Supervised Learning (COMP0078) – Coursework 1 Version 2

Due : 20 November 2024

Submission

You may work individually or in a group of two. When working in a group, both students will receive the same grade. You/your group should produce a pdf with the answer to every question/task below, code should be submitted separately in a zip file (not printed in the pdf). You will not only be assessed on the **correctness/quality** of your answers but also on **clarity of presentation**. Additionally make sure that your code is *well commented*.

Please submit on Moodle:

1. your answers as a pdf file,
2. a zip file with your source code.

Regarding the use of libraries/packages for your code, you should implement regression (with and without basis functions, kernels) using matrix algebra directly and k-NN directly (not calling a function from e.g. scikit-learn to do that for you). Otherwise, libraries are okay in general.

Questions please e-mail `sl-support@cs.ucl.ac.uk` or the moodle *Questions* forum.

1 PART I [50%]

1.1 Linear Regression

Linear regression overview: Given a set of data:

$$\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\} \tag{1}$$

where $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a vector in $\mathbb{R}^n$ and y is a real number. Linear regression finds a vector $\boldsymbol{w} \in \mathbb{R}^n$ such that the sum of squared errors

$$\text{SSE} = \sum_{t=1}^{m} (y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2 \tag{2}$$

is minimized. This is expressible in matrix form by defining X to be the $m \times n$ matrix

$$X = \begin{pmatrix} x_{1,1} & x_{1,2} & \ldots & x_{1,n} \\ x_{2,1} & x_{2,2} & \ldots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \ldots & x_{m,n} \end{pmatrix}, \tag{3}$$

and defining $\boldsymbol{y}$ to be the column vector $\boldsymbol{y} = (y_1, \ldots, y_m)$. The vector $\boldsymbol{w}$ then minimizes

$$(X\boldsymbol{w} - \boldsymbol{y})^\intercal (X\boldsymbol{w} - \boldsymbol{y}).$$

In linear regression with basis functions we fit the data sequence with a linear combination of basis functions $(\phi_1, \phi_2, \ldots, \phi_k)$ where $\phi_i : \mathbb{R}^n \to \mathbb{R}$ which defines a feature map from $\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^k$

$$\boldsymbol{\phi}(\boldsymbol{x}) = (\phi_1(\boldsymbol{x}), \ldots, \phi_k(\boldsymbol{x})), \quad \boldsymbol{x} \in \mathbb{R}^n .$$

We use the basis functions to transform the data as follows

$$\{((\phi_1(\boldsymbol{x}_1),\ldots,\phi_k(\boldsymbol{x}_1)),y_1),\ldots,((\phi_1(\boldsymbol{x}_m),\ldots,\phi_k(\boldsymbol{x}_m)),y_m)\}, \tag{4}$$

and then applying linear regression above to this transformed dataset. In matrix notation we may denote this transformed dataset as

$$\Phi := \begin{pmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1) \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_m) \end{pmatrix} = \begin{pmatrix} \phi_1(\boldsymbol{x}_1) & \ldots & \phi_k(\boldsymbol{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\boldsymbol{x}_m) & \ldots & \phi_k(\boldsymbol{x}_m) \end{pmatrix}, \quad (m \times k) \tag{5}$$

Linear regression on the transformed dataset thus finds a k-dimensional vector $\boldsymbol{w} = (w_1,\ldots,w_k)$ such that

$$(\Phi\boldsymbol{w} - y)^T(\Phi\boldsymbol{w} - y) = \sum_{t=1}^{m}(y_t - \sum_{i=1}^{k} w_i\phi_i(\boldsymbol{x}_t))^2 \tag{6}$$

is minimized.

A common basis ($n = 1$) used is the polynomial basis $\{\phi_1(x) = 1, \phi_2(x) = x, \phi_3(x) = x^2, \phi_4(x) = x^3, \ldots, \phi_k(x) = x^{k-1}\}$ of dimension k (order $k-1$) in the figure below we give a simple fit of four points produced by a linear ($k = 2$) and cubic ($k = 4$) polynomial.

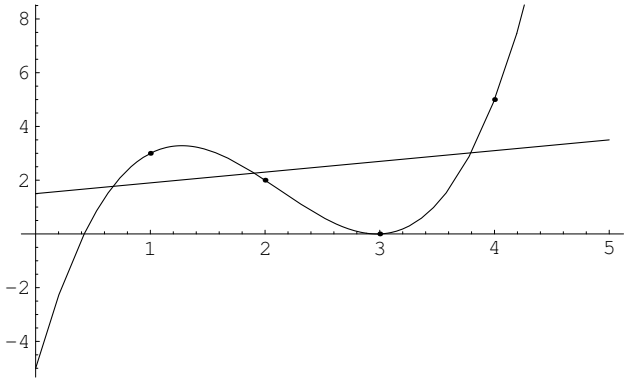


Figure 1: Data set $\{(1,3),(2,2),(3,0),(4,5)\}$ fitted with basis $\{1,x\}$ and basis $\{1,x,x^2,x^3\}$

1. **[5 pts]:** For each of the polynomial bases of dimension $k = 1,2,3,4$ fit the data set of Figure 1 $\{(1,3),(2,2),(3,0),(4,5)\}$.

   (a) Produce a plot similar to Figure 1, superimposing the four different curves corresponding to each fit over the four data points.

   (b) Give the equations corresponding to the curves fitted for $k = 1,2,3$. The equation corresponding to $k = 4$ is $-5 + 15.17x - 8.5x^2 + 1.33x^3$.

   (c) For each fitted curve $k = 1,2,3,4$ give the mean square error where MSE $= \frac{\text{SSE}}{m}$.

2. **[10 pts]:** In this part we will illustrate the phenomena of *overfitting*.

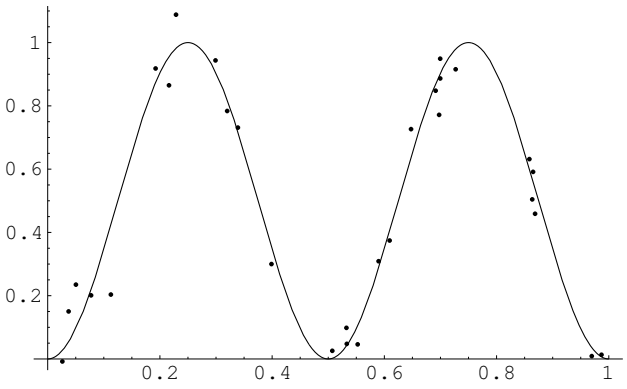
   (a) Define

   $$g_\sigma(x) := \sin^2(2\pi x) + \epsilon. \tag{7}$$

   where ϵ is a random variable distributed normally with mean 0 and variance σ^2 thus $g_\sigma(x)$ is a *random* function such that $\sin^2(2\pi x)$ is computed and then the normal noise is added on each "call" of the function. We then sample "x_i" uniformly at random from the interval $[0,1]$ 30 times creating $(x_1,\ldots,x_{30})$ and apply $g_{0.07}$ to each x creating the data set

   $$S_{0.07,30} = \{(x_1, g_{0.07}(x_1)),\ldots,(x_{30}, g_{0.07}(x_{30})\}. \tag{8}$$

   i. Plot the function $\sin^2(2\pi x)$ in the range $0 \le x \le 1$ with the points of the above data set superimposed. The plot should resemble



ii. Fit the data set with a polynomial bases of dimension $k = 2, 5, 10, 14, 18$ plot each of these 5 curves superimposed over a plot of data points.[1]

(b) Let the training error $\text{te}_k(S)$ denote the MSE of the fitting of the data set S with polynomial basis of dimension k. Plot the natural log (ln) of the training error versus the polynomial dimension $k = 1, \ldots, 18$ (this should be a decreasing function).

(c) Generate a test set T of a thousand points,

$$T_{0.07,1000} = \{(x_1, g_{0.07}(x_1)), \ldots, (x_{1000}, g_{0.07}(x_{1000})\}. \tag{9}$$

Define the test error $\text{tse}_k(S, T)$ to be the MSE of the test set T on the polynomial of dimension k fitted from training set S. Plot the ln of the test error versus the polynomial dimension $k = 1, \ldots, 18$. Unlike the training error this is not a decreasing function. This is the phenomena of *overfitting*. Although the training error decreases with growing k the test error eventually increases since rather than fitting the function, in a loose sense, we begin to fit to the noise.

(d) For any given set of random numbers we will get slightly different training curves and test curves. It is instructive to see these curves smoothed out. For this part repeat items (b) and (c) but instead of plotting the results of a single "run" plot the average results of a 100 runs (note: plot the ln(avg) rather than the avg(ln)).

3. **[5 pts]:** Now use basis (for $k = 1, \ldots, 18$)

$$\{\sin(1\pi x), \sin(2\pi x), \sin(3\pi x), \ldots, \sin(k\pi x)\}.$$

Repeat the experiments in 2 (b-d) with the above basis.

1.2 Filtered Boston housing and kernels

In this section we will use kernel methods to extend linear regression. Boston housing is a classic dataset where you are given 13 values and a goal is to predict the 14th which is the median house price. Instead of the original dataset we will instead use a modified dataset removing the ethically-suspect "column B." `Thus we will use 12 attributes to predict the 13th.` The unmodified dataset is described in more detail at `http://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html`. There are 506 entries which we will split into train and test.

