Programming Assignment

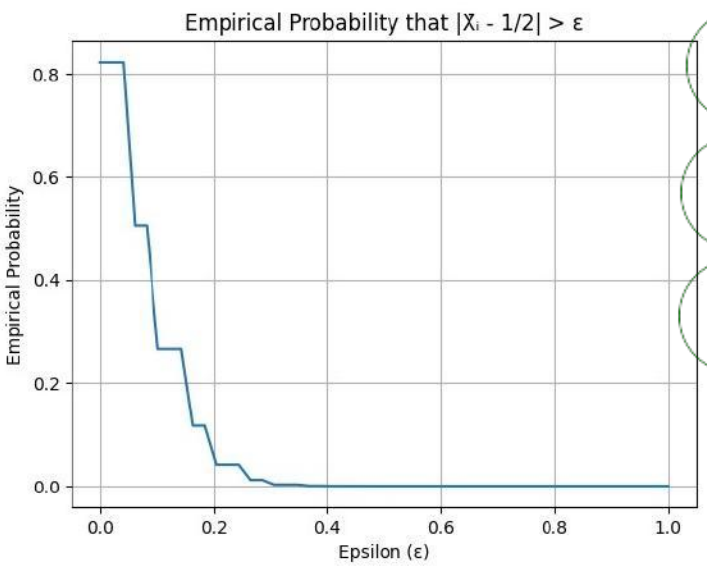
1. **Visualizing the Hoeffding bound (10 pts).**

   (a) Use **numpy** to generate an $N \times n$ matrix of samples from $Bernoulli(1/2)$. Calculate for each row i the empirical mean, $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j}$, where $N = 200000$ and $n = 20$.

   (b) Take 50 values of $\epsilon \in [0,1]$ (**numpy.linspace(0,1, 50)**), and calculate the empirical probability that $|\bar{X}_i - 1/2| > \epsilon$. Plot the empirical probability as a function of ϵ.

   (c) Add to your plot the Hoeffding bound of that probability, as a function of ϵ.

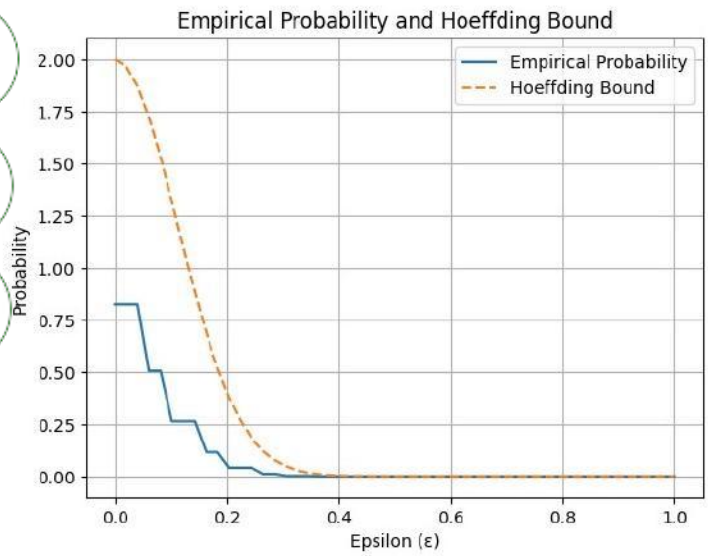
   Submit your plots (no need to submit code for this question).



(6.1)

(6.2)

(6.3)



2. **Nearest Neighbor (20 pts).** In this question, we will study the performance of the Nearest Neighbor (NN) algorithm on the MNIST dataset. The MNIST dataset consists of images of handwritten digits, along with their labels. Each image has 28×28 pixels, where each pixel is in gray-scale, and can get an integer value from 0 to 255. Each label is a digit between 0 and 9. The dataset has 70,000 images. Although each image is square, we treat it as a vector of size 784.

(7.1)

   (b) Run the algorithm using the first $n = 1000$ training images, on each of the test images, using $k = 10$. What is the accuracy of the prediction (i.e. the percentage of correct classifications)? What would you expect from a completely random predictor?

