Big Data – Assignment 1

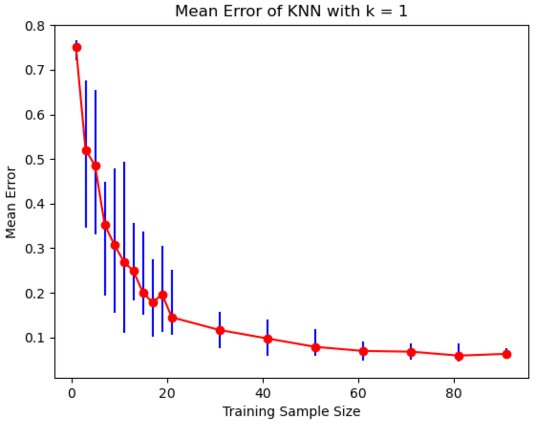
Yotam Lifschytz – 209579077, Pan Eyal – 208722058

Question 1:

The code files are attached in the zip file.

Question 2:

a)

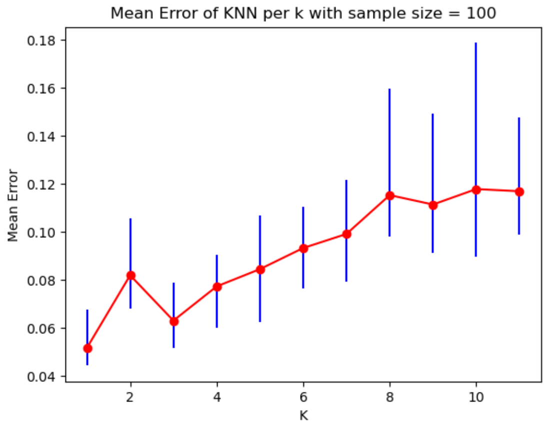


b) We can observe that as the Training Sample Size grows, the Mean Error reduces.
When we see more examples, our sample represents the distribution more accurately, therefor we expect to receive a smaller error for an ERM algorithm such as K-NN. Also, we theorize that a handwritten-digit distribution is pretty much "c-Lipchitz", as close examples have the same label, so in this case as $m \to \infty$ the error will go to the bayes optimal.

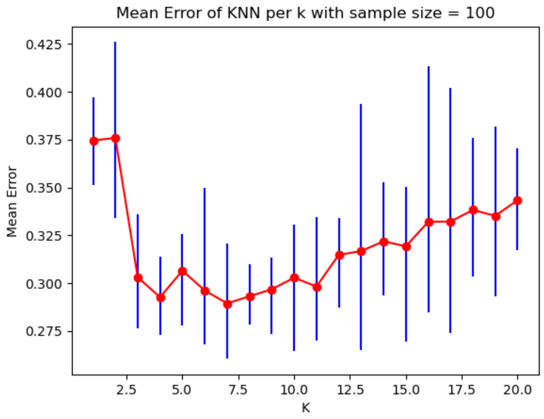
c) We get different results on different runs with the same sample size because the sample is chosen probabilistically, thus is different each time, and the accuracy of the internal model depends on how well the sample represents the distribution. When sample size is low, the variance of accuracy can be large.

d) The error-bars start large and decrease with sample size. This is exactly as stated in (c), because the variance is caused by large variances in accuracy of the model, which depends on how well the sample represents the distribution. With small sample sizes, this can vary greatly between samples. As the sample sizes grow, the accuracy of representation converges and variance lessens (pretty much the law of large numbers), thus error-bars grow smaller.

e)



f)



g) We can see the optimal value of k for the first case is 1 and in the second case (corrupted data) is between 5 and 9.

The difference between the two datasets is that in the first example we notice the error rises with k monotonically, and in the second example, at start the error improves and around the value $k = 7$ the error starts to rise monotonically, as in case 1.

Our explanation is that in the first case, the error rises as the value of k rises, because the sample size is small – thus k constitutes a high percentage of the number of examples from each digit (≈25). This may lead to a high percentage of examples with a label l being labeled by knn as l’ ($\neq l$), because they are close to the “border” with neighboring areas. In the second case, in the realm with lower k values, the prediction is too dependent on the corrupted data and therefore gains relatively high error. As k grows, the error gets averaged and therefore the error lowers, until it reaches some “saturation” in the sense of how much averaging can fix it, then starts growing monotonically as in the first case.

