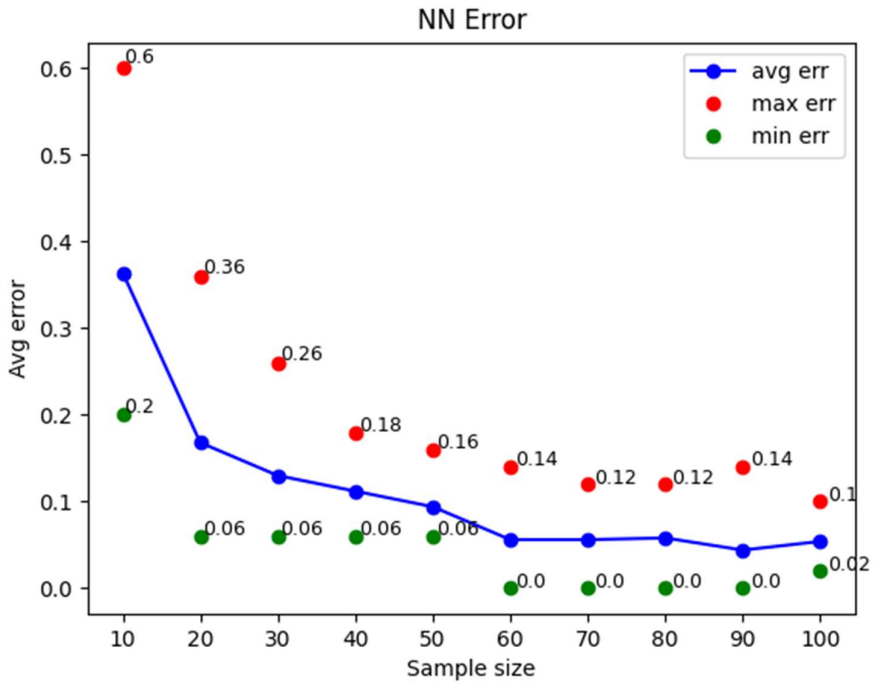
Assignment 1

Question 2:

a) The following plot shows the results for NN algorithm where the y axis represents the average error and the x axis represents the sample size. Red/green dots are the maximal/minimal error we got on each sample:



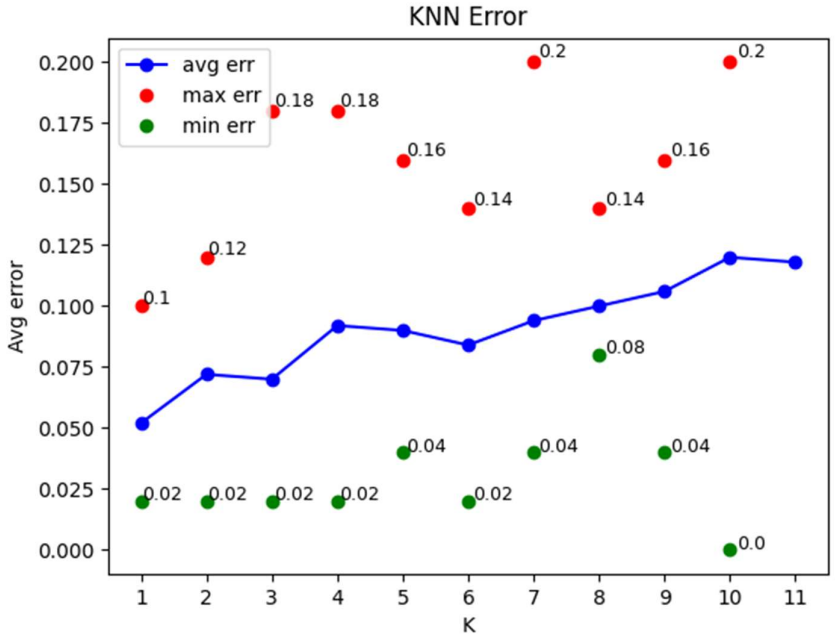
b) We can see from the plot that as the sample size gets larger, the average error decreases. Since the distribution has a deterministic labels conditioned on the example, as the sample size grows the probability that each image has a close neighbor grows. A close neighbor in this case suggests that the digits were written similar which means they have the same label, thus larger sample size reduces the average error.

c) We got different results for different runs with the same sample size. The reason is because the samples were chosen randomly from the data set, where each sample may yield a different classifier. If for instance a sample was chosen such that a lot of labels are '1' and only a few labels are '4', the probability that a test image with label '4' will have a close neighbor with label '1' grows, which in turn raises the error average. Thus different samples from the same size may yield different results.

