Second Home Assignment
Data Analysis

Nikolaj Dybdahl Rathcke (rfq695)

May 10, 2015

Question 1

(1)

We want to prove that

$$\mathbb{P}\left\{(X - E[X])^2 \geq \varepsilon\right\} \leq \frac{Var[X]}{\varepsilon} \tag{1}$$

Markovs inequality states that

$$\mathbb{P}\left\{X \geq a\right\} \leq \frac{E[X]}{a} \tag{2}$$

for some constant ε. If we substitute X with $(X - E[X])^2)$ and a for ε in (2) we get

$$\mathbb{P}\left\{(X - E[X])^2) \geq \varepsilon\right\} \leq \frac{E[(X - E[X])^2)]}{\varepsilon}$$

Now by definition, we know that $Var[X] = E[(X - E[X])^2)]$. So we can rewrite the right hand side to get

$$\mathbb{P}\left\{(X - E[X])^2) \geq \varepsilon\right\} \leq \frac{Var[X]}{\varepsilon}$$

which means we have shown (1).

(2)

Now we want to prove that

$$\mathbb{P}\left\{|(X - E[X])| \geq \varepsilon\right\} \leq \frac{Var[X]}{\varepsilon^2} \tag{3}$$

We square both sides of the inequality on the left hand side to get

$$\mathbb{P}\left\{(X - E[X])^2 \geq \varepsilon^2\right\} \leq \frac{Var[X]}{\varepsilon^2}$$

We can by definition write $Var[X]$ as $E[(X - E[X])^2)]$, so

$$\mathbb{P}\left\{(X - E[X])^2 \geq \varepsilon^2\right\} \leq \frac{E[(X - E[X])^2)]}{\varepsilon^2}$$

This matches Markovs inequality (2) where $X = (X - E[X])^2)$ and $a = \varepsilon^2$, i.e

$$\mathbb{P}\left\{X \geq a\right\} \leq \frac{E(X)}{a}$$

And thus we have shown (3).

(3)

We will prove that

$$\mathbb{P}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}X_i-\mu\right)^2\geq\varepsilon\right\}\leq\frac{\sigma^2}{N\varepsilon} \tag{4}$$

The summation on the LHS is actually the definition for variance, so

$$\mathbb{P}\left\{\left(\frac{1}{N}\sum_{i=1}^{N}X_i-\mu\right)^2\geq\varepsilon\right\}\leq\frac{\sigma^2}{N\varepsilon}$$

The sum is equal to the sample mean, $\bar{X}$, and $\mu = E[\bar{X}]$, so

$$\mathbb{P}\left\{\left(\bar{X}-E[\bar{X}]\right)^2\geq\varepsilon\right\}\leq\frac{\sigma^2}{N\varepsilon}$$

By definition $Var[\bar{X}] = \frac{\sigma^2}{N}$, which mean we can rewrite the RHS

$$\mathbb{P}\left\{\left(\bar{X}-E[\bar{X}]\right)^2\geq\varepsilon\right\}\leq\frac{Var[\bar{X}]}{\varepsilon}$$

And using definition $Var[X] = E[(X - E[X])^2)$

$$\mathbb{P}\left\{\left(\bar{X}-E[\bar{X}]\right)^2\geq\varepsilon\right\}\leq\frac{E[(\bar{X}-E[\bar{X}])^2)}{\varepsilon}$$

Replace $(\bar{X} - E[\bar{X}])^2$ with X and set $\varepsilon = a$ and we get the Markov inequality (2).

$$\mathbb{P}\left\{X\geq a\right\}\leq\frac{E[X]}{a}$$

And we have shown that (4) holds.

Question 4

An example of low euclidian distance similarity (far from each other) and high cosine similarity is the vectors

$$x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } x' = \begin{pmatrix} 100 \\ 0 \end{pmatrix}$$

These have an euclidian distance of

$$\sqrt{(1-100)^2 + (0-0)^2} = 99$$

while it has a cosine similarity of

$$\frac{\begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} 100 \\ 0 \end{pmatrix}}{\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} \| \| \begin{pmatrix} 100 \\ 0 \end{pmatrix} \|} = \frac{100}{101}$$

That they have a cosine similarity of 1 means it has the highest value possible, while the euclidian distance of 99 is quite high.