Question 3:

a) $\{x_1, x_2\} \in \mathcal{X} \subseteq \mathbb{R}^2$ is the set of vectors that describes students height and age where: height is $0 \leq x_1 \leq 2.5$ (meters) and age is $0 \leq x_2 \leq 120$ (years) (the set of all possible examples).
$\mathcal{Y} = \{drama, comedy\}$ is the group containing the labels 'drama' and 'comedy' (the set of all possible labels).

b)

Height (x_1)	Age (x_2)	$h_{bayes}(\{x_1, x_2\})$
160	20	Drama
160	40	comedy
180	25	Drama
180	35	comedy

c) We have seen in class that:

$$err_{bayes}(D) = err(h_{bayes}, D) = \mathbb{P}_{(X,Y)\sim D}\left[h_{bayes}(X) \neq Y\right] =$$

$$= \sum_{(x,y)\in X\times Y: h(X)\neq Y} \mathbb{P}[X = x, Y = y] \;=\; \dots \;=\; \sum_{x\in\mathcal{X}} \mathbb{P}[X = x](1 - \eta_{h(x)}(x))$$

Where:

$$\eta_{h(x)} = \mathbb{P}[Y = y|X = x]$$

Therefore:

$$err_{bayes}(D) = \mathbb{P}[X = \{160{,}20\}](1 - \mathbb{P}[Y = 'drama'|X = \{160{,}20\}]) +$$

$$\mathbb{P}[X = \{160{,}40\}](1 - \mathbb{P}[Y = 'comedy'|X = \{160{,}40\}]) +$$

$$\mathbb{P}[X = \{160{,}40\}](1 - \mathbb{P}[Y = 'drama'|X = \{160{,}40\}]) +$$

$$\mathbb{P}[X = \{180{,}25\}](1 - \mathbb{P}[Y = 'drama'|X = \{180{,}25\}]) +$$

$$\mathbb{P}[X = \{180{,}25\}](1 - \mathbb{P}[Y = '\ comedy'|X = \{180{,}25\}]) +$$

$$\mathbb{P}[X = \{180{,}35\}](1 - \mathbb{P}[Y = 'comedy'|X = \{180{,}35\}]) =$$

$$= \frac{13}{100} * (0) + \frac{30}{100} * \left(1 - \frac{30}{50}\right) + \frac{20}{100} * \left(1 - \frac{20}{50}\right) + \frac{17}{100} * \left(1 - \frac{17}{22}\right) + \frac{5}{100} * \left(1 - \frac{5}{22}\right) + \frac{35}{100} * (0)$$

$$= \frac{69.8}{220} \approx 0.32$$

d)

Height	genre	$probability$
160	drama	33%
160	comedy	30%
180	drama	17%
180	comedy	20%

e) We will show the bayes optimal predictor:

Height (x_1)	$h_{bayes}(\{x_1, x_2\})$
160	Drama
180	comedy

We have seen in class that:

$err_{bayes}(D) = \ \dots\ = \sum_{x \in \mathcal{X}} \mathbb{P}[X = x](1 - \eta_{h(x)}(x))$

Where:

$$\eta_{h(x)} = \mathbb{P}[Y = y | X = x]$$

Therefore:

$$err_{bayes}(D) = \mathbb{P}[X = 160](1 - \mathbb{P}[Y = 'drama' | X = 160]) +$$

$$\mathbb{P}[X = 160](1 - \mathbb{P}[Y = 'comedy' | X = 160]) +$$

$$\mathbb{P}[X = 180](1 - \mathbb{P}[Y = 'drama' | X = 180]) +$$

$$\mathbb{P}[X = 180](1 - \mathbb{P}[Y = 'comedy' | X = 180]) =$$

$$= \frac{63}{100} * (0) + \frac{63}{100} * \left(1 - \frac{33}{63}\right) + \frac{37}{100} * \left(1 - \frac{20}{37}\right) + \frac{37}{100} * (0)$$

$$= \frac{30}{100} + \frac{17}{100} = 0.47$$

We can see from the result that optimal bayes error has grown, this is since we narrowed the representation of the examples, thus omitting relevant information – this increases the relative proportion of the probability of all non-maximal-probability labels per each example, thus enlarging the Bayes optimal error – meaning $\mathbb{P}\left[Y \neq h_{bayes}(x) \middle| X = x\right]$ increases for each x, and we know that:

$$err_{bayes}(D) = \sum_{x \in X} \mathbb{P}[X = x]\mathbb{P}\left[Y \neq h_{bayes}(x) \middle| X = x\right]$$