The **filtered** boston housing data set as a ".csv' file is located at

`http://www.cs.ucl.ac.uk/staff/M.Herbster/boston-filter`

4. **[10 pts]:** "Baseline versus full linear regression."
Rather than use all of our attributes for prediction it is often useful to see how well a baseline method works for a problem. In this exercise, we will compare the following:

[1] Depending on you implementation you may have numerical errors for large values of k this is 'ok' for the purposes of this exercise.

(a) Predicting with the mean y-value on the training set.

(b) Predicting with a single attribute and a bias term.

(c) Predicting with all the attributes

The training set should be 2/3, and the test set should be 1/3, of your data in (a)-(c). In the following average your results over 20 runs (each run based on a different (2/3,1/3) random split).

a. Naive Regression. Create a vector of ones that is the same length as the training set using the function `ones`. Do the same for the test set. By using these vectors we will be fitting the data with a constant function. Perform linear regression on the training set. Calculate the MSE on the training and test sets and note down the results.

b. Give a simple interpretation of the constant function in 'a.' above.

c. Linear Regression with single attributes. For each of the twelve attributes, perform a linear regression using only the single attribute but incorporating a bias term so that the inputs are augmented with an additional 1 entry, $(\boldsymbol{x}_i, 1)$, so that we learn a weight vector $\boldsymbol{w} \in \Re^2$. For each of the twelve regressions, calculate the MSE on the training and test sets and note down the results.

d. Linear Regression using all attributes. Now we would like to perform linear regression using all of the data attributes at once.

   Perform linear regression on the training set using this regressor, and incorporate a bias term as above. Calculate the MSE on the training and test sets and note down the results. You should find that this method outperforms any of the individual regressors.

1.3 Kernelised ridge regression

For nonlinear regression, the dual version will prove important. Obviously, linear regression is not capable of achieving a good predictive performance on a nonlinear data set. Here, the dual formulation will prove extremely useful, in combination with the *kernel trick*.

For our exercises we will use a slight variation on the optimisation (as compared to the notes) that defines ridge regression for a training set with ℓ examples,

$$\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w}\in\mathbb{R}^n} \frac{1}{\ell}\sum_{i=1}^{\ell}(\boldsymbol{x}_i^\intercal \boldsymbol{w} - y_i)^2 + \gamma \boldsymbol{w}^\intercal \boldsymbol{w}. \tag{10}$$

i.e, we have replace the sum of the square errors with the mean square error (MSE).[2] For a given kernel function K define the kernel matrix $\boldsymbol{K}$ for a training set of size ℓ elementwise via

$$K_{i,j} := K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

The dual optimisation formulation after kernelization is

$$\boldsymbol{\alpha}^* = \text{argmin}_{\boldsymbol{\alpha}\in\mathbb{R}^\ell} \frac{1}{\ell}\sum_{i=1}^{\ell}(\sum_{j=1}^{\ell}\alpha_j K_{i,j} - y_i)^2 + \gamma \boldsymbol{\alpha}^\intercal \boldsymbol{K}\boldsymbol{\alpha}. \tag{11}$$

Now if we use $\boldsymbol{y} := (y_1, \ldots, y_\ell)^\intercal$ to denote a vector that contain the y-values of training set we may solve in the dual as follows

$$\boldsymbol{\alpha}^* = (\boldsymbol{K} + \gamma\ell\mathbf{I}_\ell)^{-1}\boldsymbol{y} \tag{12}$$

[2]The motivation is that we wish the regularisation parameter γ to 'scale' somewhat independently of training set size.

where $\mathbf{I}_\ell$ denotes the $\ell \times \ell$ identity matrix. The evaluation of the regression function on a test point can be reformulated as:

$$y_{\text{test}} = \sum_{i=1}^{\ell} \alpha_i^* K(\boldsymbol{x}_i, \boldsymbol{x}_{\text{test}}) \tag{13}$$

where the K is the kernel function.

5. [**20 pts**] "Kernel Ridge Regression"
In this exercise we will perform kernel ridge regression (KRR) on the data set using the Gaussian kernel,

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left(-\frac{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma^2}\right). \tag{14}$$

For this question, you will hold out 2/3 of data for training and report the test results on the remaining 1/3.

   a. Create a vector of γ values $[2^{-40}, 2^{-39}, \ldots, 2^{-26}]$ and a vector of σ values $[2^7, 2^{7.5}, \ldots, 2^{12.5}, 2^{13}]$ (recall that σ is a parameter of the Gaussian kernel see equation (14)). Perform kernel ridge regression on the *training* set using five-fold cross-validation to choose among all pairing of the values of γ and σ. Choose the γ and σ values that perform the best to compute the predictor (by then retraining with those parameters on the training set) that you will use to report the test and training error.

   b. Plot the "cross-validation error" (mean over folds of validation error) as a function of γ and σ.

   c. Calculate the MSE on the training and test sets for the best γ and σ.

   d. Repeat "exercise 4a,c,d" and "exercise 5a,c" over 20 random $(2/3, 1/3)$ splits of your data. Record the train/test error and the standard deviations (σ') of the train/test errors and summarise these results in the following type of table. *Additional clarification:* when repeating 5a,c: you will thus be generating 20 best (γ, σ) pairs.

Method	**MSE train**	**MSE test**
Naive Regression	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 1)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 2)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 3)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 4)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 5)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 6)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 7)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 8)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 9)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 10)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 11)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (attribute 12)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Linear Regression (all attributes)	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?
Kernel Ridge Regression	??.?? $\pm$ σ'?	??.?? $\pm$ σ'?

2 PART II [20%]

2.1 k-Nearest Neighbors

In this section you will implement the k-NN algorithm and explore its performance as a function of k. Given a new data point, the k-NN algorithm attributes its label to the majority label of its k nearest neighbors. See here for details.

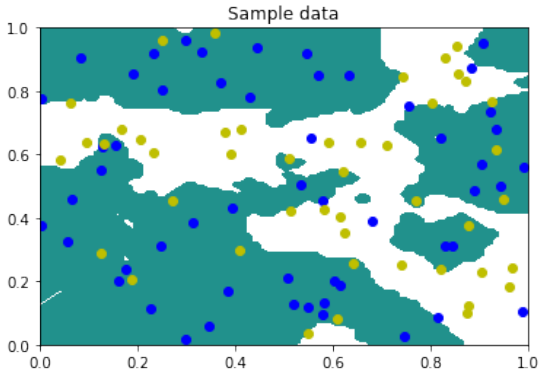


Figure 1: A hypothesis $h_{S,v}$ visualized with $|S| = 100$ and $v = 3$.

2.1.1 Generating the data

A voted-center hypothesis is a function $h_{S,v} : [0,1]^2 \to \{0, 1, \square\}$ where $S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_{|S|}, y_{|S|})\}$ is a set of labeled centers with each $\boldsymbol{x}_i \in [0,1]^2$ and $y_i \in \{0,1\}$ and v is a positive integer. In this coursework we will always set $v := 3$ for all of of the exercises in this subsection. Where $h_{S,v}(\boldsymbol{x}) = 0$ if the majority of the v nearest $\boldsymbol{x}_i$'s to $\boldsymbol{x}$ in S are '0', similarly for $h_{S,v}(\boldsymbol{x}) = 1$ and finally in certain "corner" cases the "the majority of v ..." will not be well-defined and in that case $h_{S,v}(\boldsymbol{x}) = \square$. See Section 2.1.4 for addressing those corner cases.

In Figure 1 one such hypothesis is visualised with $|S| = 100$ and $v = 3$. The white area is the mapping to 0 and the turquoise is the mapping to 1, the corresponding centers are green and blue.

6. **[5 pts]** *Produce a visualisation of an $h_{S,v}$ similar to the figure.*

   First we will need to generate a random h. Do this by sampling 100 centers uniformly at random from $[0,1]^2$ with 100 corresponding labels sampled uniformly at random from $\{0,1\}$. Later we will need to generate more random hypotheses. Call this distribution over hypotheses generated by this random process $p_{\mathcal{H}}$.

2.1.2 Estimated generalization error of k-NN as a function of k

In this section we visualise the performance as a function of k. You will produce a figure where the vertical axis is the estimated generalisation error and the horizontal axis is $k = 1, \ldots, 49$.

Generating our noisy data. Given an $h_{S,v}$ the underlying probability distribution $p_h(\boldsymbol{x}, y)$ is determined by sampling an $\boldsymbol{x}$ uniformly at random from $[0,1]^2$ and then its corresponding y value is then generated by flipping biased coin $P(\text{heads}) = 0.8/P(\text{tails}) = 0.2$ if heads comes up then $y = h_{S,3}(\boldsymbol{x})$ otherwise y is sampled uniformly at random from $\{0,1\}$. When you generate training and test sets they will both come from this distribution.

7. **[7 pts]** a) Produce a visualisation using Protocol A (see Figure 2). b) Using roughly 5 sentences explain why the figure is the approximate shape that it is and comment on any interesting features of the figure.

2.1.3 Determine the optimal k as a function of the number of training points (m)

In this section you will estimate the optimal k as a function of the number of training points. Perform the following experimental protocol.