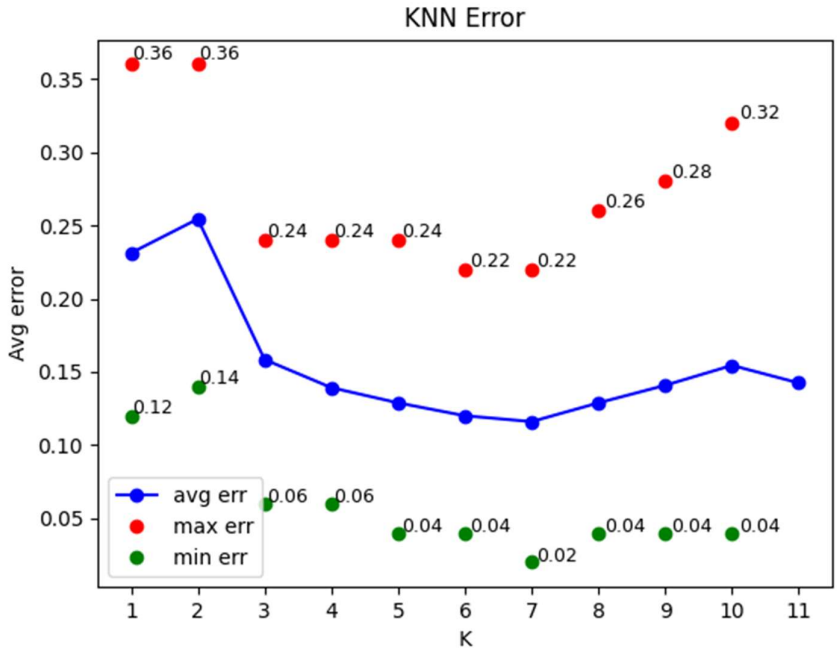
d) The general trend of the maximal and minimal error was to decrease as the sample size gets larger. The reason is same as we explained in section 'b' where we explained why the average decreasing for larger sample size. Notice that in some sample sizes the minimal error is low while the maximal error is significantly higher. It can be explained due to the fact that different runs with the same sample size may yield different results. For each sample size, it might be that the samples chosen for the run were not well distributed, meaning that for some of the images had a relative far

nearest neighbor which in turn leads to less accurate results from the classifier. The same for a well distributed sample that for its run gave much accurate results.

e) The plot bellow shows the error of KNN classifier for an increasing K with constant sample size:



f) The plot bellow shows the error of corrupted KNN classifier for an increasing K with constant sample size:



Assignment 1

g) Comparing the two graphs we got for the experiments, we can see that for the non-corrupted train the average error is minimal when k=1, and for the corrupted train when k=7. One of the differences between the graphs is the trending of the average error function. In the non-corrupted case, the error was increasing as K increased while in the corrupted case the error decreased at the beginning and slowly increased again for k>7. We can see that k=1 corresponds for high error for the corrupted case while it is optimal in the non-corrupted case. The bad result on k=1 makes sense because 20% of the training labels are corrupted, meaning that some of the digits in the test that are neighbors to the corrupted can be identified with a label of a digit that looks completely different. The optimum at k=7 may suggest that digits that were identified wrong when k=1 has enough close neighbors with a non-corrupted label which makes the majority of the neighbors make a correct prediction. Still notice that the optimum in the non-corrupted case gives better results from the optimum in the corrupted case.