So total error increases.

f) We saw in class that the expected error of the Memorize rule is:

$$\mathbb{E}_{s \sim G^m}[err(h_s^{mem}, G)] = \frac{k-1}{k}\sum_{x \in X} p_x(1 - p_x)^m$$

And therefor in our case is:

$$\mathbb{E}_{s \sim G^5}[err(h_s^{mem}, G)] = \frac{2-1}{2}\sum_{x \in X} p_x(1 - p_x)^5$$

$$= \frac{1}{2}(\mathbb{P}[x = \{160,20\}] * (1 - \mathbb{P}[x = \{160,20\}])^5 + \mathbb{P}[x = \{170,40\}]$$
$$* (1 - \mathbb{P}[x = \{170,40\}])^5 + \mathbb{P}[x = \{180,25\}] * (1 - \mathbb{P}[x = \{180,25\}])^5$$
$$+ \mathbb{P}[x = \{180,35\}] * (1 - \mathbb{P}[x = \{180,35\}])^5) =$$

$$= \frac{1}{2}\left(\frac{20}{100} * \left(1 - \frac{20}{100}\right)^5 + \frac{30}{100} * \left(1 - \frac{30}{100}\right)^5 + \frac{10}{100} * \left(1 - \frac{10}{100}\right)^5 + \frac{40}{100} * \left(1 - \frac{40}{100}\right)^5\right) =$$

$$= \frac{1}{2}\left(\frac{20}{100} * \left(\frac{80}{100}\right)^5 + \frac{30}{100} * \left(\frac{70}{100}\right)^5 + \frac{10}{100} * \left(\frac{90}{100}\right)^5 + \frac{40}{100} * \left(\frac{60}{100}\right)^5\right) \approx 0.103$$

We are allowed to use it because G has a deterministic label conditioned on the example.

Question 4:

a) Our objective is to show that for nn_1, the output from 1-nearest-neighbor algorithm on $\mathcal{X} \times \mathcal{Y}$ with distance $p = \Delta$,satisfies the conditions of an ERM algorithm for hypothesis class H_1. meaning:

$$nn_1 \in \underbrace{argmin}_{h \in H_1} (err(h, S))$$

First, we will show that $nn_1 \in H_1$ and then, that $err(nn_1, S) = 0$ and thus we will conclude:

$$nn_1 \in \underbrace{argmin}_{h \in H_1} (err(h, S))$$

On any given $S^m = \big((x_1, y_1), \dots, (x_m, y_m)\big)$ sample size: $nn_1(x_i) = y_i$

To show that $nn_1 \in H_1$ we will find an equivalent function $\hat{f}$ from H_1 that will receive the same output for any given $x \in \mathbb{R}$.

For any $1 \le i \le m - 1$ we will choose: $a_i = \frac{y_i + y_{i+1}}{2}$ and define function $\hat{f}$ as follows:

$$\hat{f}(x) = f_{m-1, a_1, \dots, a_{m-1}, y_1, \dots, y_m}(x) = f_{m-1, \frac{y_1+y_2}{2}, \dots, \frac{y_{m-1}+y_m}{2}, y_1, \dots, y_m}(x) =$$

$$= \begin{cases} y_i & , 1 \le i \le m-1 \ and\ i\ is\ the\ smallest\ index\ such\ that\ x \le a_i \\ y_m & , x > a_n \end{cases}$$

Claim: $\hat{f} = nn_1$

Proof:

Given $\tilde{x} \in \mathbb{R}$,

For some $i \in [1, m-1]$:

$x_i \le \tilde{x} \le x_{i+1}$ or $\tilde{x} < x_1$ or $x_m < \tilde{x}$

Case 1: $\Delta(\tilde{x}, x_i) \le \Delta(\tilde{x}, x_{i+1})$:

$$nn_1(\tilde{x}) = \left\{ y_j \middle| \min_{0 \le j \le m} \Delta(\tilde{x}, x_j) \text{ and } (x_j, y_j) \in S \right\} = y_i$$

And:

$$\hat{f}(x) = y_j\ s.t{:}\ 1 \le j \le m-1\ and\ j\ is\ the\ smallest\ index\ such\ that\ \tilde{x} \le \frac{y_j + y_{j+1}}{2} = y_i$$

Therefor $nn_1(\tilde{x}) = \hat{f}(\tilde{x})$.

