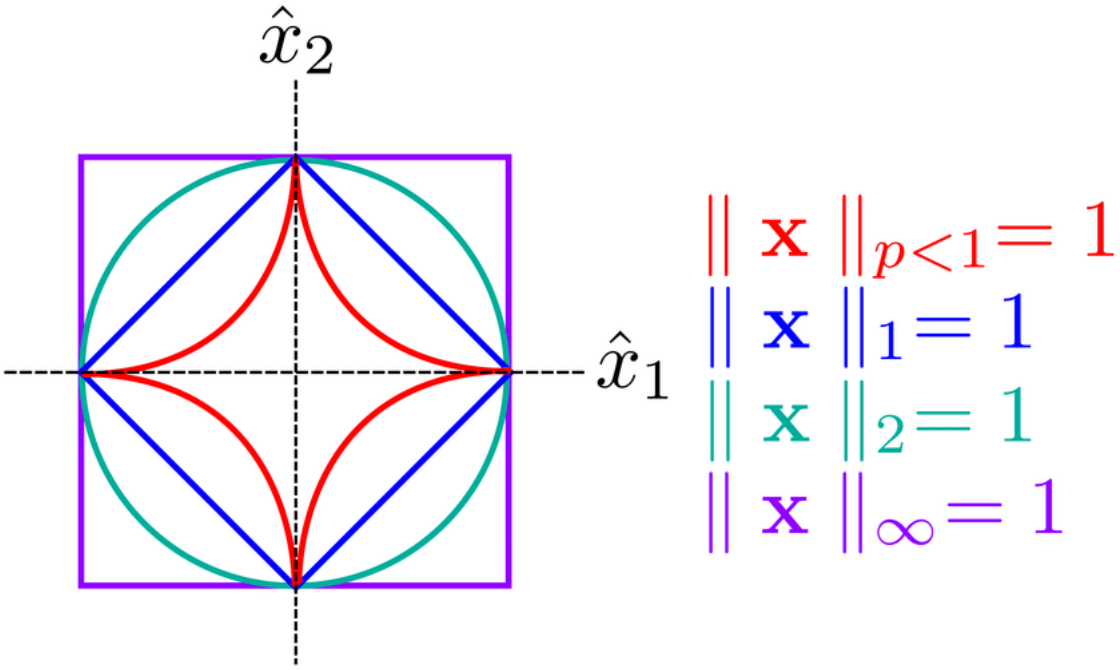
We, got a square which is rotated $45°, i.e$ a rhombus.(unit circle for ℓ_1 norm)
For ℓ_2,

$$\ell_2 = \|X\|_2 = (\sum_{i=1}^{d} x_i^2)^{\frac{1}{2}} \iff x_1^2 + x_2^2 = 1 \tag{9}$$

This is an equation of a circle with radius one.(unit circle for ℓ_2 norm)
For ℓ_∞,

$$\begin{aligned}\ell_\infty &= max_i\{|1|, |x_2|\}\ \forall\, x_2 \leq 1 \implies x_1 = 1\\ &= max_i\{|-1|, |x_2|\}\ \forall\, x_2 \leq 1 \implies x_1 = -1\\ &= max_i\{|x_1|, |1|\}\ \forall\, x_1 \leq 1 \implies x_2 = 1\\ &= max_i\{|x_1|, |-1|\}\ \forall\, x_1 \leq 1 \implies x_2 = -1\end{aligned} \tag{10}$$

This gives us a square (unit circle for ℓ_∞ norm). They are shown in the below *figure*.



Source:*Research gate*
Figure 1 : Plot of different norms.

We can notice that unit circle for ℓ_1 norm is less than ℓ_2 norm, which indeed is less than ℓ_∞ norm. This tells us,

$$\Rightarrow \boxed{\|x\|_1 \geq \|x\|_2 \geq \|x\|_\infty} \tag{11}$$

Why is that so? Because the unit circle for $\|x\|_1$ norm is smaller than the unit circle for $\|x\|_2$ norm, *i.e*, for same values of x_1 and x_2, we get larger value for ℓ_1 compared to ℓ_2. Similarly, this applies for ℓ_∞ norm as well. For example,

$$\begin{aligned}&For\ x_1 = 0.6\ and\ x_2 = 0.3,\\ &\qquad \|x\|_1 = 0.6 + 0.3 = 0.9,\ and\\ &\qquad \|x\|_2 = \sqrt{0.6^2 + 0.3^2} = 0.67\\ &\qquad \|x\|_\infty = max_i\{0.6, 0.3\} = 0.6\end{aligned} \tag{12}$$

This, helped us to complete the first half of the question, which told us to prove $\boxed{\|x\|_\infty \leq \|x\|_2}$ and $\boxed{\|x\|_\infty \leq \|x\|_1}$.

For the second half, instead of me proving for d and $\sqrt{d}$ individually, we will take a generalised approach.
We know, that below equation is true as we approximated the ℓ_∞ norm.

$$|x_i| \leq \|x\|_\infty,\ \forall\, i = \{1, 2, .., d\}$$

Now let's raise the power of the above equation to p. As, $1 \leq p < \infty$ the sign of inequality does not change.

$$|x_i|^p \leq \|x\|_\infty^p, \; \forall \, i = \{1, 2, .., d\}$$

We have this equation for every $i = \{1, 2, .., d\}, i.e$

$$\begin{aligned} |x_1|^p &\leq \|x\|_\infty^p \\ |x_2|^p &\leq \|x\|_\infty^p \\ &. \\ &. \\ &. \\ |x_d|^p &\leq \|x\|_\infty^p \end{aligned} \tag{13}$$

Adding all the terms in the above equation set (13), we get the below equation. Notice here that d gets multiplied in the RHS as there are d possible values of x_i.

$$|x_1|^p + |x_2|^p + ... + |x_d|^p \leq d.\|x\|_\infty^p$$

Taking the p^{th} root and simplifying further.

$$\begin{aligned} &(|x_1|^p + |x_2|^p + ... + |x_d|^p)^{\frac{1}{p}} \leq (d.\|x\|_\infty^p)^{\frac{1}{p}} \\ &(\sum_{i=1}^{d} |x_1|^p)^{\frac{1}{p}} \leq (d.\|x\|_\infty^p)^{\frac{1}{p}} \\ &\implies \|x\|_p \leq d^{\frac{1}{p}} (\|x\|_\infty^p)^{\frac{1}{p}} \\ &\implies \boxed{\|x\|_p \leq d^{\frac{1}{p}}.\|x\|_\infty} \end{aligned} \tag{14}$$

Now, if we substitute $p = 1$ for $\|x\|_1$ norm and $p = 2$ for $\|x\|_2$ norm. We get,

$$\boxed{\|x\|_1 \leq d.\|x\|_\infty} \; and \; \boxed{\|x\|_2 \leq \sqrt{d}.\|x\|_\infty}$$

Combining equation **(11)** and **above** , we obtain our answers.

$$\boxed{\|x\|_\infty \leq \|x\|_2 \leq \sqrt{d}\|x\|_\infty \, , \; and, \; \|x\|_\infty \leq \|x\|_1 \leq d\|x\|_\infty}$$