8. **[8 pts]** a) Produce a visualisation using Protocol B (see Figure 3). I.e, plotting m versus optimal k b) Using roughly 4 sentences explain why the figure is the approximate shape that it is.

```
For each k ∈ {1, ..., 49}
    Do 100 runs ...
        Sample a h from p_H
        Build a k-NN model with 4000 training points sampled from p_h(x, y).
        Run k-NN estimate generalisation error (for this run)
          using 1000 test points sampled from p_h(x, y)
    The estimated generalization error (y-axis) is then the mean
      of these 100 ''run'' generalisation errors.
```

Figure 2: Experimental Protocol A: k-NN. For computational purposes, you may reorder the loops in the protocol as long as this is done in a principled way.

```
For each m ∈ {100, 500, 1000, 1500, ..., 4000}
    Do 100 runs ...
         For each k ∈ {1, ..., 49}
              Sample a h from p_H
              Build a k-NN model with m training points sampled from p_h(x, y).
              Run k-NN estimate generalisation error (for this run)
                using 1000 test points sampled from p_h(x, y)
         The estimated optimal k (for this run) is then the k
           with minimal estimated generalisation error.
    The estimated optimal k (y-axis) is then the mean
      of these 100 ''run'' optimal 'k's.
```

Figure 3: Experimental Protocol B: k-NN. For computational purposes, you may reorder the loops in the protocol, as long as this is done in a principled way.

2.1.4 Additional Clarifications

Note both the k-NN algorithm and the hypothesis h can produced "undefined" results in certain corner cases for individual $\boldsymbol{x}$'s. Resolve these case by generating either the label or prediction uniformly at random.

3 PART III [30%]

3.1 Questions

Notation: We overload $[\cdot]$ as follows $[n] := \{1, 2, \ldots, n\}$ if n is a positive integer and $[\texttt{pred}] = 1$ if `pred` is a logical predicate which is true and $[\texttt{pred}] = 0$ otherwise.

9. **[5 pts]** *Kernel modification* Consider the function $K_c(\boldsymbol{x}, \mathbf{z}) := c + \sum_{i=1}^{n} x_i z_i$ where $\boldsymbol{x}$, $\mathbf{z} \in \mathbb{R}^n$.

   (a) For what values of $c \in \mathbb{R}$ is K_c a positive semidefinite kernel? Give an argument supporting your claim. ("The closer your argument is to a proof the more likely it is to receive full credit.")

   (b) Suppose we use K_c as a kernel function with linear regression (least squares). Explain how c influences the solution.

10. **[10 pts]** Suppose we perform linear regression with a Gaussian kernel $K_\beta(\boldsymbol{x}, \mathbf{t}) = \exp(-\beta\|\boldsymbol{x} - \mathbf{t}\|^2)$ to train a classifier on a dataset $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m) \in \mathbb{R}^n \times \{-1, 1\}$. Thus obtaining a function $f : \mathbb{R}^n \to \mathbb{R}$ which is of the form $f(\mathbf{t}) = \sum_{i=1}^{m} \alpha_i K_\beta(\boldsymbol{x}_i, \mathbf{t})$. The corresponding classifier is then $\text{sign}(f(\mathbf{t}))$. This classifier depends on the parameter β selected for the kernel. In what scenario will chosen β enable the trained linear classifier to simulate a 1-Nearest Neighbor classifier trained on the same dataset?

Note $\beta \in \mathbb{R}$ is a fixed scalar which may be selected as a function of $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ and the test point $\mathbf{t}$, i.e., we may use a function so that $\beta = \hat{\beta}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m, \mathbf{t})$ an exact function $\hat{\beta}(\cdot)$ need not be given just an argument that one exists.

Give an argument supporting your reasoning. ("The closer your argument is to a proof the more likely it is to receive full credit.")

11. **[15 pts]** We will show that there is no *"free lunch"* in machine learning, namely that without making assumptions on the learning setting, we cannot provide guarantees on how good an algorithm can be at solving a learning problem.

    We consider a binary classification setting with $\mathcal{X}$ and $\mathcal{Y} = \{-1, 1\}$ the input and output sets. For any probability distribution ρ on $\mathcal{X} \times \mathcal{Y}$ and any function $f : \mathcal{X} \to \mathcal{Y}$, we denote

    $$\mathcal{E}_\rho(f) = \int_{\mathcal{X}\times\mathcal{Y}} \mathbf{1}_{\{f(x)\neq y\}} \, d\rho(x, y)$$

    the misclassification excess risk with respect to ρ. Here $\mathbf{1}_{\{f(x)\neq y\}}$ is equal to 0 when $f(x) = y$ and 1 otherwise. In the following, always assume $\mathcal{X}$ to have infinite cardinality (i.e. infinite number of elements). Denote $S = (x_i, y_i)_{i=1}^n$ a dataset of n input-output pairs. In the following you will be showing that, for any algorithm $A : S \mapsto (f : \mathcal{X} \to \mathcal{Y})$ and for any integer $n \in \mathbb{N}$, there exists a distribution ρ such that

    - $\inf_{f:\mathcal{X}\to\mathcal{Y}} \ \mathcal{E}_\rho(f) = 0$
    - $\mathbb{P}_{S\sim\rho^n}\left(\mathcal{E}_\rho(A(S)) > \frac{1}{8}\right) \geq \frac{1}{7}$

    where $\mathbb{P}_{S\sim\rho^n}$ (and also $\mathbb{E}_{S\sim\rho^n}$ in the following) denotes the probability (resp. expectation) with respect to the random variable $S = (x_i, y_i)_{i=1}^n$ with each pair (x_i, y_i) independently sampled from ρ. We organized this proof along multiple steps (if you find you are not able to prove one step, don't get stuck! Go to the next step and assume the previous one to be true). Give an argument supporting your reasoning. ("The closer your argument is to a proof the more likely it is to receive full credit."):

    (a) Let $C \subset \mathcal{X}$ be a subset of $\mathcal{X}$ with cardinality (i.e. number of elements) $|C| = 2n$. Denote with $\mathcal{Y}^C = \{f_1, \ldots, f_T\}$ the set of all possible functions $f : C \to \mathcal{Y}$. We have $T = |\mathcal{Y}^C| = 2^{2n}$. For any $i = 1, \ldots, T$ denote with ρ_i the distribution over $C \times \mathcal{Y}$ such that

    $$\rho_i(\{(x, y)\}) = \begin{cases} \frac{1}{2n} & \text{if } y = f_i(x) \\ 0 & \text{otherwise} \end{cases} \qquad \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

    Show that $\mathcal{E}_{\rho_i}(f_i) = \inf_{f:\mathcal{X}\to\mathcal{Y}} \ \mathcal{E}_{\rho_i}(f) = 0$

    (b) Let C^n be the set of all possible *input* datasets (of size n) of points in C. Then $C^n = \{S_1, \ldots, S_k\}$ with $k = |C^n| = (2n)^n$. For every $j = 1, \ldots, k$ and input set $S_j = (x_1, \ldots, x_n) \in C^n$, denote $S_j^i = \{(x_1, f_i(x_1)), \ldots, (x_n, f_i(x_n))\}$ the corresponding input-output dataset for every $i = 1, \ldots, T$. Show that

    $$\max_{i=1,\ldots,T} \ \mathbb{E}_{S\sim\rho_i^n} \ \mathcal{E}_{\rho_i}(A(S)) \geq \min_{j=1,\ldots,k} \ \frac{1}{T}\sum_{i=1}^T \mathcal{E}_{\rho_i}(A(S_j^i))$$

    (Hint: show that for any $i = 1, \ldots, T$

    $$\mathbb{E}_{S\sim\rho_i^n} \ \mathcal{E}_{\rho_i}(A(S)) = \frac{1}{k}\sum_{j=1}^k \mathcal{E}_{\rho_i}(A(S_j^i))$$

    Hint 2: use the fact for any set of scalars $\alpha_1, \ldots, \alpha_m$ we have $\max_\ell \alpha_\ell \geq \frac{1}{m}\sum_{\ell=1}^m \alpha_\ell \geq \min_\ell \alpha_\ell$).

    (c) Fix $j \in \{1, \ldots, k\}$ with $S_j = \{(x_1, \ldots, x_n\}$ and denote $R_j = \{v_1, \ldots, v_p\}$ the subset of points of C that do not belong to S_j. Show that

    $$\frac{1}{T}\sum_{i=1}^T \mathcal{E}_{\rho_i}(A(S_j^i)) \geq \frac{1}{2} \ \min_{v\in R_j} \ \frac{1}{T}\sum_{i=1}^T \mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}}$$

(hint: lower bound the risk with respect to the errors only over S'_j. Use again the fact that $\frac{1}{m}\sum_{\ell=1}^{m}\alpha_m \geq \min_m \alpha_m$)

(d) With the same assumptions as above, show that for any $v \in R_j$

$$\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}} = \frac{1}{2}.$$

(hint: show that for any $v \in R_j$ we can partition $\mathcal{Y}^C$ into $T/2$ pairs $(f_i, f_{i'})$ such that $f_i(x) \neq f_{i'}(x)$ if and only if $x = v$. Show that $\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}} + \mathbf{1}_{\{A(S_j^{i'})(v)\neq f_{i'}(v)\}} = 1$.)

(e) We prove an auxiliary result: let Z be a random variable with values in $[0,1]$ and let $\mathbb{E}[Z] = \mu$. Show that for any $a \in (0,1)$

$$\mathbb{P}(Z > 1-a) \geq \frac{\mu-(1-a)}{a}$$

(hint: use Markov's inequality.)

(f) Combine the results of previous excercises to show that for any algorithm A and any integer n there exists a distribution ρ over $\mathcal{X}\times\mathcal{Y}$ such that

$$\mathbb{P}_{S\sim\rho^n}\left(\mathcal{E}_\rho(A(S)) > \frac{1}{8}\right) \geq \frac{1}{7}$$

(hint: use $Z = \mathcal{E}_\rho(A(S))$ in the previous bound. Use previous results to lower bound $\mathbb{E}_{S\sim\rho_i^n} Z$ for all i).