Case 2: $\Delta(\tilde{x}, x_i) > \Delta(\tilde{x}, x_{i+1})$:

$$nn_1(\tilde{x}) = \left\{ y_j \middle| \min_{0 \le j \le m} \Delta(\tilde{x}, x_j) \text{ and } (x_j, y_j) \in S \right\} = y_{i+1}$$

And:

$$\hat{f}(x) = y_j \; s.t{:}\, 1 \le j \le m-1 \; and \; j \; is \; the \; smallest \; index \; such \; that \; \tilde{x} \le \frac{y_j + y_{j+1}}{2} = y_{i+1}$$

Therefor $nn_1(\tilde{x}) = \hat{f}(\tilde{x})$.

Case 3: $\tilde{x} < x_1$:

$$nn_1(\tilde{x}) = \left\{ y_j \middle| \min_{0 \le j \le m} \Delta(\tilde{x}, x_j) \text{ and } (x_j, y_j) \in S \right\} = y_1$$

And:

$$\hat{f}(x) = y_j \; s.t{:}\, 1 \le j \le m-1 \; and \; j \; is \; the \; smallest \; index \; such \; that \; \tilde{x} \le \frac{y_j + y_{j+1}}{2} = y_1$$

Therefor $nn_1(\tilde{x}) = \hat{f}(\tilde{x})$.

Case 4: $x_m < \tilde{x}$:

$$nn_1(\tilde{x}) = \left\{ y_j \middle| \min_{0 \le j \le m} \Delta(\tilde{x}, x_j) \text{ and } (x_j, y_j) \in S \right\} = y_m$$

And:

$$\hat{f}(x) = y_m$$

Therefor $nn_1(\tilde{x}) = \hat{f}(\tilde{x})$.

Claim: $err(nn_1, S) = 0$

Proof:

The error defined as:

$$err(h, S) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[h(x_i) \ne y_I]$$

From nn_1 algorithm, because $\Delta(x_i, x_i) = 0$, for any $i \in [1, m]$: $nn_1(x_i) = y_i$

Thus:

$$err(nn_1, S) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}[nn_1(x_i) \ne y_I] = \frac{1}{m} * 0 = 0$$

b) For Sample size $S^3 = ((x_1, 0), (x_2, 0), (x_3, 1))$

And for 3-nearest-neighbor algoritem nn_3 :

$$err(nn_3, S) = \frac{1}{3} + \mathbb{1}[nn_3(x_1) \neq 0] + \mathbb{1}[nn_3(x_2) \neq 0] + \mathbb{1}[nn_3(x_3) \neq 0]$$

$$= \frac{1}{3} + \mathbb{1}[0 \neq 0] + \mathbb{1}[0 \neq 0] + \mathbb{1}[1 \neq 0] = \frac{1}{3} + 0 + 0 + 1 = \frac{1}{3}$$

Since $err(nn_1, S) = 0 < \frac{1}{3} = err(nn_3, S)$, 3-nearest-neighbor algoritem nn_3 do not behave like an ERM algorithm for the hypothesis class H_1.

Question 5:

a) We've seen in class that if there are only two labels, $Y = \{0, 1\}$, we can set

$$h_{bayes}(x) = \mathbb{1}\left[\eta_1(x) \geq \frac{1}{2}\right] = \mathbb{1}\left[\alpha \geq \frac{1}{2}\right] = \begin{cases} 1, & \alpha \geq 0.5 \\ 0, & \alpha < 0.5 \end{cases}$$

$$err\left(h_{bayes}, D\right) = \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[h_{bayes}(\mathrm{x}) \neq \mathrm{y_x}\right] =$$

$$= \int_0^1 \mathbb{1}\left[h_{bayes}(x) \neq y_x \right] dx = \int_0^1 \mathbb{1}\left[\mathbb{1}\left[\alpha \geq \frac{1}{2}\right] \neq y_x \right] dx =$$

$$= \begin{cases} \int_0^1 \mathbb{1}[1 \neq y_x]\, dx, & \alpha \geq 0.5 \\ \int_0^1 \mathbb{1}[0 \neq y_x]\, dx, & \alpha < 0.5 \end{cases} = \begin{cases} \int_0^1 \mathbb{1}[y_x = 0]\, dx, & \alpha \geq 0.5 \\ \int_0^1 \mathbb{1}[y_x = 1]\, dx, & \alpha < 0.5 \end{cases} =$$

$$= \begin{cases} \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[\mathrm{y_x} = 0 | X = x], & \alpha \geq 0.5 \\ \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[\mathrm{y_x} = 1 | X = x], & \alpha < 0.5 \end{cases} = \begin{cases} 1 - \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[\mathrm{y_x} = 1 | X = x], & \alpha \geq 0.5 \\ \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[\mathrm{y_x} = 1 | X = x], & \alpha < 0.5 \end{cases}$$

