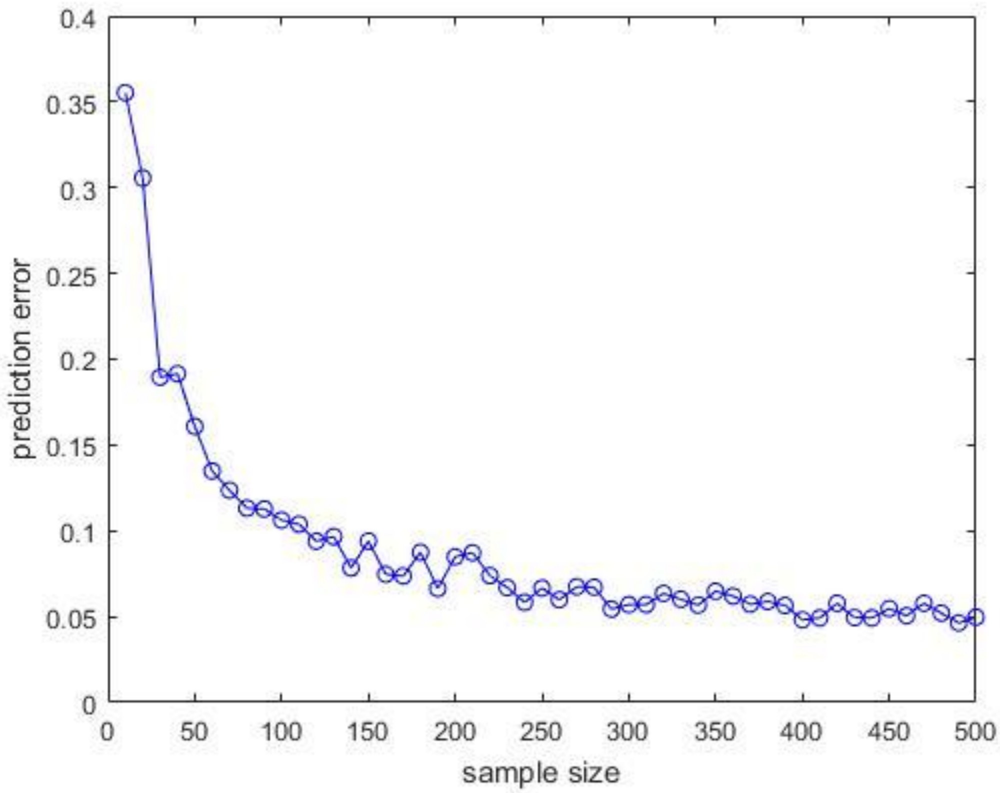
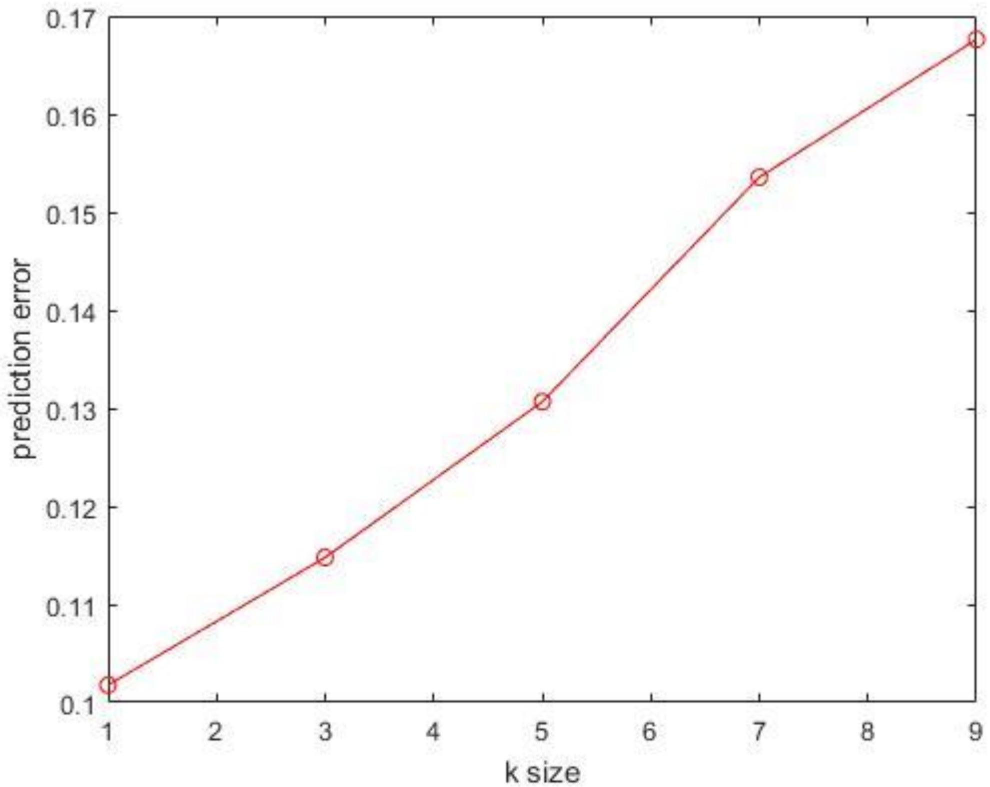
2. a.



2. b. Yes. As the sample size increases, the prediction error generally decreases. This is due to the fact that with growing sample size, we have a better probability of encountering an example that is either already in the sample, or very close to one. The algorithm then predicts the label correctly for that example.