(g) **No-Free-Lunch (Open-ended questions).** You just proved the No-Free-Lunch theorem for machine learning.

i. Summarize the above results in the formal statement of a theorem.

ii. Consider the definition of *learnable space of function*:
DEFINITION. A space $\mathcal{H}$ of functions $f : \mathcal{X} \to \mathcal{Y}$ is **learnable** if there exists an algorithm A such that, for any $\epsilon, \delta \in (0,1)$ and any distribution ρ over $\mathcal{X}\times\mathcal{Y}$ there exists an integer $\bar{n}$ such that for any $n \geq \bar{n}$

$$\mathbb{P}_{S\sim\rho^n}\left(\mathcal{E}_\rho(A(S)) - \inf_{f\in\mathcal{H}}\mathcal{E}_\rho(f) \leq \epsilon\right) \geq 1-\delta$$

Compare this definition with the results above. Does the No-Free-Lunch theorem imply that the space $\mathcal{Y}^{\mathcal{X}}$ of all functions $f : \mathcal{X} \to \mathcal{Y}$ is (or is not) *learnable*? Why?

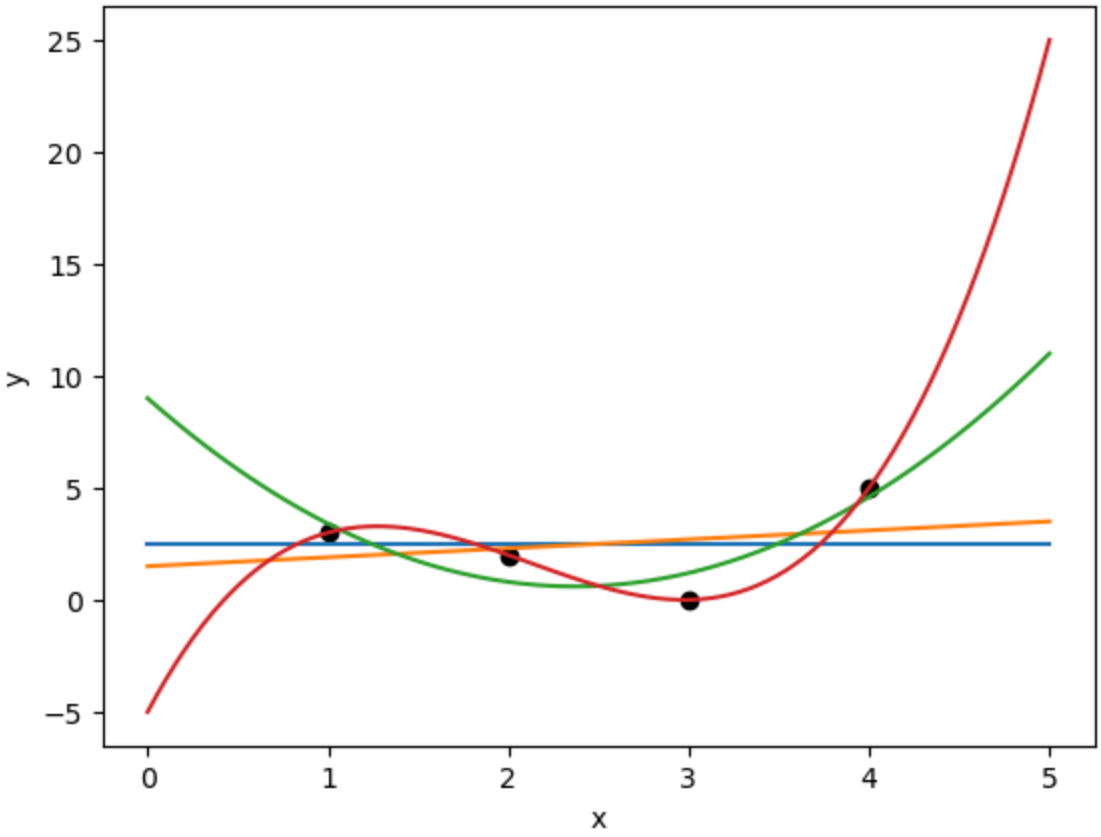
iii. Comment on the implications of the No-Free-Lunch theorem with respect to the design of a machine learning algorithm.

Lecture Title: Coursework - 1

Part 1

1.1 Linear Regression

1.1.1 (a)



1.1.1 (b)

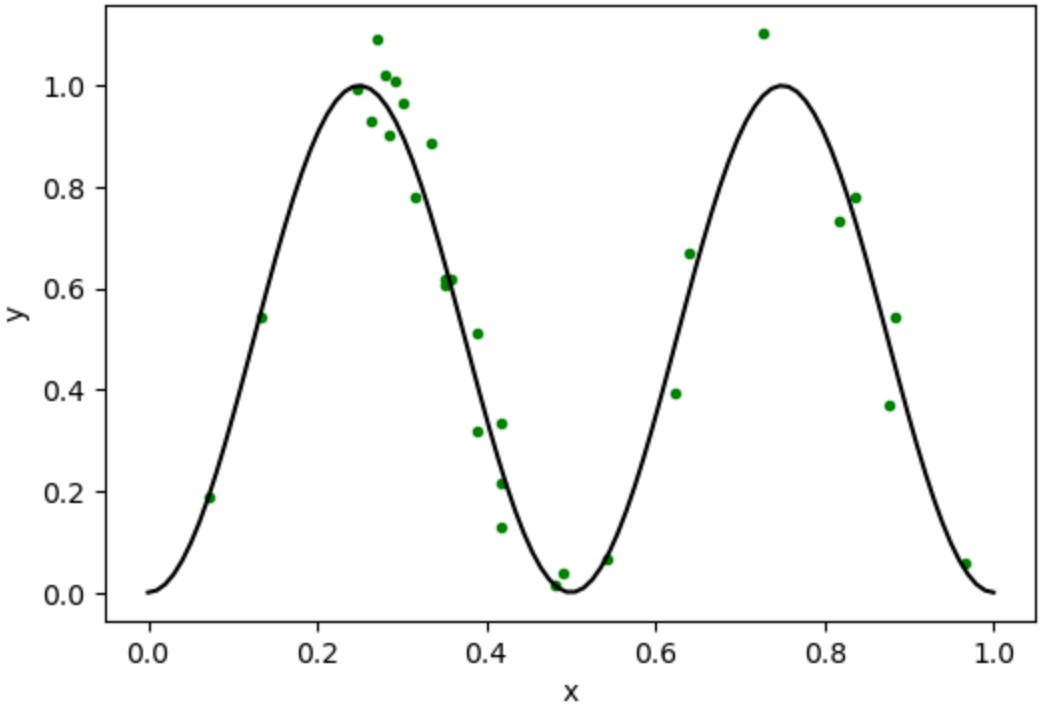
```
The equation corresponding to K: 1:
2.50
The equation corresponding to K: 2:
1.50 + 0.40x^1
The equation corresponding to K: 3:
9.00 - 7.10x^1  + 1.50x^2
```

1.1.1 (c)

```
 k: 1 MSE: 3.25
 k: 2 MSE: 3.05
 k: 3 MSE: 0.7999999999999999
 k: 4 MSE: 1.8374029875272024e-24
```