Thus, the Bayes-optimal predictor error of D as a function of α is:

$$\begin{cases} 1 - \alpha \ , & \alpha \geq 0.5 \\ \alpha, & \alpha < 0.5 \end{cases}$$

b) We know that we have a homogenous distribution s.t $\mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[Y = 1 | X = x] = \alpha$.

Let us denote the nearest neighbor of some x as $nn(x) \in \{x_1, \dots, x_m\}$, and we denote its tag as $Y_{nn(x)}$. Thus, we also know that $\mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[Y_{nn(x)} = 1 | X = x\right] = \alpha$.

So, we see that $\mathbb{P}[h(X) \neq Y]$ is independent of the identity of $x, x_1, \dots, x_m$.

Finally, we can say that:

$$f(\alpha) \equiv \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[h(X) \neq Y \mid X = x \ and\ examples\ in\ S^m\ are\ x_1, \dots, x_m] =$$

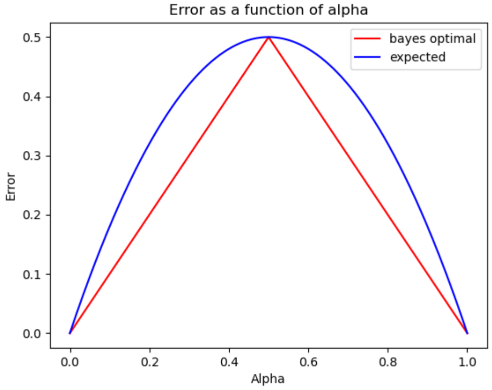
$$\mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[h(X) \neq Y] = \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[h(X) = 0 \ \wedge \ Y = 1] + \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[h(X) = 1 \ \wedge \ Y = 0] =$$

$$\mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[Y_{nn(x)} = 0 \ \wedge \ Y = 1\right] + \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[Y_{nn(x)} = 1 \ \wedge \ Y = 0\right] =$$

$$\mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[Y_{nn(x)} = 0\right] \cdot \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[Y = 1] + \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}\left[Y_{nn(x)} = 1\right] \cdot \mathbb{P}_{(\mathrm{X,Y}) \sim \mathrm{D}}[Y = 0] =$$

$$2\alpha(1 - \alpha)$$

c)



(The way we plot this graph can be seen in the attached python 'Q5.py' file)

d) Let there be $\alpha \in [0,1]$.

Claim 1:

$$\begin{cases} 1-\alpha & ,\alpha \geq 0.5 \\ \alpha & ,\alpha < 0.5 \end{cases} = err\left(h_{bayes}, D\right) \leq \mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right] = 2\alpha(1-\alpha)$$

Proof:

If $\alpha \geq 0.5$ then:

$$err\left(h_{bayes}, D\right) = 1-\alpha$$

And then:

$$\alpha \geq \frac{1}{2} \rightarrow\ 1 \leq 2\alpha \rightarrow\ \frac{1-\alpha}{1-\alpha} \leq 2\alpha \rightarrow\ \ 1-\alpha \leq 2\alpha(1-\alpha) \rightarrow$$
$$err\left(h_{bayes}, D\right) \leq \mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right]$$

If $\alpha < 0.5$ then:

$$err\left(h_{bayes}, D\right) = \alpha$$

And then:

$$\alpha < \frac{1}{2} \rightarrow 1-\alpha > 1-\frac{1}{2} \rightarrow 1-\alpha > \frac{1}{2} \rightarrow 2\alpha(1-\alpha) > 2\alpha * \frac{1}{2} \rightarrow\ 2\alpha(1-\alpha) > \alpha \rightarrow$$
$$err\left(h_{bayes}, D\right) \leq \mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right]$$