2. c. When the sample sizes are small, the graph generally appears non-smooth. This is because when the sample size is small, we have a large probability of encountering examples that aren't similar to anything in our sample. In that case, our error will be high. This means that between several runs with small sample sizes, the prediction error can vary greatly, which will make the graph appear non-smooth.

2. d.



2. e. Yes. As k increases, so does the prediction error. When k increases, we take into account more examples which may be further away from our example. This increases the probability of false prediction.

2. f. for m=100, k=1:

91.6832	0.8911	6.0396	1.3861
1.0183	93.6864	2.8513	2.4440
14.5740	1.9058	77.4664	6.0538
0.5219	1.3570	2.7140	95.4071

We can observe that on the main diagonal of the confusion matrix, the percentages are significantly higher than the other cells. This means most of the numbers were predicted correctly.

Furthermore, the cell not in the main diagonal with the highest percentage is (3,1), which correspond to the number 5 being confused by 3. This is also not surprising, since 5 and 3 are quite similar in their shapes. Also, the cell on the main diagonal with the lowest percentage is (3,3) which corresponds to the number 5 being mistaken the most.

$(*)\ P_{(X,Y)\sim D_a}[X=a]=1 \rightarrow \eta(a) \overset{(**)}{=} P_{(X,Y)\sim D_a}[Y=1 \mid X=a] \underset{(*)}{=} P_{(X,Y)\sim D_a}[Y=1]$

ⓐ $P[\hat{h}_S(x)=1] \underset{(*)}{=} P[\hat{h}_S(a)=1] = P[y_{NN}(a)=1] =$

(כיוון שמתן כי הבחירה אקראית של הסמפלים אותה נקודה) $= \dfrac{\sum_{(a,Y_i)\in S} \mathbb{I}(Y=1)}{m} = \dfrac{|\{(a,Y)\in S \mid Y=1\}|}{m} \equiv \beta_S$

ⓑ $err(\hat{h}_S, D_a) := P_{(X,Y)\sim D_a}[\hat{h}_S \neq Y] \underset{(*)}{=} P_{(X,Y)\sim D_a}[\hat{h}_S(a)\neq Y] =$

$= P[(\hat{h}_S(a)=0 \wedge Y=1) \vee (\hat{h}_S(a)=1 \wedge Y=0)] =$ (ז״ר)

$= P[\hat{h}_S(a)=0 \wedge Y=1] + P[\hat{h}_S(a)=1 \wedge Y=0] =$ (ב״ת)

$= P[\hat{h}_S(a)=0]\cdot P[Y=1] + P[\hat{h}_S(a)=1]\cdot P[Y=0] =$

$\underset{(**)}{=} (1-\beta_S)\,\eta(a) + \beta_S(1-\eta(a)) = \eta(a) - 2\beta_S\eta(a) + \beta_S$

ⓒ $E_{S\sim D_a^m}[\beta_S] = E_{S\sim D_a^m}\left[\dfrac{|\{(a,Y)\in S \mid Y=1\}|}{m}\right] =$ (לינאריות תוחלת

$= \dfrac{1}{m} E_{S\sim D_a^m}[\,|\{(a,Y)\in S \mid Y=1\}|\,] = \left(\sim Bin\ B(m,\eta(a))\right)$

$= \dfrac{m\cdot\eta(a)}{m} = \boxed{\eta(a)}$

ⓓ $\overline{err} = E_{S\sim D_a^m}[err(\hat{h}_S, D_a)] = E(\eta(a) - 2\beta_S\eta(a) + \beta_S) =$

לינאריות התוחלת $= \eta(a) - 2\eta(a)\cdot E(\beta_S) + E(\beta_S) = \eta(a) - 2\eta^2 a + \eta(a) = 2\eta(a) - 2\eta^2(a)$

(e)

נחלק למקרים:

אם $\eta(a)>\frac{1}{2}$, $h^*(x)=1$ (כי $P_{(x,y)\sim D_a}(Y=1)>\frac{1}{2}$)

אם $\eta(a)\le\frac{1}{2}$, $h^*(x)=0$ (כי $P_{(x,y)\sim D_a}[Y=1]\le\frac{1}{2}$)

לכן:

$$err_{bayes}=err(h^*,D_a)=P_{(x,y)\sim D_a}[h^*(x)\neq Y]=\begin{cases}P(Y=0) & \eta(a)>\frac{1}{2}\\ P(Y=1) & \eta(a)\le\frac{1}{2}\end{cases}$$

$$=\begin{cases}1-\eta(a) & \eta(a)>\frac{1}{2}\\ \eta(a) & \eta(a)\le\frac{1}{2}\end{cases}$$

(f)

יהי $\varepsilon>0$ נקבע $\eta(a)=\max\{\frac{3}{4},1-\frac{\varepsilon}{4}\}$

אז כיוון ש- $\eta(a)\ge\frac{3}{4}>\frac{1}{2}$ נקבל:

$$\frac{err}{err_{bays}}=\frac{2\eta(a)-2\eta^2(a)}{1-\eta(a)}=2\eta(a)>2-\varepsilon$$

כיוון ש- עבור $1>\varepsilon>0$ מתקיים:

וגם $\eta(a)=1-\frac{\varepsilon}{4}>\frac{3}{4}$ $\quad$ $2\eta(a)=2-\frac{\varepsilon}{2}>2-\varepsilon$

ועבור $1\le\varepsilon$ מתקיים:

וגם $\eta(a)=\frac{3}{4}>1-\frac{\varepsilon}{4}$ $\quad$ $2\eta(a)=\frac{3}{2}>2-\varepsilon$