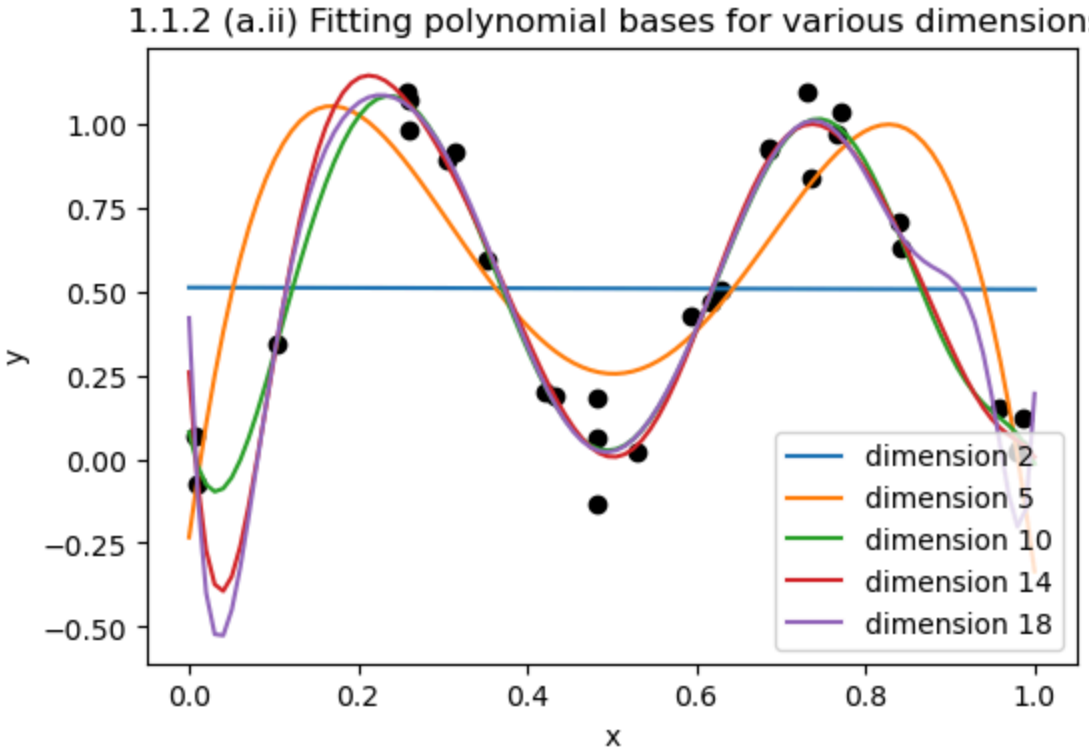
1.1.2

a.i

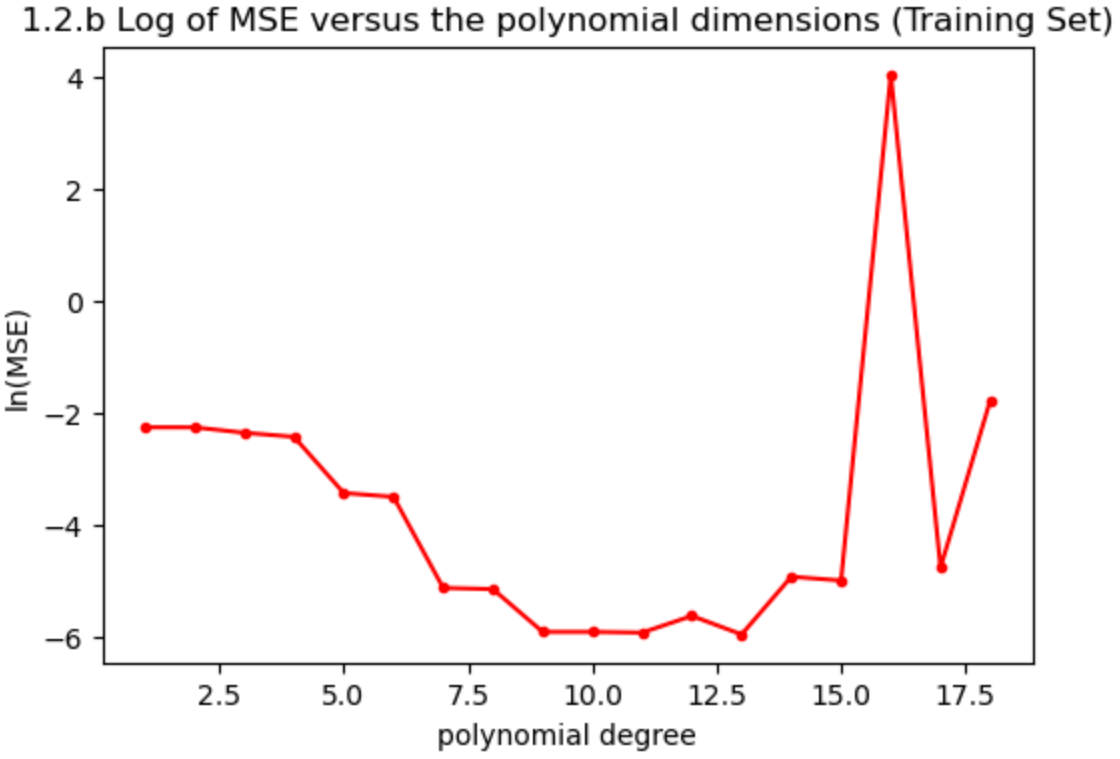


1.1.2

a.ii

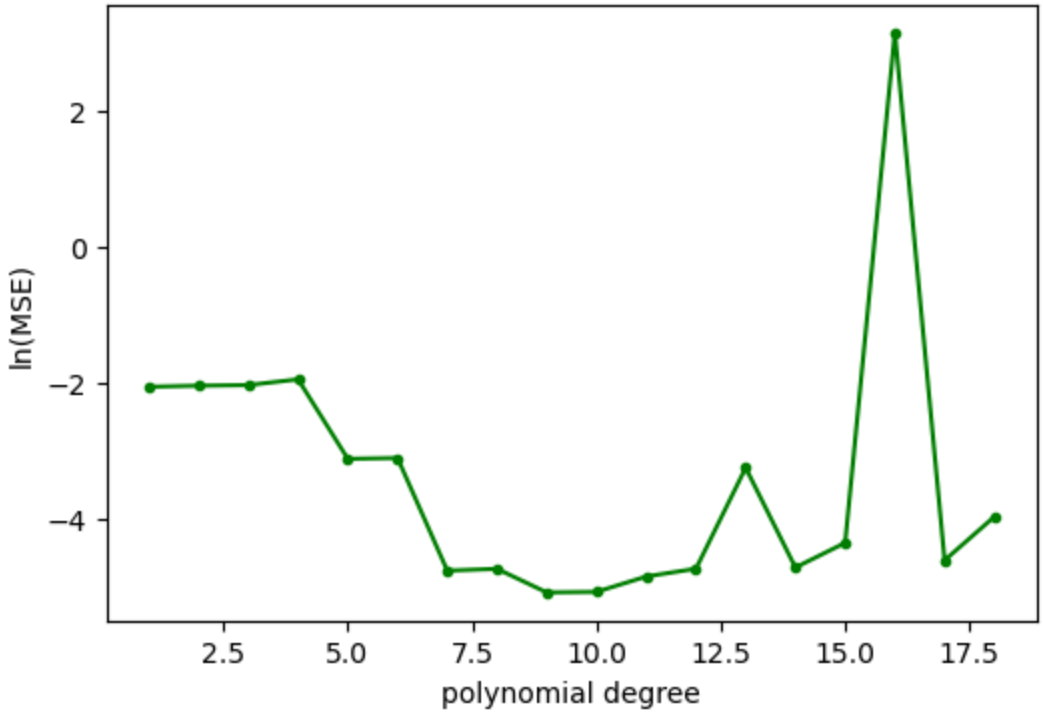

1.1.2 (a.ii) Fitting polynomial bases for various dimensions
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
y
0.0
0.2
0.4
0.6
0.8
1.0
x
dimension 2
dimension 5
dimension 10
dimension 14
dimension 18

1.1.2


1.2.b Log of MSE versus the polynomial dimensions (Training Set)
4
2
0
−2
−4
−6
ln(MSE)
2.5
5.0
7.5
10.0
12.5
15.0
17.5
polynomial degree

1.1.2 (c)

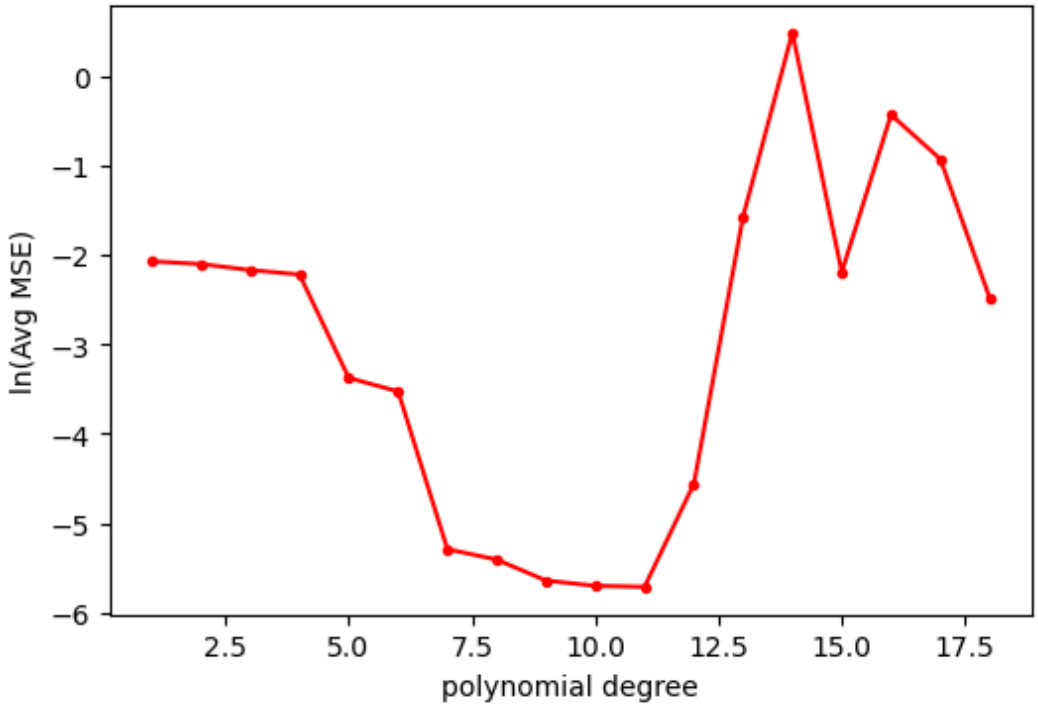
1.2.c Demonstrating Overfitting in Polynomial Regression (Testing Set)