Claim 2:

$$\mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right] = 2\alpha(1-\alpha) \leq \begin{cases} 2*(1-\alpha) & ,\alpha \geq 0.5 \\ 2*\alpha & ,\alpha < 0.5 \end{cases} = 2 * err\left(h_{bayes}, D\right)$$

Proof:

If $\alpha \geq 0.5$ then:

$$2 * err\left(h_{bayes}, D\right) = 2*(1-\alpha)$$

And then:

$$\alpha \leq 1 \rightarrow \alpha * 2(1-\alpha) \leq 1 * 2(1-\alpha) \rightarrow 2\alpha(1-\alpha) \leq 2(1-\alpha) \rightarrow$$
$$\mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right] \leq 2 * err\left(h_{bayes}, D\right)$$

If $\alpha < 0.5$ then:

$$2 * err\left(h_{bayes}, D\right) = 2 * \alpha$$

And then:

$$\alpha \geq 0 \rightarrow 1-\alpha \leq 1 \rightarrow (1-\alpha) * 2\alpha \leq 2\alpha \rightarrow 2\alpha(1-\alpha) \leq 2\alpha \rightarrow$$
$$\mathbb{E}_{S\sim D^m}\left[err\left(\hat{h}_S, D\right)\right] \leq 2 * err\left(h_{bayes}, D\right)$$

Thus, from claim 1 and clam 2 we conclude that the expected error is always between the Bayes-optimal error and twice the Bayes-optimal error.

We will find the values of α where the expected error equal to the Bayes-optimal error:

$$\begin{cases} 1-\alpha & ,\alpha \geq 0.5 \\ \alpha & ,\alpha < 0.5 \end{cases} = 2\alpha(1-\alpha)$$

If $\alpha \geq 0.5$ then:

$$1-\alpha = 2\alpha(1-\alpha)$$
$$2\alpha - 2\alpha^2 + \alpha - 1 = 0$$
$$2\alpha^2 - 3\alpha + 1 = 0$$
$$\alpha_1 = 0.5, \alpha_2 = 1$$

If $\alpha \leq 0.5$ then:

$$\alpha = 2\alpha(1-\alpha)$$
$$2\alpha - 2\alpha^2 - \alpha = 0$$
$$2\alpha^2 - \alpha = 0$$
$$\alpha_1 = 0, \alpha_2 = 0.5$$

Thus, when $\alpha = 0, 0.5, 1$ the expected error equal to the Bayes-optimal.

We will find the values of α where the expected error equal to twice the Bayes-optimal error:

$$\begin{cases} 2 * (1-\alpha) & ,\alpha \geq 0.5 \\ 2 * \alpha & ,\alpha < 0.5 \end{cases} = 2\alpha(1-\alpha)$$

If $\alpha \geq 0.5$ then:

$$2 * (1-\alpha) = 2\alpha(1-\alpha)$$
$$2\alpha - 2\alpha^2 + 2\alpha - 2 = 0$$
$$2\alpha^2 - 4\alpha + 2 = 0$$
$$\alpha = 1$$

If $\alpha < 0.5$ then:

$$2\alpha = 2\alpha(1-\alpha)$$
$$2\alpha - 2\alpha^2 - 2\alpha = 0$$
$$2\alpha^2 = 0$$
$$\alpha = 0$$

Thus, when $\alpha = 0, 1$ the expected error equal to twice the Bayes-optimal.

Question 6:

We define:

$$\mathcal{X}_n = \{G = (E, V) \mid G\ is\ an\ undirected\ graph, |V| = n \in \mathbb{N}, \forall v \in V \rightarrow \deg(v) \leq 5\}$$

$$\mathcal{Y} = \{0,1\}$$

And for $x \in \mathcal{X}_n$, we define:

$$g(x) = \{\deg(v_1), \dots, \deg(v_n)\}$$

And we define our hypotheses class:

$$\mathcal{H} = (h_v(x) \mid v \in \mathbb{N}^n , h_v(x) = \mathbb{1}[g(x) = v])$$

a) If $g(x)$ is deterministic, so is $h_v(x)$ for any v. **But** we can observe a subtle point – the labeling in $\mathcal{D}$ is done via a deterministic function of $g(x)$, let's call it $f\big(g(x)\big)$, but this does not mean that we cannot have $g(x_1) \neq g(x_2)$ but $f\big(g(x_1)\big) = f\big(g(x_2)\big)$.

