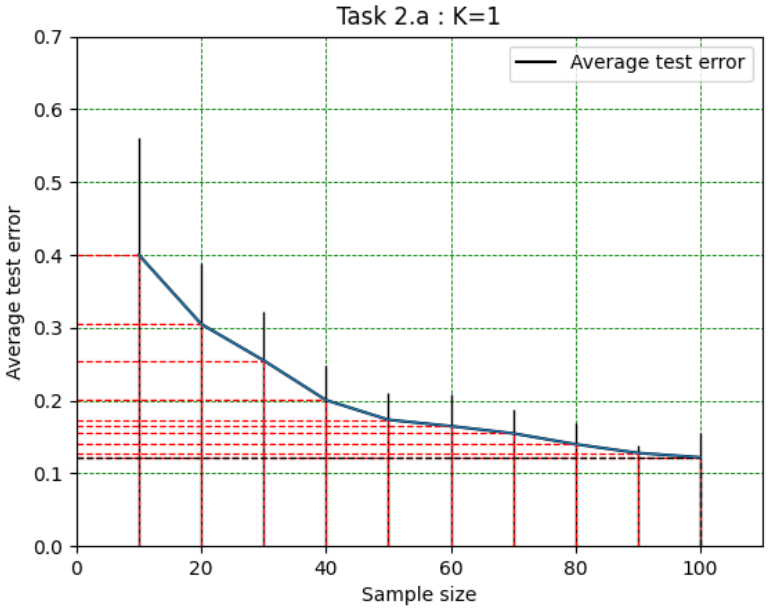
Introduction to Machine Learning
Exercise 1



2b

המגמה שאנו מבחינים בה בגרף הינה מגמת ירידה כאשר ככל שנגדיל את גודל המדגם כך תרד השגיאה הממוצעת. אפשר להסביר זאת בכך שכאשר גודל המדגם גדל, כך יש לנו יותר מידע שבאמצעותו נאמן את המודל שלנו ליצירת פונקציה מדויקת יותר.

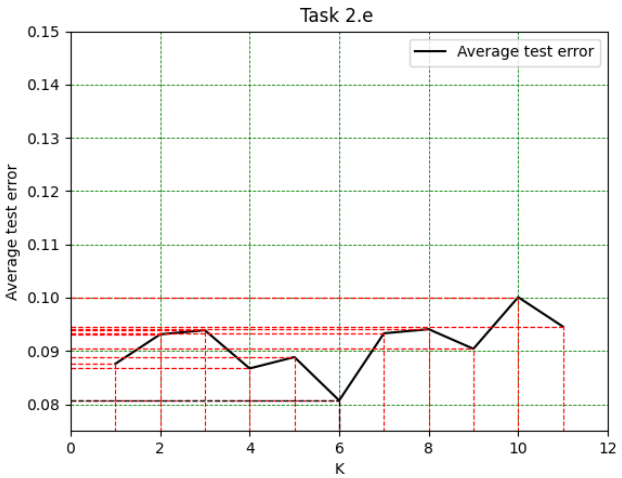
2c

כן, והסיבה לכך היא שהמדגם נבחר באופן רנדומלי, ואפילו שהגודל זהה- הדוגמאות עצמן שונות- והן אלה שקובעות את איכות האימון והפונקציה שלנו.

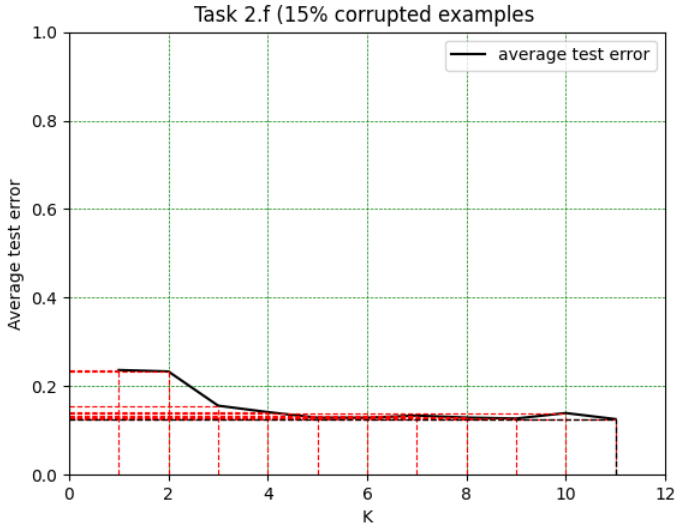
2d

ניתן לראות שככל שגודל המדגם גדל כך טווח בר הטעות קטן, ניתן להסביר זאת בכך שככל שנגדיל את המדגם כך יהיו לנו יותר דוגמאות שבאמצעותן נאמן את המודל שלנו, לכן הוא יהיה מדויק יותר וכך הוא יהיה פחות רגיש לדוגמאות קיצוניות אותן נצטרך לסווג.