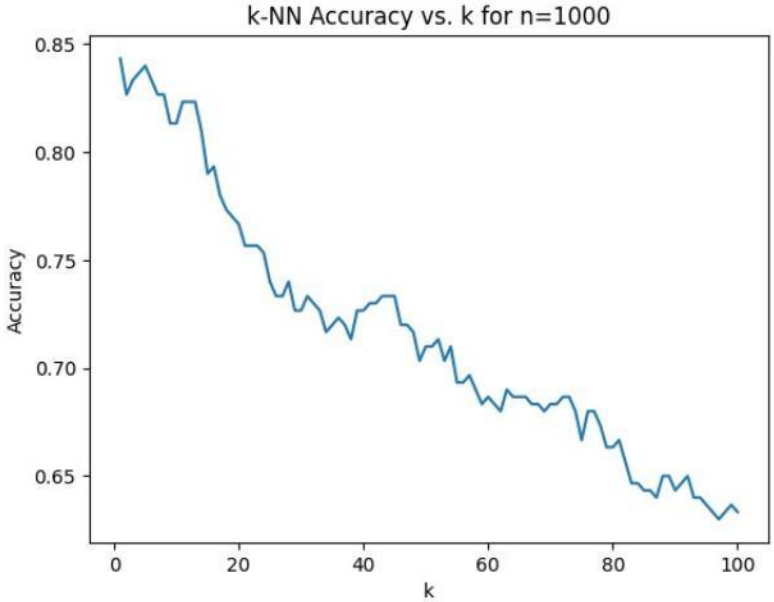
(7.2)

• בכל שבמות הדגימות שואפת ל- 86% האמון מתקבץ על

• מאחר ורנדומלי נצפה שהוא יבחר מספר בין 0-9 באופן שרירותי ולכן לכל דגימה יש 10% דיוק

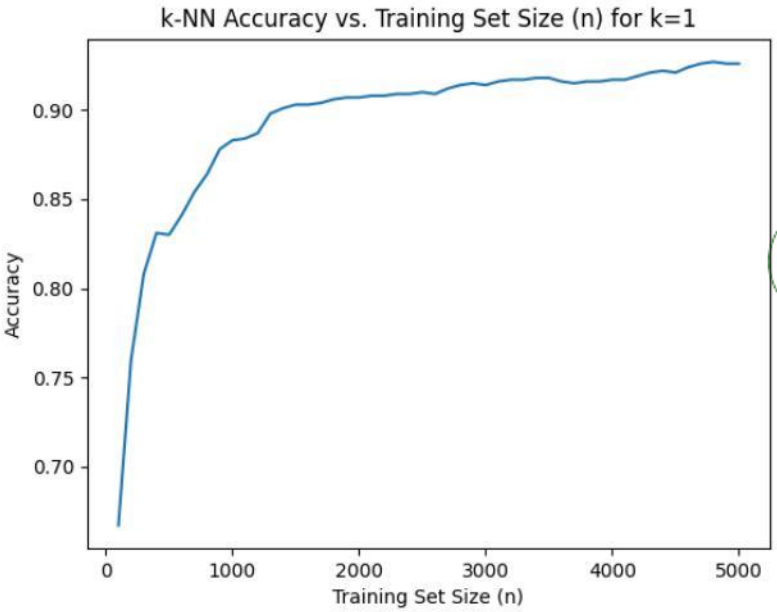
   (c) Plot the prediction accuracy as a function of k, for $k = 1, \ldots, 100$ and $n = 1000$. Discuss the results. What is the best k?

לפי התוצאות ניראה כי ה-K הכי טוב הוא K=5 או K=1. בעצם ה-K שנתן את הדיוק הגבוה ביותר.
ניראה כי ערכים הקטנים של K (1-5) הנו דיוק גבוה. יש הסבר מאחורי זה כי כאשר ה-K גדול, המאוזן מתחיל לשקול שכנים
רחוקים מידי שאינם קרובים לווקטור (מידע חשוב). עם זאת בתרחיש K=1 יכולה לפגוע להגיע ל- overfitting. לכן 1 ניראה
אופטימלי יותר אבל ההפרש בין 1 ל-5 מצער וגם הכל משתנה אם כמות ה-test Data. הגרף שיצא בעמוד הבא:



(7.3)

(d) Using $k = 1$, run the algorithm on an increasing number of training images. Plot the prediction accuracy as a function of $n = 100, 200, \ldots, 5000$. Discuss the results.



(7.4)

כפי שניתן לראות בגרף ככל שעולים במספר ה- Training Data הדיוק עולה עבור K=1 אבל בשל כלשהו (באזור 1200) יש האטה בקצב הדיוק שנפל לאט בעצר ה-90%. מסקנות:

1. חשיבות ה-Training Data - העלייה המהירה בהתחלה מעידה כמה חשוב להחזיק באוסף גדול לצורך דיוק האלגוריתם.

2. עקומת למידה: בהתחלה המודל לומד במהירות אבל לבסוף חווה ירידה בדיוק עקב רוויה של גודל הקלט.

3. Overfitting עבור n קטנים: ב- n = 100 רואים שבחירת K = 1 שגוי.

4. גודל אופטימלי של קבוצת Training Data: בגלל שנראה כי יותר נתונים עדיפים, תלויה בזאת מראה Trade off, יש כמות אופטימלית של תמונות שבה אתה מסיג דיוק מול מעבר לנק' האופטימלית גם יתכן שלאת עיבוד ואיחסון הנתונים אינה מדהקת לשפר האלגוריתם.

(8)