Question 3:

a) $\mathcal{X} = \{(n,m) \in \mathbb{N}^2 | n \leq 250, m \leq 120\}$

$\mathcal{Y} = \{drama, comedy\}$

b) We will find the Bayes-optimal predictor as follows:

Without loss of generality define: $drama = 0, comedy = 1$. Define η_i as follows: $\eta_i(x) = \mathbb{P}_{(X,Y)\sim D}[Y = i | X = x] = \frac{\mathbb{P}[Y=i \wedge X=x]}{\mathbb{P}[X=x]}$. We proofed in class that the best prediction rule must satisfy: $h_{bayes}(x) \in argmax_{y\in\mathcal{Y}}\, \eta_y(x)$. Since $\mathcal{Y}$ contains only two labels we can define $h_{bayes}(x) = \mathbb{I}[\eta_1(x) > \frac{1}{2}]$. Calculating the values of the predictor gives:

$$h_{bayes}(160,20) = \mathbb{I}\left[\mathbb{P}_{(X,Y)\sim D}[Y = 1 | X = (160,20)] \geq \frac{1}{2}\right] =$$
$$= \mathbb{I}\left[1 - \mathbb{P}_{(X,Y)\sim D}[Y = 0 | X = (160,20)] \geq \frac{1}{2}\right] = \mathbb{I}\left[1 - \frac{0.13}{0.13} \geq \frac{1}{2}\right] = 0 = drama$$

$$h_{bayes}(160,40) = \mathbb{I}\left[\eta_1(160,40) \geq \frac{1}{2}\right] = \mathbb{I}\left[\frac{0.3}{0.5} \geq \frac{1}{2}\right] = 1 = comedy$$

$$h_{bayes}(180,25) = \mathbb{I}\left[\eta_1(180,25) \geq \frac{1}{2}\right] = \mathbb{I}\left[\frac{0.05}{0.22} \geq \frac{1}{2}\right] = 0 = drama$$

$$h_{bayes}(180,35) = \mathbb{I}\left[\eta_1(180,35) \geq \frac{1}{2}\right] = \mathbb{I}\left[\frac{0.15}{0.15} \geq \frac{1}{2}\right] = 1 = comedy$$

c) Since h_{bayes} we defined in the previous section is optimal, using the calculations we saw in class, we can calculate the Bayes-optimal error as follows:

$$\mathrm{err}\left(\mathrm{h_{bayes}}, \mathcal{D}\right) = \mathbb{P}_{(\mathrm{X,Y})\sim \mathrm{D}}\left[\mathrm{h_{bayes}(X)} \neq \mathrm{Y}\right] = \sum_{x\in \mathrm{X}} \mathbb{P}[X = x]\left(1 - \eta_{h_{bayes}(x)}(x)\right) =$$
$$\mathbb{P}[X = (160,20)]\left(1 - \eta_0(160,20)\right) + \mathbb{P}[X = (160,40)]\left(1 - \eta_1(160,40)\right)$$
$$+ \mathbb{P}[X = (180,25)]\left(1 - \eta_0(180,25)\right) + \mathbb{P}[X = (180,35)]\left(1 - \eta_1(180,35)\right) =$$
$$= 0.13\left(1 - \frac{0.13}{0.13}\right) + 0.5\left(1 - \frac{0.3}{0.5}\right) + 0.22\left(1 - \frac{0.17}{0.22}\right) + 0.15\left(1 - \frac{0.15}{0.15}\right) = 0.25$$

Assignment 1

d) Frist calculate:

$\mathbb{P}[X_2 = 160, Y = 0] = \mathbb{P}[(X = (160,20), Y = 0) \uplus (X = (160,40), Y = 0)] = \mathbb{P}[(X = (160,20), Y = 0)] +$
$+\mathbb{P}[(X = (160,40), Y = 0)] = 0.33$

$\mathbb{P}[X_2 = 160, Y = 1] = \mathbb{P}[(X = (160,20), Y = 1) \uplus (X = (160,40), Y = 1)]$
$= \mathbb{P}[(X = (160,20), Y = 1)] + \mathbb{P}[(X = (160,40), Y = 1)] = 0.30$