2e



2f



2g

ניתן לראות כי בניסוי הראשון ה-k האופטימלי הוא 6 ואילו בניסוי השני ה-k האופטימלי הוא 11.
אכן קיים הבדל בין שני הניסויים והוא שהשגיה גדלה כאשר שגינו במדגם. ה-k משפיע על איכות הלמידה ועל השגיאה שלה, לכן ה-k האופטימלי גדל.

3a

Let $S \sim D^m$, $y_1 \neq y_2$ and $Y = \{0,1\}$ and let $(x_1, y_1),(x_2, y_2) \in S$.

D is c-Lipschitz, thus $x_1, x_2 \in SUPP(D)$ and $|\eta(x_1) - \eta(x_2)| \leq c \cdot \rho(x_1, x_2)$.

D has a Bayes-error of zero and it's deterministic, and because of that $\eta(x) = \{0,1\}$.

$y_1 \neq y_2$ so $\eta(x_1) \neq \eta(x_2)$ and therefore $|\eta(x_1) - \eta(x_2)| = 1$.

so $1 = |\eta(x_1) - \eta(x_2)| \leq c \cdot \rho(x_1, x_2)$ $\rightarrow$ $1 \leq c \cdot \rho(x_1, x_2)$

$1 \leq c \cdot ||x_1 - x_2||$ $\rightarrow$ $\frac{1}{c} \leq ||x_1 - x_2||$

3b

Let B be a set of balls of radius $r = \frac{1}{3c}$ that cover the space of points X.

Let's assume that $err(f_s^{nn}, D) \neq 0$, so that there is an $x \in X$ that for him $f_s^{nn}(x) \neq y$.

We'll look at $x's$ nearest neighbor: x'.

According to **3a** if $||x - x'|| \leq \frac{1}{c}$ then $y = y'$, which means: $f_s^{nn}(x') \neq y$, or in other words: $y \neq y'$.

In the worst case x' is not in the same ball as x. So, since $r = \frac{1}{3c}$: $||x - x'|| \leq \frac{2}{3c}$.

$\rightarrow$ $||x - x'|| \leq \frac{2}{3c} < \frac{1}{c}$ and because of c-Lip we can conclude that: $y = y'$.

That is contrary to our assumption.

4a

$\chi = \{x \mid x \in R^2\}$ so that the first dimension of x is the rabbit's age, and the second is its weight.

$Y = \{0, 1\}$ so that black is 1, white is 0.

4b

$h_{bayes}(5, 2) = 0$, $h_{bayes}(12, 1) = 1$, $h_{bayes}(12, 2) = 0$

4c

$$err(h_{bayes}, D) = \sum_{x \in X} P[h_{bayes}(x) \neq y] = 0.08 + 0.04 = 0.12$$

4d

$h_1(x)$ *is the function that labels each* x *to be* 0. $h_2(x)$ *is the function that labels each* x *to be* 1.

$err(h_1) = 0.28$, $err(h_2) = 0.72$

$err_{app}(H) = inf_{h \in H}\ err(h, D) = err(h_1) = 0.28$

4e

$$h_{bayes} = \begin{cases} 0, & x(1) \leq 25 \\ 1, & otherwise \end{cases}$$

4f

No, we can't calculate the error of the Bayes optimal predictor we provided because we are missing essential information about the probability of each rabbit to be in a certain age, and the function is not continuous.

4g

$$E_{S \sim D^m}[err(h_{bayes}, D)] = \frac{k-1}{k} \sum_{x \in X} p_x (1 - p_x)^m =$$

$$= \frac{1}{2}(0.06 * 0.94^5 + 0.12 * 0.88^5 + 0.53 * 0.47^5 + 0.29 * 0.71^5) = \frac{1}{2} * 0.1784 = 0.0859$$

We can use this formula for D'' because it is deterministic, contrary to D which is not.