Now, an example of low cosine similarity but high euclidan distance similarity (close to each other) would be

$$x = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \text{ and } x' = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$$

These have an euclidian distance of

$$\sqrt{(1-0)^2 + (0+1)^2} = \sqrt{2}$$

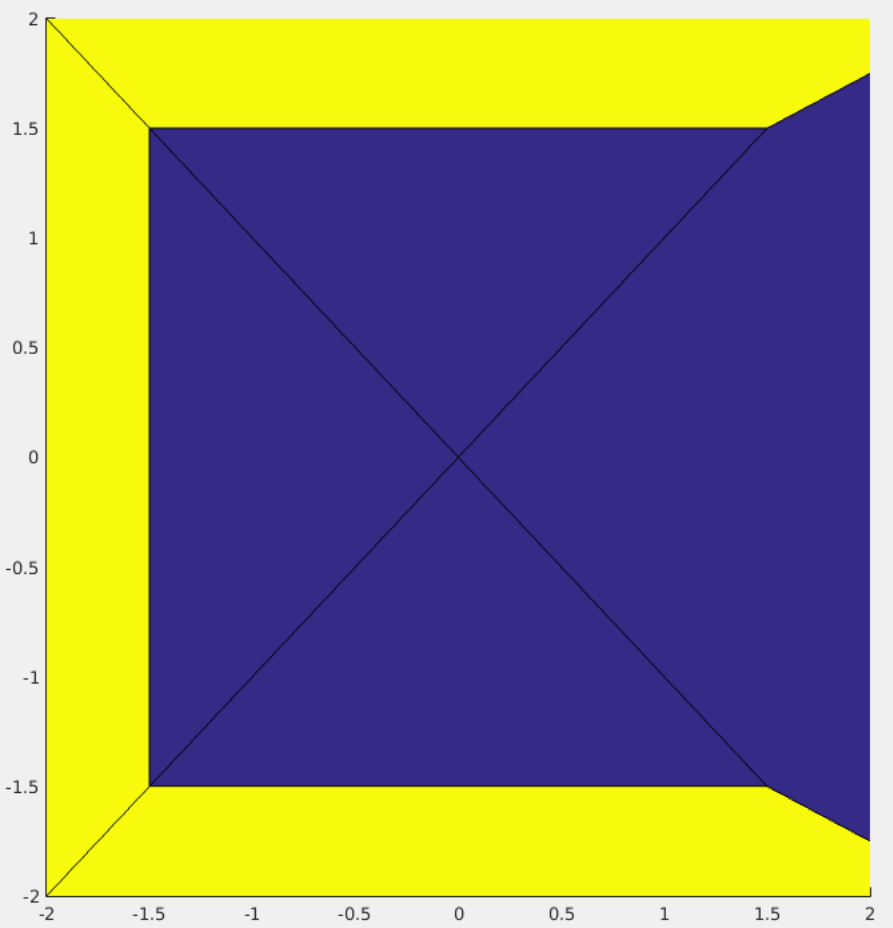
while it has a cosine similarity of

$$\frac{\begin{pmatrix} 1 \\ 0 \end{pmatrix} \cdot \begin{pmatrix} -1 \\ 0 \end{pmatrix}}{\| \begin{pmatrix} 1 \\ 0 \end{pmatrix} \| \| \begin{pmatrix} -1 \\ 0 \end{pmatrix} \|} = -1$$

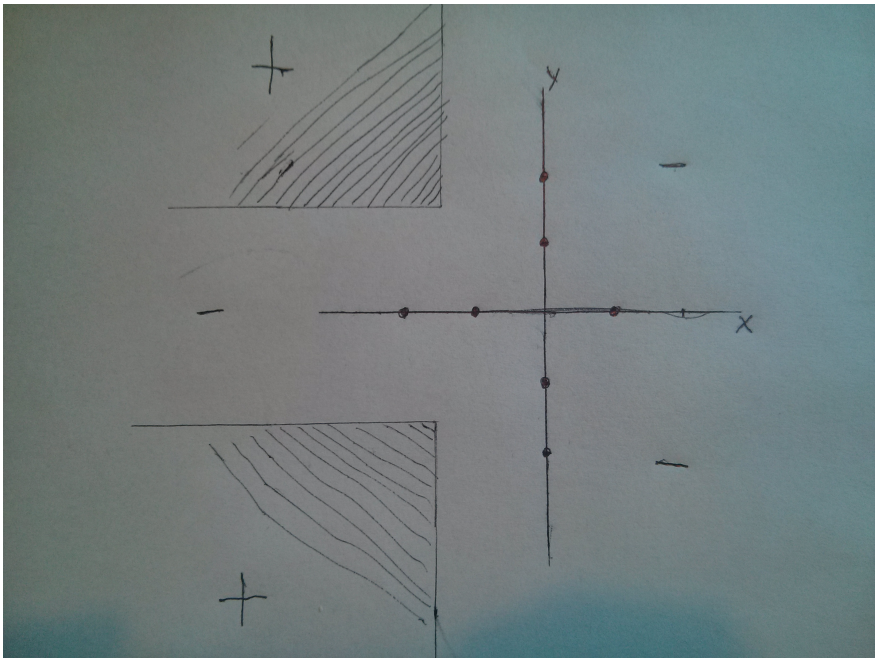
So they have the lowest possible cosine similarity, while the euclidian distance is very low.