$\mathbb{P}[X_2 = 180, Y = 0] = \mathbb{P}[(X = (180,25), Y = 0) \uplus (X = (180,35), Y = 0)]$
$= \mathbb{P}[(X = (180,25), Y = 0)] + \mathbb{P}[(X = (180,35), Y = 0)] = 0.17$

$\mathbb{P}[X_2 = 180, Y = 1] = \mathbb{P}[(X = (180,25), Y = 1) \uplus (X = (180,35), Y = 1)]$
$= \mathbb{P}[(X = (180,25), Y = 1)] + \mathbb{P}[(X = (180,35), Y = 1)] = 0.20$

So the table of probabilities is:

height	genre	probability
160	drama	0.33
160	comedy	0.30
180	drama	0.17
180	comedy	0.20

e) First we define a Bayes-optimal predictor. Like before: $\eta_i(x) = \mathbb{P}_{(X_2,Y)\sim D}[Y = i | X_2 = x] = \frac{\mathbb{P}[Y=i \wedge X_2=x]}{\mathbb{P}[X_2=x]}$

and $h_{bayes}(x) = \mathbb{I}[\eta_1(\mathrm{x}) > \frac{1}{2}]$ and we calculate:

$$h_{bayes}(160) = \mathbb{I}\left[\eta_1(160) > \frac{1}{2}\right] = \mathbb{I}\left[\frac{0.3}{0.63} > \frac{1}{2}\right] = 0 = drama$$

$$h_{bayes}(180) = \mathbb{I}\left[\eta_1(180) > \frac{1}{2}\right] = \mathbb{I}\left[\frac{0.2}{0.37} > \frac{1}{2}\right] = 1 = comedy$$

Now we can calculate the Bayes optimal error similar to section c:

$$\mathrm{err}\left(\mathrm{h_{bayes}}, \mathcal{D}_2\right) = \mathbb{P}_{(\mathrm{X_2,Y})\sim \mathrm{D_2}}\left[\mathrm{h_{bayes}(X_2)} \neq \mathrm{Y}\right] = \sum_{x \in \mathrm{X_2}} \mathbb{P}[X_2 = x]\left(1 - \eta_{h_{bayes}(x)}(x)\right) =$$

$$\mathbb{P}[X_2 = 160]\left(1 - \eta_0(160)\right) + \mathbb{P}[X_2 = 180]\left(1 - \eta_1(180)\right) = 0.63\left(1 - \frac{0.33}{0.63}\right) + 0.37\left(1 - \frac{0.2}{0.37}\right) = 0.47$$

f) Since $\mathcal{D}$ has a deterministic label conditioned on the example we can use the formula we have seen in class for the expected missing mass: $\mathbb{E}_{S\sim D^5}[M_S] = \sum_{x \in X} \mathbb{P}[X = x](1 - \mathbb{P}[X = x])^5 =$
$= \mathbb{P}[X = (160,20)](1 - \mathbb{P}[X = (160,20)])^5 + \mathbb{P}[X = (170,40)](1 - \mathbb{P}[X = (170,40)])^5$
$+ \mathbb{P}[X = (180,25)](1 - \mathbb{P}[X = (180,25)])^5 + \mathbb{P}[X = (180,35)](1 - \mathbb{P}[X = (180,35)])^5$
$= 0.2 * 0.8^5 + 0.3 * 0.7^5 + 0.1 * 0.9^5 + 0.4 * 0.6^5 = 0.20611$

Also we saw in class that $err(f_S^{mem}, \mathcal{D}) = \frac{M_S(k-1)}{k}$ for $k = |\mathcal{Y}|$. So the expected error of the memorizer algorithm is:

$$\mathbb{E}[err(f_S^{mem}, \mathcal{D})] = \mathbb{E}\left[\frac{M_S(k-1)}{k}\right] = \frac{1}{2}\mathbb{E}[M_S] = 0.103055$$