1.1.2

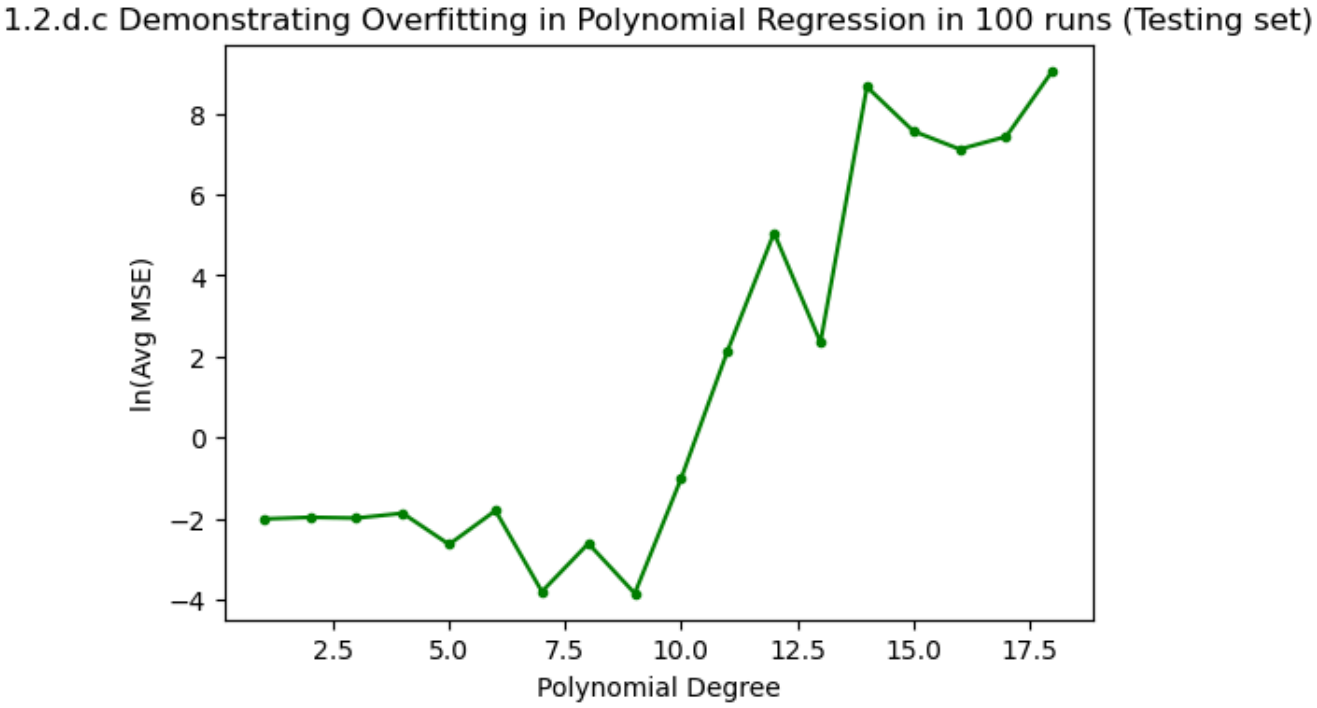
d-b

1.2.d.b Log of MSE versus the polynomial dimension for 100 runs (Training Set)

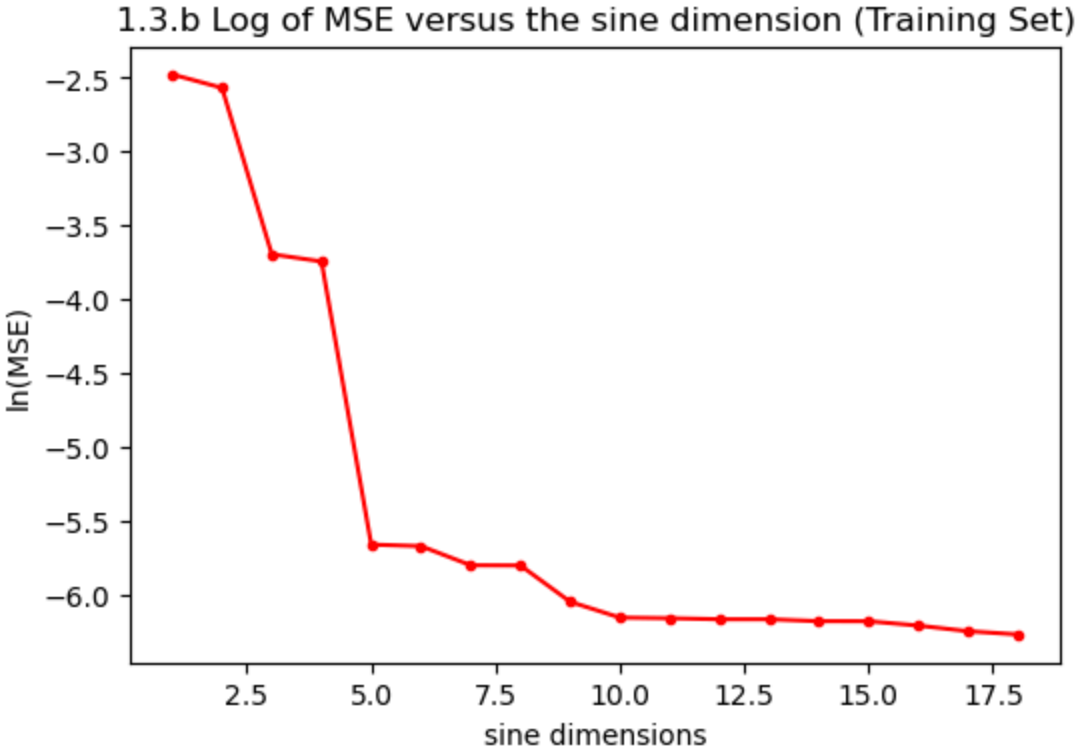


1.1.2

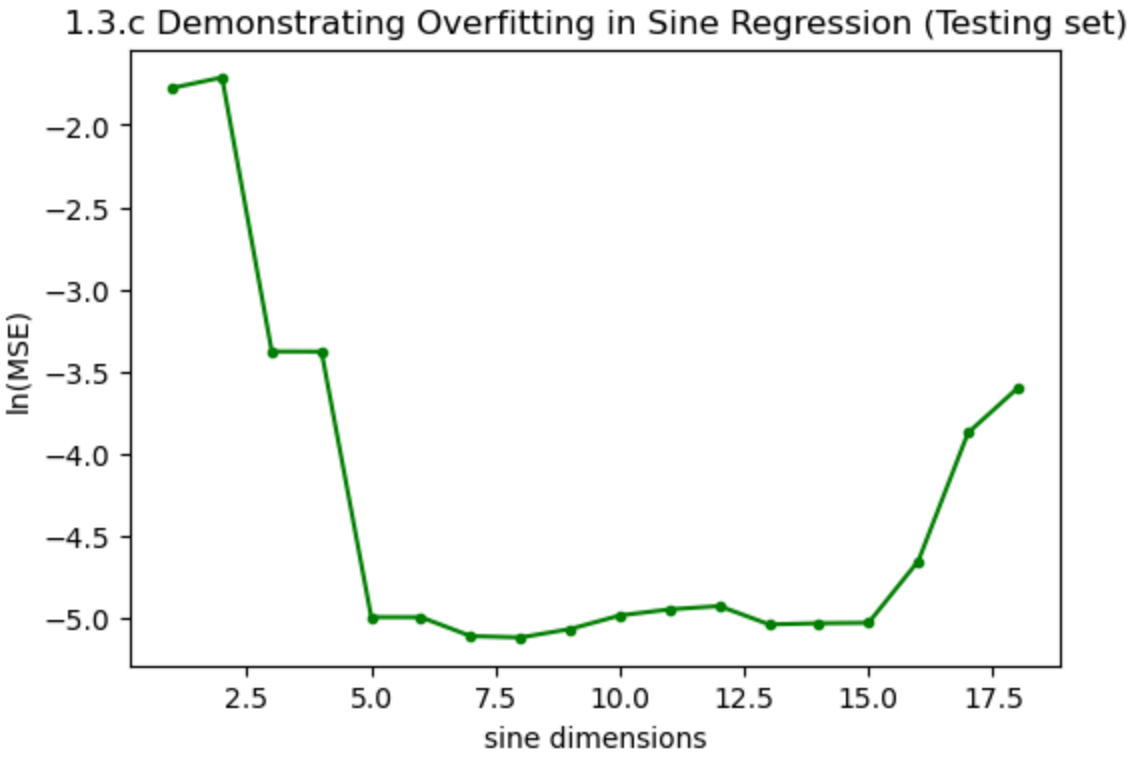
d-c


1.2.d.c Demonstrating Overfitting in Polynomial Regression in 100 runs (Testing set)
ln(Avg MSE)
8
6
4
2
0
−2
−4
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Polynomial Degree

1.1.3 (b)

