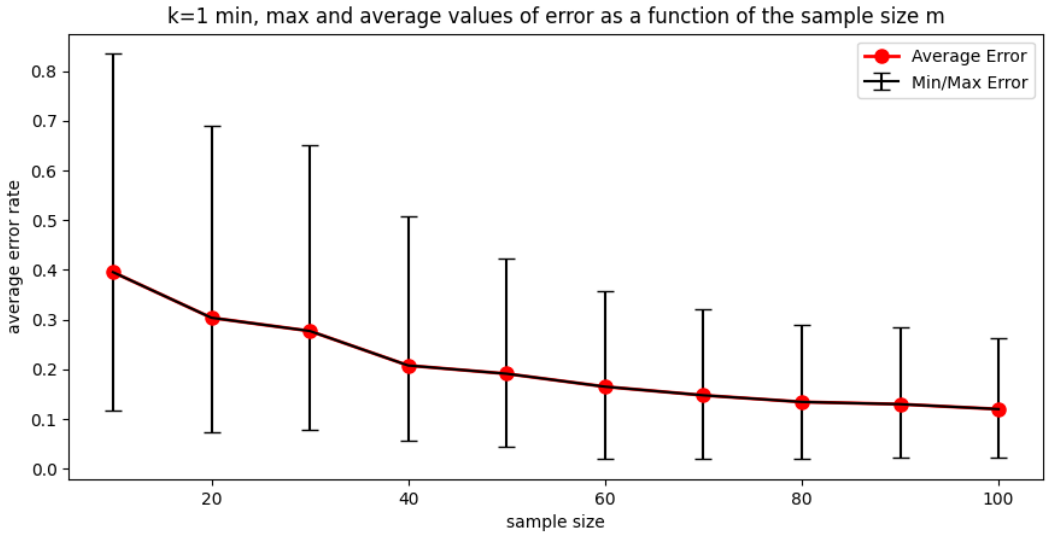
Introduction to Machine Learning

Exercise 1

Question 2

(a)

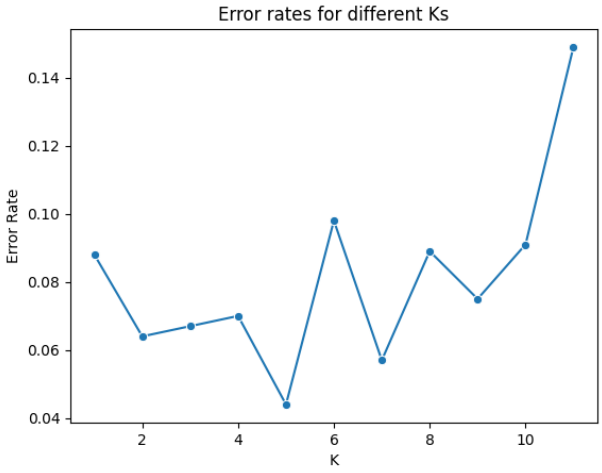


(b) Yes, we can see a trend of decreasing error as we increase the sample size. It makes sense because when we have more samples, each test sample will have closer training samples to look at their labels. The nearest neighbour algorithm (knn with k=1) relies on the assumption that the closer a training sample is to the test sample, the higher probability it will share its label and generate a better/precise result.

(c) Yes, we got different error results with the same sample size as the differences between the maximum and the minimum on each sample size shows. It happens because the training sample is chosen randomly upon each run so the learning process runs over a different sample each time, thus produce a different result.

(d) Yes, the size of the error bars shrinks as the sample size grows. The reason for that is that when the sample size grows the learning algorithm learns from more examples, hence generates more precise rule that lower the error.

(e)



As this graph shows, the K that gives the best results is K = 5. We learned that when the K is too low it doesn't take into account enough data from the sample, and when K is too high it takes other samples that are too far and it causes a distortion, which leads to higher error.

Question 3

a. **Given**: η of $\mathcal{D}$ is c-Lipschitz with respect to the Euclidean distance,
therefore, by definition $|\eta(x1) - \eta(x2)| \leq C * \rho(x1, x2) = C * \|x1 - x2\|$
Given: $\mathcal{D}$ has a Bayes-error of zero, **therefore**, we can assume that $\mathcal{D}$ has deterministic labels (otherwise, it was impossible to achieve such error).
Let $(x1, y1), (x2, y2) \in S$ and $y1 \neq y2$.
Without loss of generality, y1 = 1 and y2 = 0 (since Y = {0,1} and $y1 \neq y2$).
Since $\mathcal{D}$ has deterministic labels and Y = {0,1} we got that $\eta(x) := \eta(1) \Rightarrow \eta(x1) = 1$ and $\eta(x2) = 0 \Rightarrow |\eta(x1) - \eta(x2)| = 1 \leq C * \|x1 - x2\| \Rightarrow \|\boldsymbol{x1 - x2}\| \geq \frac{\mathbf{1}}{\boldsymbol{C}}$

b. Poof by contradiction:
Assume that $err(f_S^{nn}, \mathcal{D}) > 0$ it means there is a pair $(x, y) \in X$ so that $f_S^{nn}(x) \neq y$.
By the definition $f_S^{nn}(x) = Y_{nn}(x)$ it means that there is some $(x', y') \in S$ that x is a nearest neighbor of x' and $f_S^{nn}(x) \neq Y_{nn}(x) = y' \neq y$.
It's given in the question that:

- the set of balls $\mathbb{B}$ covers the space of points in X with balls with radius $\frac{1}{3c}$
- Every point $(x, y) \in X$ exist in at least one ball $\beta \in \mathbb{B}$.
- Every ball $\beta \in \mathbb{B}$ contains at least one point $(x', y') \in S$.