For example, if we define $f_a\big(g(x)\big)$ to be the function (for some $a \in \mathbb{R}$):

$$f_a\big(g(x)\big) = \begin{cases} 1 \ , & \big||g(x)|\big| \geq a \\ 0 \ , & \big||g(x)|\big| \leq a \end{cases}$$

Then for $g(x_1) \neq g(x_2)$ we could have $f\big(g(x_1)\big) = f\big(g(x_2)\big)$.

For any chosen v', $\exists x' \in \mathcal{X}_n \rightarrow g(x') = v'$ and we can choose x'' s.t $g(x'') \neq g(x')$, but $f\big(g(x')\big) = f\big(g(x'')\big)$.

Thus:

$$h_{v'}(x') = \mathbb{I}[g(x') = v'] = 1$$
$$h_{v'}(x'') = \mathbb{I}[g(x'') = v'] = 0$$

But we have $f\big(g(x')\big) = f\big(g(x'')\big)$, so $\mathcal{D}$ is not realizable by $\mathcal{H}$.

Thus, this is the **agnostic** setting.

b) First, we determine the size of $\mathcal{H}$:

The size of $\mathcal{H}$ is exactly the number of possible $v \in \mathbb{N}^n = 5^n$, as all other parameters are constant.

This means $|\mathcal{H}| = 5^n$.

So, the PAC bound for the agnostic case is, for given ϵ, δ:

$$m(n) \geq \frac{2\log(|\mathcal{H}(n)|) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = \frac{2\log(5^n) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = n\left(\frac{2}{\epsilon^2}\log(5)\right) + \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$$

$$m(n) \geq n\left(\frac{2}{\epsilon^2}\log(5)\right) + \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right) = O(n)$$

c) $\underline{VC(\mathcal{H}) < 2:}$

For $x_1, x_2 \in \mathcal{X}_n$ we have 4 different ways to label:

$$(y_1, y_2) = (0,0), (0,1), (1,0), (1,1)$$

Case 1 – $\boldsymbol{g(x_1) \neq g(x_2)}$:

In this case, no v we choose can label x_1, x_2 as $(1,1)$, as $h_v(x) = \mathbb{I}[g(x) = v]$, thus if $w.l.o.g$ $g(x_1) = v$, than $g(x_2) \neq v$ and we have:

$$h_v(x_1) = \mathbb{I}[g(x_1) = v] = 1$$
$$h_v(x_2) = \mathbb{I}[g(x_2) = v] = 0$$

Case 2 – $\boldsymbol{g(x_1) = g(x_2)}$:

In the same fashion we have that there is no way to label x_1, x_2 differently ($(1,0)$ or $(0,1)$).

Thus, there is no set of size 2 that $\mathcal{H}$ shatters.

$\underline{VC(\mathcal{H}) \geq 1:}$

For a given sample x_0 , we can shatter this set (of size 1) with $v_1 = g(x_0)$ and $v_2 \neq g(x_0)$, which label:

$$h_{v_1}(x_0) = \mathbb{I}[g(x_0) = v_1] = 1$$
$$h_{v_2}(x_0) = \mathbb{I}[g(x_0) = v_2] = 0$$

Such v_1, v_2 exist of course, and this is trivial to show.

So, we see that:

$$VC(\mathcal{H}) = 1$$

<u>Pac bound:</u>

$$m \geq \frac{2VC(\mathcal{H}) - 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = \frac{2 - 2\log(2) + 2\log(\delta)}{\epsilon^2} = \frac{2\log(\delta)}{\epsilon^2}$$

We remember that:

$$m \geq n\left(\frac{2}{\epsilon^2}\log(5)\right) + \frac{2}{\epsilon^2}\log\left(\frac{2}{\delta}\right)$$

$$\log\left(\frac{2}{\delta}\right) \leq \frac{\epsilon^2 m}{2} - n\log(5)$$

$$1 + n\log(5) - \frac{\epsilon^2 m}{2} \leq \log(\delta)$$

$$\Rightarrow m \geq \frac{2\log(\delta)}{\epsilon^2} \geq \frac{2}{\epsilon^2}\left(1 + n\log(5) - \frac{\epsilon^2 m}{2}\right)$$

$$\Rightarrow 2m \geq \frac{2}{\epsilon^2} + \frac{2}{\epsilon^2} n\log(5)$$

$$\Rightarrow m \geq \frac{1 + n\log(5)}{\epsilon^2}$$