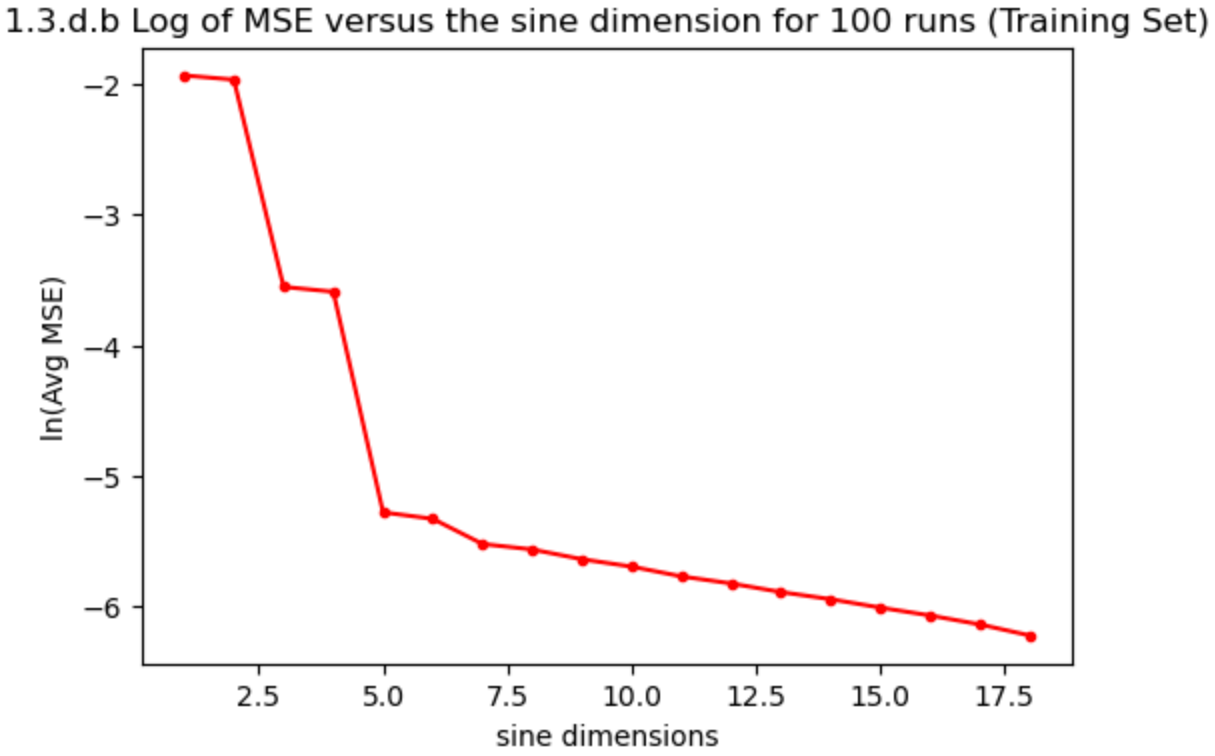

1.3.b Log of MSE versus the sine dimension (Training Set)
ln(MSE)
−2.5
−3.0
−3.5
−4.0
−4.5
−5.0
−5.5
−6.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
sine dimensions

1.1.3 (c)


1.3.c Demonstrating Overfitting in Sine Regression (Testing set)
ln(MSE)
−2.0
−2.5
−3.0
−3.5
−4.0
−4.5
−5.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
sine dimensions

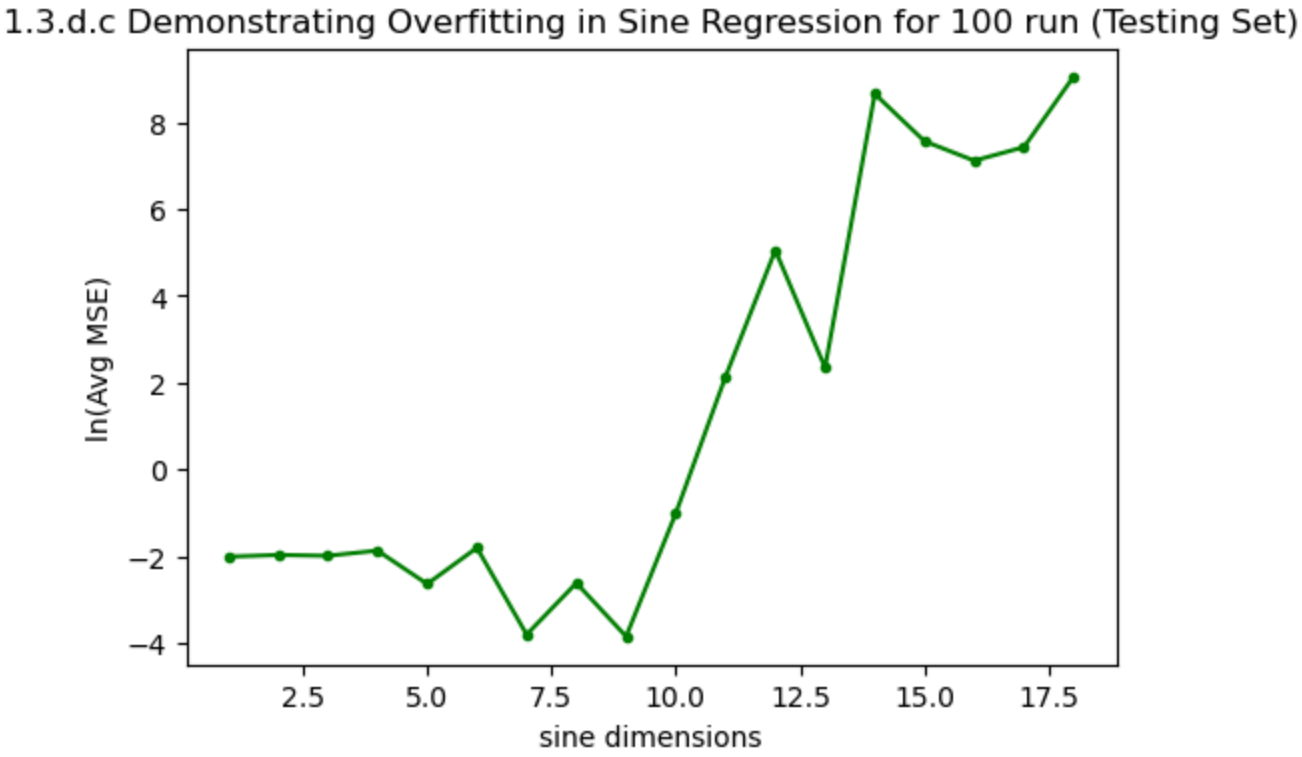
1.1.3

d-b


1.3.d.b Log of MSE versus the sine dimension for 100 runs (Training Set)
ln(Avg MSE)
−2
−3
−4
−5
−6
2.5
5.0
7.5
10.0
12.5
15.0
17.5
sine dimensions

1.1.3

d-c



1.2 Filtered Boston housing and kernels

(a)

```
 *Naive Regression*
Average Training MSE: 84.06, Average Test MSE: 85.35
```

(b)

Give a simple interpretation of the constant function in 'a.' above.

In the solution above, we used least squares regression as a constant function on vector of ones, this approach averages the housing price in the training dataset to predict the target. The constant function in this case provides a good mean prediction by minimizing the MSE. Mathematically, this can be represented as:

$$f(x_i) = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

(c)

Linear Regression with single attributes

```
Feature (CRIM): Average Training MSE: 69.36,  Average Test MSE: 76.84
Feature ( ZN ): Average Training MSE: 72.19,  Average Test MSE: 76.41
Feature (INDUS ): Average Training MSE: 66.27,  Average Test MSE: 61.79
Feature (CHAS): Average Training MSE: 82.54,  Average Test MSE: 81.18
Feature (NOX): Average Training MSE: 67.86,  Average Test MSE: 71.65
Feature (RM): Average Training MSE: 42.98,  Average Test MSE: 45.37
Feature (AGE): Average Training MSE: 71.37,  Average Test MSE: 74.91
Feature (DIS): Average Training MSE: 78.56,  Average Test MSE: 80.86
Feature (RAD): Average Training MSE: 71.40,  Average Test MSE: 74.05
Feature (TAX): Average Training MSE: 64.71,  Average Test MSE: 68.58
Feature (PTRATIO): Average Training MSE: 61.37,  Average Test MSE: 65.75
Feature (LSTAT): Average Training MSE: 38.20,  Average Test MSE: 39.26
```

(d)

```
Linear Regression using all attributes
Training MSE: 22.53, Average Test MSE: 23.69
```