From the above we can say that for every point $(x, y) \in X$ there's a point $(x', y') \in S$ that resides in the same ball and since x is a nearest neighbour of x' , it means that the reside in the same ball and that $\rho(x, x') = \|x - x'\| \leq 2 * \frac{1}{3c}$ (worst case is when the 2 points are on the edges of the ball).

So formally, we have 2 points x, x' with different labels y, y' that satisfy $\|\boldsymbol{x1 - x2}\| \leq \frac{2}{3c} < \frac{\mathbf{1}}{\boldsymbol{C}}$ **in contradiction to the proof in item a above.**

Question 4

a. We can define X and Y in this problem as follows:

   X represents a rabbit, which is basically a vector of 2 features in $\mathbb{R}^2$ – the age and weight of the rabbit. Since the age of a rabbit is bound by 42 months and the weight is bounded by 5kg, we can define X := [0, 42] x [0, 5].

   The label set Y in this case is the color of the rabbit, which can be either black or white, hence Y:= {black, white}.

b. According to the Bayes-optimal rule definition: $h_{bayes}(x) \in argmax_{y \in Y} \eta_y(x)$

   In our case, based on the probability table, we can define $h_{bayes}(x)$ as follows:

$$h_{bayes}(x) = \begin{cases} black, x \in \{(8,4), (15,1) \\ white, x = (15,2) \end{cases}$$

c. $err(h, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[h(x) \neq Y] = \sum_{(x,y)\in X\ x\ Y: h(x)\neq y} \mathbb{P}[X = x, Y = y] =$

   $p(x = (8,4) \cap y = white) + p(x = (15,1) \cap y = white)$ = 0.06 + 0.07 = 0.13

d. So we will evaluate the error of every possible $h \in \mathcal{H}$ for that:

   Let $h_{black}: X \to Y$ so that $h_{black}(x) = black, \forall x \in X$ and $h_{white}: X \to Y$ so that $h_{white}(x) = white, \forall x \in X$.

   $\mathcal{H}$ is an hypothesis class that contains only constant functions, hence we can define:

   $\mathcal{H} = \{ h_{black}, h_{white}\}$ since it's the only possible constant function over the label set Y.

   Now we will evaluate the $err(h, d)$ for each $h \in \mathcal{H}$:

$$err(h_{black}, \mathcal{D}) = p(y = black \mid x = (8,4) + p(y = black \mid x = (15,1) + p(y = black \mid x = (15,2) = 0.06 + 0.07 + 0.24 = 0.37$$

$$err(h_{white}, \mathcal{D}) = p(y = white \mid x = (8,4) + p(y = white \mid x = (15,1) = 0.42 + 0.21 = 0.63$$

   According to the definition, the approximation error $\mathcal{H}$ is:

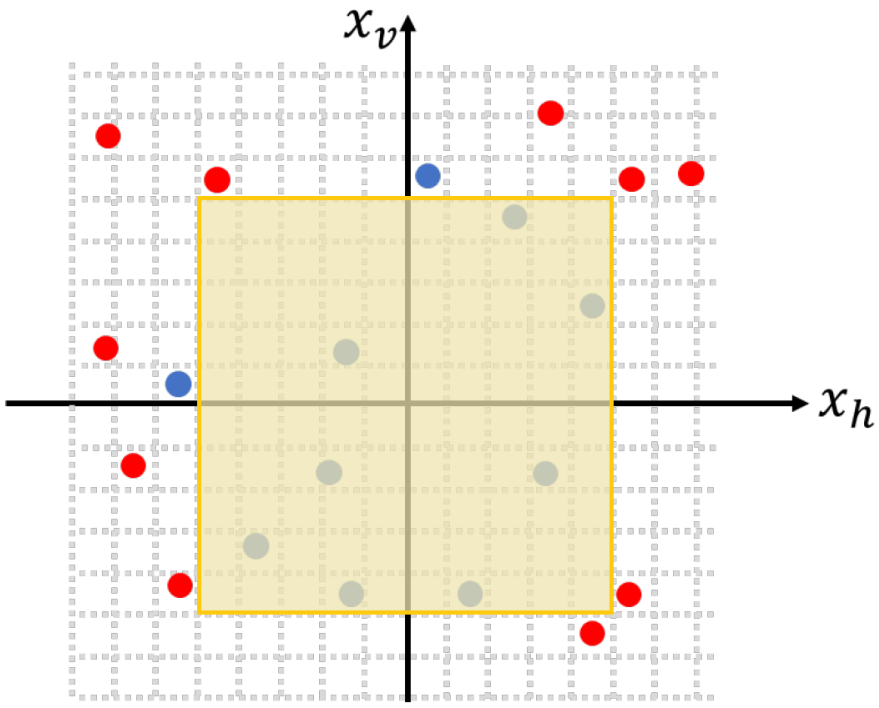
$$err_{app} := \min_{h\in\mathcal{H}} err(h,d) = \min\{ err(h_{black}, \mathcal{D}), err(h_{white}, \mathcal{D})\} = 0.37$$