Assignment 1

Question 4:

a) Denote the hypothesis of NN algorithm with distance function Δ for $S = ((x_1, y_1), \dots, (x_n, y_n))$ by h_S^{nn}, and since S is non-conflicting w.l.g assume $x_1 < x_2 < \cdots < x_n$ (if $x_i = x_j$ then h_S^{nn} stays that same and it does not change that proof, so we can assume the claim above).Define the function $f := f_{n-1,x'_1,x'_2,\dots,x'_{n-1},y_1,y_2,\dots,y_n} \in \mathcal{H}_1$where $\forall i\ x'_i = \frac{x_i + x_{i+1}}{2}$, note that $x'_1 < x'_2 \dots < x'_{n-1}$ so f is well-defined and indeed a function from $\mathcal{H}_1$. If we will proof that $f \equiv h_S^{nn}$ then we can derive that NN algorithm with distance function Δ is an ERM algorithm for $\mathcal{H}_1$ since $err(f, S) = err(h_S^{nn}, S) = 0$.

Let $x \in \mathcal{X}$. If $x \geq x'_{n-1}$ then $f(x) = y_n$ and $\min_{x_1,\dots,x_n} |x - x_i| = |x - x_n|$ so $h_S^{nn}(x) = y_n$.
Else, denote note k to be the smallest index such that $x \leq x'_k$, so $f(x) = y_k$. But by the definition of x'_k we know that $\min_{x_1,\dots,x_n} |x - x_i| = |x - x_k|$ because x_k is closest to x, then $h_S^{nn}(x) = y_k$. Hence, $f \equiv h_S^{nn}$ which is exactly what we wanted to show.

b) Assume we run a KNN algorithm for $k > 1$ and we got the hypothesis h_S^{knn} for a non-conflicting sample $S = \{(x_1, y_1), (x_2, y_2), (x_3, y_2), \dots, (x_m, y_2)\}$, here the labels of $x_2, \dots, x_m$ is y_2 and $y_1 \neq y_2$. So since $k > 1$ we have that $h_S^{knn}(x_1) = y_2$ hence $err\left(h_S^{knn}, S\right) > 0$. But in section a' we proofed that there exists $h \in \mathcal{H}_1$ such that $err(h, S) = 0$ which means that KNN algorithm did not return a hypothesis with minimal error on the sample from the class $\mathcal{H}_1$ thus it is not an ERM algorithm for the class.

Question 5:

a) Let h_{bayes} be the Bayes-optimal predictor. We proofed in class that such a function must satisfy: $h_{bayes}(x) \in argmax_{y \in \mathcal{Y}} \eta_y(x)$. Since $\mathcal{Y}$ has only two labels we can set:
$h_{bayes}(x) = \mathbb{I}\left[\eta(x) \geq \frac{1}{2}\right] = \mathbb{I}\left[\alpha \geq \frac{1}{2}\right]$, hence $h_{bayes} \equiv 1\ if\ \alpha \geq \frac{1}{2}\ and\ h_{bayes} \equiv 0\ else$. So by definition of Bayes-optimal rule, the Bayes-optimal error is:

$$err\left(h_{bayes}, \mathcal{D}\right) = \mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[h(x) \neq y] = \begin{cases} \mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 0] & if\ \alpha \geq \frac{1}{2} \\ \mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 1] & if\ \alpha < \frac{1}{2} \end{cases}$$

From the law of total probability we know:

$$\mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 0] = \int_0^1 \mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 0 | X = x] f_X(x) dx = \int_0^1 (1 - \alpha) f_X(x) dx = 1 - \alpha$$

$$\mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 1] = \int_0^1 \mathbb{P}_{\mathcal{X} \times y \sim \mathcal{D}}[y = 1 | X = x] f_X(x) dx = \int_0^1 \alpha f_X(x) dx = \alpha$$