Question 5

The decision regions for the $1 - NN$ rule can be seen below



The code used to generate it can be found in the `q5` file.
The regions for the $3 - NN$ rule is seen below.



The 'corner points' that separate those classified as -1 (noted as -) and $+1$ (noted as +) are located in $(-1.5, 1.5)$ and $(-1.5, -1.5)$.

Question 6

(1)

The code that implements this is in `q6`. It takes the first 500 images (as there weren't 1000 of each digit) corresponding to a digit from the test set, and for each of these it compares to the images from the training set (which is a mix of 2 digits). Measuring the euclidian distance it finds the K closest matches from the training set. If the majority of them are the right digit, it's considered a success.
Running the code with labels 0 and 1, 0 and 8 and 5 and 6 will provide the graphs (figure 1, 2 and 3) where E_{in} is a function of iterations N. Note that it's only compared one way, i.e. 0's are picked from the test set and compared to the mix of 0's and $1's$ from the training and the same with the other two cases.

Note: I realize now that my E_{in} is not calculated correctly as the calculation I made for them is outside the right for-loop. So it will always divide by 500 instead of dividing by how many times we have looped through (how many observations we have made). I do not have time to change this, but I imagine the plots will look a lot different. They will most likely converge toward some percentage falling under and over until it evens out. Conclusions could then be made on how large E_{in} is at different iterations. We would want a large E_{in} as that means it often guesses the correct digit.

K-NN algorithm not implemented.

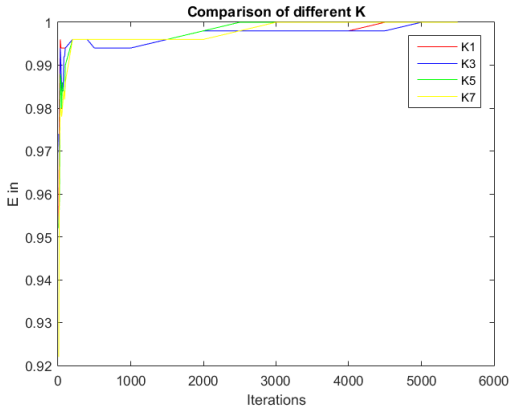


Figure 1: Comparison of 0's and 1's - E_{in} as a function of iterations

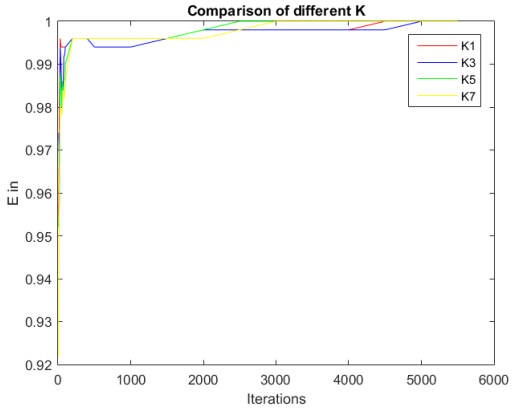


Figure 2: Comparison of 0's and 8's - E_{in} as a function of iterations

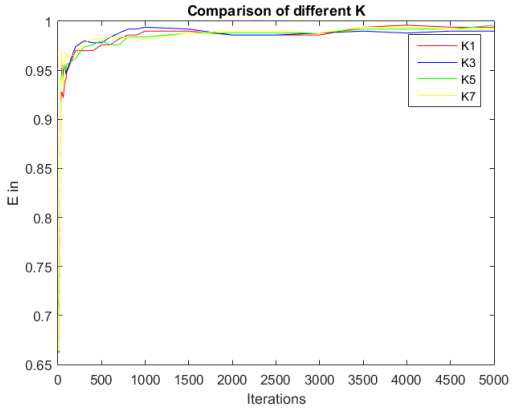


Figure 3: Comparison of 5's and 6's - E_{in} as a function of iterations