e. $\mathbb{E}_{s\sim\mathcal{D}^m}\left[err\left(\hat{h}_s, \mathcal{D}\right)\right] = \frac{k-1}{k}\sum_{x\in X} p_x\,(1-p_x)^m = \frac{1}{2}(0.08 * 0.92^3 + 0.15 * 0.85^3 + 0.47 * 0.53^3 + 0.3 * 0.7^3)$ = 0.1636.

   The reason we can use this formula here is that it requires deterministic label conditioned on the input, so while $\mathcal{D}''$ have it, $\mathcal{D}$ not.

Question 5

(a) The empirical error achieved by ERM with the hypothesis class of rectangular is $\frac{1}{10}$, as shown in the following figure:

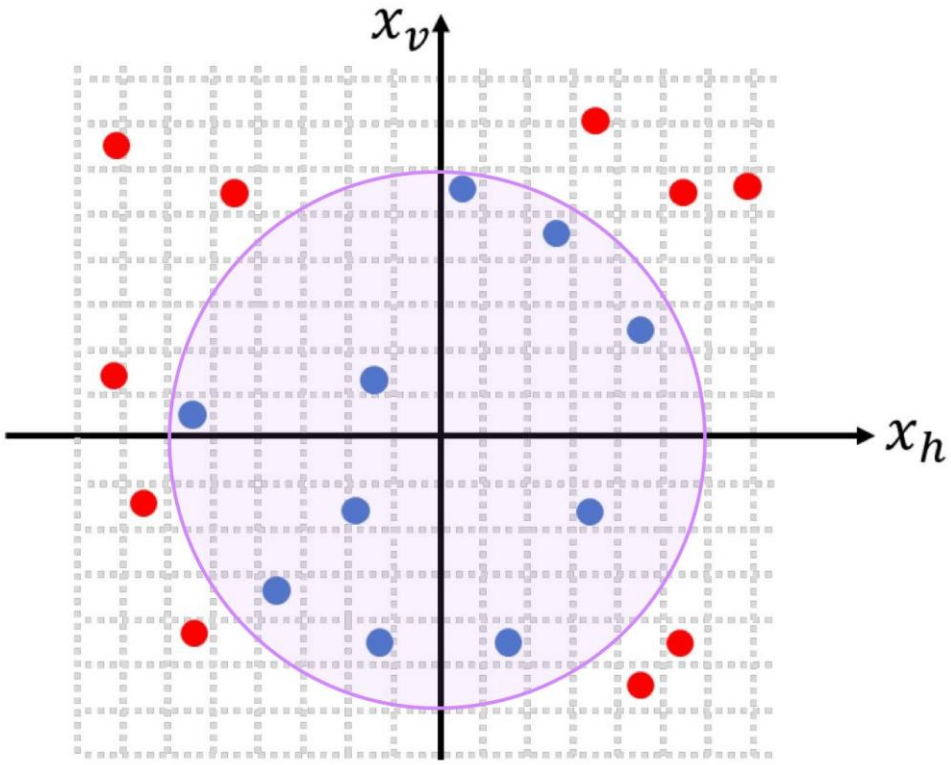


This rectangle in the figure is the threshold. For a sample point with offset $[x_h, x_v]$, if the offset is inside the rectangle, the sample point will be labelled $y = 1$, and $y = 0$ otherwise. Because the threshold for each axis is symmetrical, this rectangle is the best we can create, with the thresholding function $\tau_h = 5,\ \tau_v = 5$ (according to the measures of the gray grid), giving us 2 samples out of 20 with their wrong label.

(b) Lets define:

$$g(x_h, x_v) = \sqrt{x_h^2 + x_v^2}$$

The formula of g gives the distance of an input (x_h, x_v) from the origin, or the tree trunk. Using this g, the hypothesis class $\mathcal{H}_g$ has circular thresholding functions, where the center of the circle is the origin. The ERM in that case will pick a hypothesis that keeps the blue samples inside the threshold, and the red samples out, and therefore we will achieve zero empirical error for this sample, as shown in the following figure:



This circular prediction rule is better than the rectangular prediction rule because besides the zero empirical error that we mentioned, it makes more sense as the farmer believes that apples dropping from their tree are tastier if they are close to the tree trunk, this formula computes that distance from the tree trunk.