So we can conclude

$$err\left(h_{bayes}, \mathcal{D}\right) = \begin{cases} 1 - \alpha & if\ \alpha \geq \frac{1}{2} \\ \alpha & if\ \alpha < \frac{1}{2} \end{cases} = \min(\alpha, 1 - \alpha)$$

b) Let x_k be the nearest neighbor of x where x_k is an example in S. Notice that $f(\alpha)$ is the probability that the label of x is different for the label of x_k given S and x, and since $\mathcal{D}$ is uniform over $\mathcal{X}$ the probability that $(x, y) \in S$ is zero because the probability that a finite set of examples

drawn independently according to $\mathcal{D}$ has repeating values of x is zero. So we can write the following:
$f(\alpha) = \mathbb{P}[h(X) \neq Y|x,S] = \mathbb{P}[(x,0),(x_k,1)|x,S] + \mathbb{P}[(x,1),(x_k,0)|x,S]$

From here we use the fact that x and x_k were chosen independently and again with probability 1 they are different:

$f(\alpha) = \mathbb{P}[(x_k,1)|x,S]\mathbb{P}[(x,0)|x,S] + \mathbb{P}[(x,1)|x,S]\mathbb{P}[(x_k,0)|x,S]$

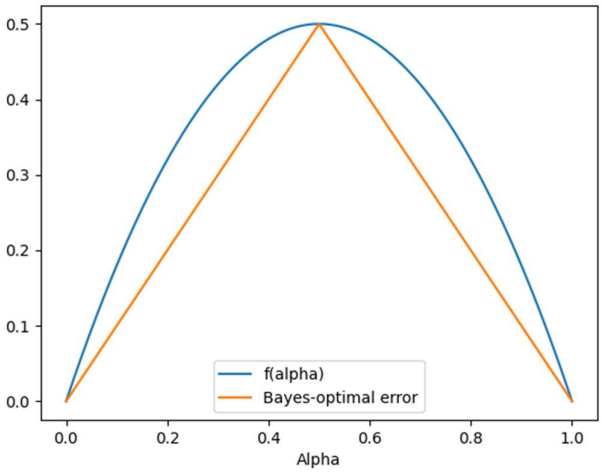
Now we can use the fact that x is not dependent on S and that x_k is not dependent on x to derive:

$f(\alpha) = \mathbb{P}[(x_k,1)|S]\mathbb{P}[(x,0)|x] + \mathbb{P}[(x,1)|x]\mathbb{P}[(x_k,0)|S]$

Since $S \sim \mathcal{D}^m$ we can conclude:

$$f(\alpha) = \mathbb{P}[(x_k,1)|x_k]\mathbb{P}[(x,0)|x] + \mathbb{P}[(x,1)|x]\mathbb{P}[(x_k,0)|x_k]$$
$$= \mathbb{P}[Y=1|x_k](1-\mathbb{P}[(Y=1)|x]) + \mathbb{P}[Y=1|x](1-\mathbb{P}[Y=1|x_k]) = 2\alpha(1-\alpha)$$

c) Below there is a plot of $f(\alpha)$ and Bayes-optimal error as a function of α:



d) Since $\alpha \in [0,1]$ we know $(1-\alpha) \in [0,1]$ so $2\alpha(1-\alpha) \leq 2\alpha$ and o $2\alpha(1-\alpha) \leq 2(1-\alpha)$hence $f(\alpha)$ is always smaller from twice the Bayes-optimal error.

If $0 \leq \alpha \leq \frac{1}{2}$ then $2(1-\alpha) \geq 1$ and from here: $2\alpha(1-\alpha) \geq a$ Bayes-optimal error for this case.

Else, $1 \geq \alpha \geq \frac{1}{2}$ then $2a \geq 1$ and from here $2\alpha(1-\alpha) \geq 1-\alpha$ which is the Bayes-optimal error for this case.

For $\alpha = 1$ and $\alpha = 0$ we have $f(\alpha) = 0$ and so as the Bayes-optimal error. Also for $\alpha = \frac{1}{2}$ we have $f(\alpha) = \frac{1}{2}$ and so as the Bayes-optimal error.

For $0 < a < \frac{1}{2}$ and $\frac{1}{2} < \alpha < 1$ the inequalities for the Bayes-optimal as lower bound are strict, meaning that for such α we have that $f(\alpha)$ is strictly larger than the Bayes-optimal error.