5a

let sample size = m. according to what we've learned in class :

$$m \geq \frac{log(|H|) + log\frac{1}{\delta}}{\varepsilon} = \frac{log(|N+1|) + log\frac{1}{0.05}}{0.03}$$

5b

$err(f_\alpha, D) = P_{(x,y) \sim D}[f_\alpha(x) \neq y] = P[f_\alpha = 1 \mid y = 0] \vee P[f_a = 0 \mid y = 1] =$

$= P(\alpha \leq x < \beta) \vee P(\beta \leq x < \alpha).$

As we know $a \in [\beta - \varepsilon, \beta + \varepsilon]$, thus what we found is an exclusive or. Both cases can't appear together.

case 1: $P(\alpha \leq x < \beta) \leq P(\beta - \varepsilon \leq x < \beta) =^{*} \beta - \beta + \varepsilon = \varepsilon$
* D uniformly distributes
case 2: $P(\beta \leq x < \alpha) \leq P(\beta \leq x < \beta + \varepsilon) =^{*} \beta + \varepsilon - \beta = \varepsilon$
In both cases $err(f_{\alpha}, D) \leq \varepsilon$.

5c

Let (x_1, y_1) so that $x_1 \in [\beta, \beta + \varepsilon]$, (x_2, y_2) so that $x_2 \in [\beta - \varepsilon, \beta]$. Since D is realizable and $\beta \in [0,1]$, then $err(f_{\beta}, D) = 0$, S is distributed by D^m so also $err(f_{\beta}, S) = 0$.
Since the error on the sample is zero, $f_{\beta}(x_1) = 1$, $f_{\beta}(x_2) = 0$. We'll call h_s the output of the $ERM_{th}\ algorithm$ and therefore $h_s(x_1) = 1$, $h_s(x_2) = 0$. As we know the algorithm chooses a single classifier from H_s, thus exists $a \in [0,1]$ that satisfies $h_s(x) = f_{\beta}(x) = I[x \geq a]$.
a can't be bigger than $x_1$or smaller than x_2 since it will cause an error on the sample, thus $a \in [x_2, x_1]$. We'll remember that $x_1 \in [\beta, \beta + \varepsilon]$ and $x_2 \in [\beta - \varepsilon, \beta]$, so $a \in [\beta - \varepsilon, \beta + \varepsilon]$ and according to the last question (b) we know that $err(f_{\alpha}, D) \leq \varepsilon$.

5d

x is uniformly distributed over [0,1] and that'ss why :
$P(\forall x \in X, x \notin [\beta - \varepsilon, \beta]) =$

$$\prod_{x \in X} P(x \notin [\beta - \varepsilon, \beta]) = P(x \notin [\beta - \varepsilon, \beta])^m = (1 - P(x \in [\beta - \varepsilon, \beta]))^m =$$

$$= (1 - \varepsilon)^m.$$

In the same way:
$P(\forall x \in X, x \notin [\beta, \beta + \varepsilon]) =$

$$\prod_{x \in X} P(x \notin [\beta, \beta + \varepsilon]) = P(x \notin [\beta, \beta + \varepsilon])^m = (1 - P(x \in [\beta, \beta + \varepsilon]))^m =$$

$$= (1 - \varepsilon)^m.$$

So , $P(err(h_s, D) \leq \varepsilon) = P(\exists x \in [\beta - \varepsilon, \beta] \wedge \exists x \in [\beta, \beta + \varepsilon]) =$

$$= 1 - [P(\forall x \in X, x \notin [\beta - \varepsilon, \beta]) + P(\forall x \in X, x \notin [\beta, \beta + \varepsilon])] = 1 - 2(1 - \varepsilon)^m.$$

5e

$P_{S \sim D^m}[err(h_s, D) \leq 0.03] \geq 1 - 0.05 \geq 1 - 2(1 - 0.03)^m$

$0.95 \geq 1 - 2(0.97)^m$ **=>** $0.025 \geq (0.97)^m$ **=>** $m \geq 121.1$
Thus m, which is the minimal sample size is 122.

5f

The better approach is the one we used in e, and that's since in question a the sample size is dependable on |H| .H is infinite now because the threshold is now rational and $N \to \infty$.
Thus, we will prefer using the method of question e.