Notice that for $\alpha = 0$, $f(\alpha) = 0$ and also $f'(\alpha) = -4\alpha + 2$, so $f'(0) = 2$ meaning that 2α is a linear approximation of $f(\alpha)$ at the point 0, which is exactly twice the Bayes-optimal error, so as α

approaches to zero it approaches twice the Bayes-optimal error. Similarly, for $\alpha = 1$, $f(1) = 0$, $f'(1) = -2$ which means that $-2\alpha + 2$ is a linear approximation of $f(\alpha)$ at the point 1, which is exactly twice the Bayes-optimal error, so as α approaches to 1 it approaches twice the Bayes-optimal error. Since for all $0 < \alpha < 1$, $f(\alpha)$ is strictly less than twice the Bayes-optimal error, and $f(\alpha)$ is both continues and curving outwards we know that there is no other value α which approaches twice the Bayes-optimal error.

Question 6:

a) Given the information we have, there is no guarantee that there exists $h_v \in \mathcal{H}$ such that $err(h_v, D) = 0$. For instance, $x_1, x_2 \in \mathcal{X}$ such that $g(x_1) \neq g(x_2)$ and the deterministic function for the labeling defined as $f(g(x)) = \begin{cases} 1 & if\ g(x) = g(x_1) or\ g(x) = g(x_2) \\ 0 & otherwise \end{cases}$ then for all $h_v \in \mathcal{H}$ we get that h_v has a wrong prediction for at least one of x_1, x_2.

So given the information the problem has an agnostic setting.

b) All graphs edges have a degree at most 5 so we can deduce:
$\{v \in \mathbb{N}^n | h_v \not\equiv 0\} \subset \{(v_1, .., v_n) | \forall i,\ v_i \in [5]\}$ and since right hand side has 6^n elements, we can bound the hypothesis class by $|\mathcal{H}| \leq 6^n$. So $\log(|\mathcal{H}|) \leq nlog6$. Now we will use the PAC-learning upper bound for the sample complexity seen in class to the agnostic setting: $m \geq \frac{2\,1 \quad (|\mathcal{H}|) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2}$. Since $\log(|\mathcal{H}|)$ is $O(n)$ we conclude that the sample complexity is linear in n.

c) We claim that the VC dimension of $\mathcal{H}$ is 1 if $n > 1$ and if $n = 1$ then $VC(\mathcal{H}) = 0$.

Let $x_1, x_2 \in \mathcal{X}$ be two different graphs. If $g(x_1) = g(x_2)$ then for all $h_v \in \mathcal{H}$ we have $h_v(x_1) = h_v(x_2)$ so labeling x_1 with 1 and x_2 with 0 implies that no h_v can give a right prediction. For the second case where $g(x_1) \neq g(x_2)$, there is no $h_v \in \mathcal{H}$ such that $h_v(x_1) = h_v(x_2) = 1$. So labeling both x_1 and x_2 with 1 yields that no h_v can give a right prediction. Hence no two different elements from $\mathcal{X}$ can be labeled in all possible combination which in turn implies $VC(\mathcal{H}) \leq 1$. Also, choose any $x \in \mathcal{X}$, for $v = g(x)$ if we label x with 1 then h_v predicts x correctly and if we label x with 0, then h_w predicts x correctly for $w \neq v$(under the assumption that $n \neq 1$). Hence $VC(\mathcal{H}) \geq 1$, so $VC(\mathcal{H}) = 1$.

Now we can use $VC(\mathcal{H})$ to get a better upper bound to the sample complexity. Recall that $m(\epsilon, \delta) = \ominus\left(\frac{VC(\mathcal{H}) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$ for the agnostic setting. So in our case: $m(\epsilon, \delta) = \ominus\left(\frac{1 + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$ which is a better upper bound we showed in section b.

Notice that for the case where $n = 1$ there is only one element in $\mathcal{X}$ so $|\mathcal{H}| = 1$ and in that case $h(x) = 1$, so labeling x with 0 implies that no h predicts x right thus $VC(\mathcal{H}) = 0$.