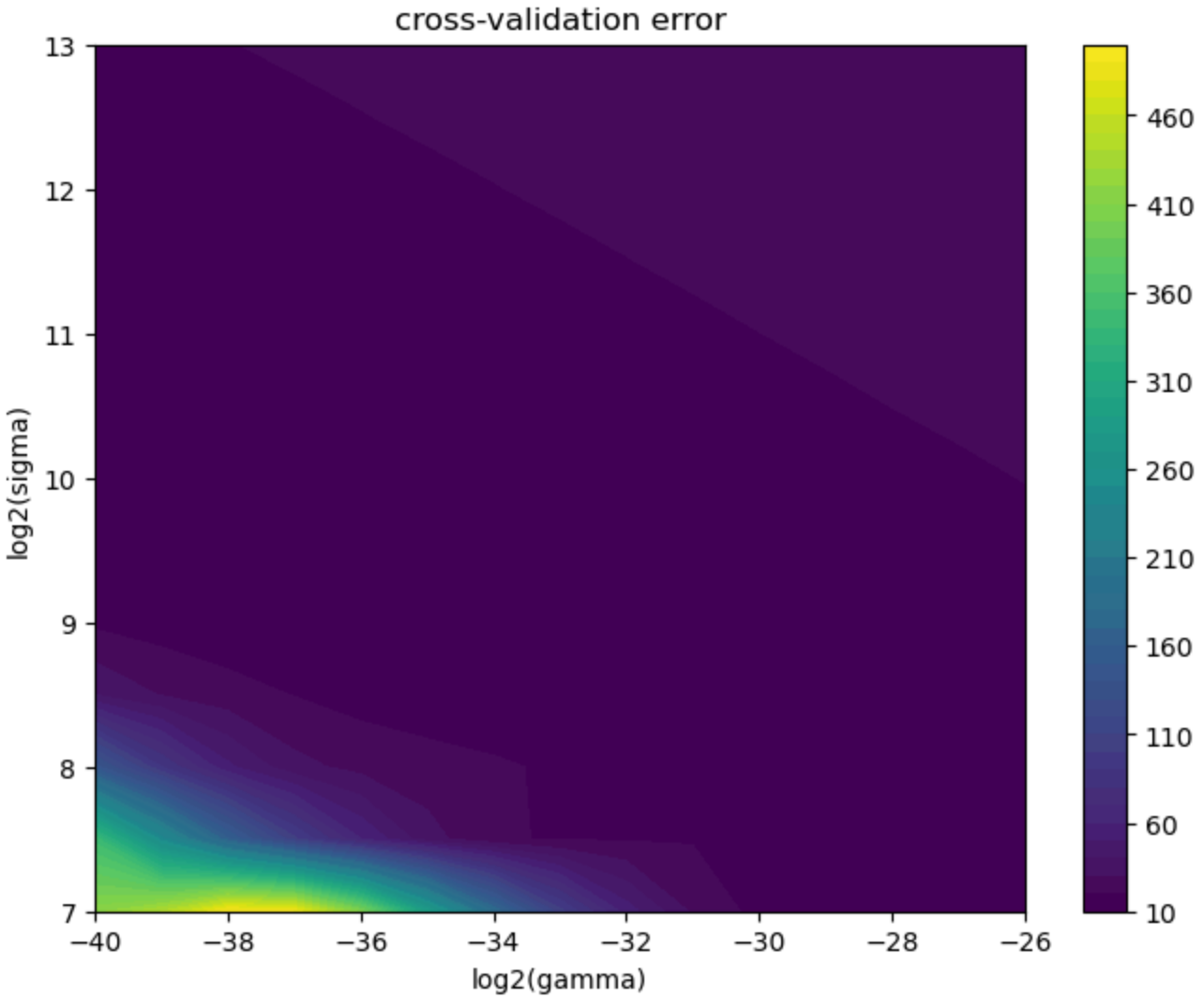
1.3 Kernelised ridge regression

(a), (c)

```
(a):  Best σ : -34.0,  Best γ: 9.0

(c): Training MSE: 4.540366776324811, Testing MSE: 15.743638334050754
```

(b)



(d)

Method	MSE train	MSE test
Naive Regression	85.30 ± 3.87	82.88 ± 7.64
Linear Regression (attribute 1)	71.67 ± 4.00	73.15 ± 8.91
Linear Regression (attribute 2)	74.38 ± 3.56	72.02 ± 7.03
Linear Regression (attribute 3)	63.01 ± 4.87	68.40 ± 9.63
Linear Regression (attribute 4)	82.07 ± 3.75	81.85 ± 7.71
Linear Regression (attribute 5)	69.60 ± 4.71	68.15 ± 9.37
Linear Regression (attribute 6)	44.18 ± 3.41	42.66 ± 6.80
Linear Regression (attribute 7)	72.47 ± 5.07	72.78 ± 10.19
Linear Regression (attribute 8)	78.88 ± 4.02	80.25 ± 8.00
Linear Regression (attribute 9)	71.12 ± 4.39	74.61 ± 8.94
Linear Regression (attribute 10)	66.77 ± 3.77	64.47 ± 7.51
Linear Regression (attribute 11)	62.97 ± 5.45	62.47 ± 11.05
Linear Regression (attribute 12)	38.86 ± 3.21	38.15 ± 6.20
Linear Regression (all attributes)	22.01 ± 0.00	24.59 ± 0.00
Kernel Ridge Regression	10.50 ± 0.96	14.14 ± 2.95

Part 2

Question 6

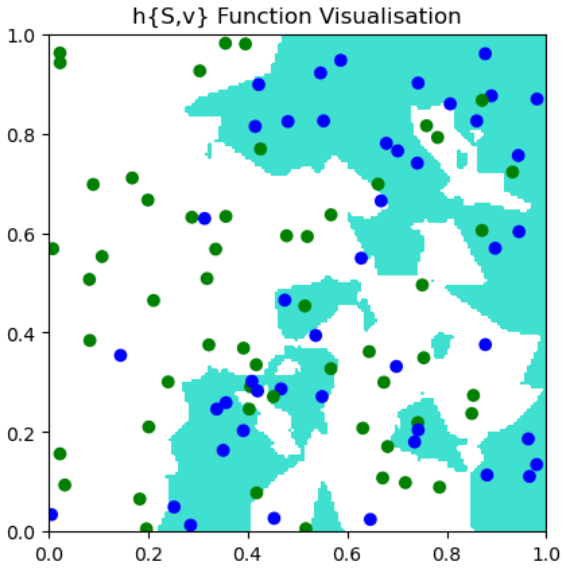


Figure 1: Visualisation of the $h_{S,3}$ ($v = 3$ and S is sampled uniformly at random)

Question 7

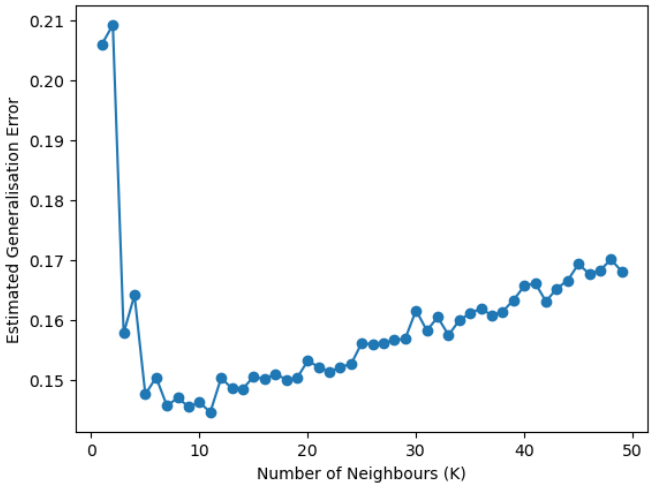


Figure 2: (Protocol A) Estimated Generalisation Error - Number of Neighbours (k) Plot

(b) From the visualisation of the function $h_{S,3}$, we see that $h_{S,3}$ splits the domain $[0,1]^2$ into regions labelled 0 or 1. So for every point $p \in [0,1]^2$ (with the exception of boundary points), there is a neighbourhood $N(p)$ containing points with the same label. As $h_{S,3}$ is used to generate the noisy datasets, given a sufficiently dense training dataset, the closest training points to a test point p_T will fall in $N(p_T)$ and their labels can be used to predict the test label through k-NN. Initially, as seen in the visualisation above, using a larger k decreases the estimated generalisation error (from approximately $k = 1$ to $k = 11$ with small fluctuations from $k = 7$ to k $= 11$) as this reduces the effect of noise in label prediction. But eventually k becomes too large and the labels of training

points outside the neighbourhood $N(p_T)$ are used leading to a steady rise in estimated generalisation error (from $k = 11$ onwards).

Question 8

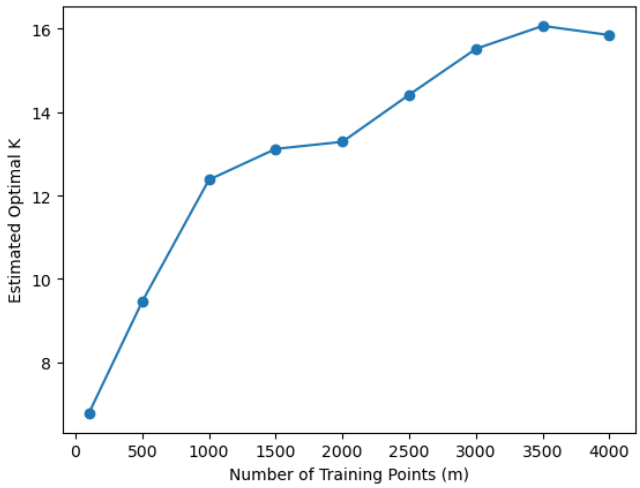


Figure 3: (Protocol B) Estimated Optimal Number of Neighbours (k) - Number of Training Points (m) Plot

(b) As explained in 7(b), every point $p \in [0,1]^2$ has a neighbourhood $N(p)$ with the same $h_{S,3}$ label. A larger number of training points (m) increases density of the training dataset on $[0,1]^2$, so for every test point p_T there are more training points in the region $N(p_T)$ and the effect of noise is decreased. This explains why, in the visualisation, the optimal value of k increases as m increases. However, as the test dataset is generated using the same method as the training dataset (noisy evaluation of $h_{S,3}$), higher values of k also fail to capture the noise in the test label (this may explain the decreasing rate of increase in the optimal k value).

Question 9

Kernel Modification Consider the function $K_c(\mathbf{x},\mathbf{z}) := c + \sum_{i=1}^{n} x_i z_i$ where $\mathbf{x},\mathbf{z} \in \mathbb{R}^n$.

(a) For what values of $c \in \mathbb{R}$ is K_c a positive semidefinite kernel? Give an argument supporting your claim.

Claim. K_c is a positive semidefinite kernel $\iff c \geqslant 0$.
proof. By definition, the kernel K_c is positive semidefinite (PSD) if and only if:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) \geqslant 0$$

holds for all $\mathbf{x_1}, ..., \mathbf{x_n} \in \mathbb{R}^n, a_1, ..., a_n \in \mathbb{R}$ and $n \in \mathbb{N}$. Using the fact that $K_c(\mathbf{x}, \mathbf{z}) := c + \sum_{i=1}^{n} x_i z_i = c + \mathbf{x^T z}$:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} c a_i a_j + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j}) \\ &= c(\sum_{i=1}^{n} a_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j})\end{aligned}$$

($\Longleftarrow$) Suppose that $c \geqslant 0$. Notice $\mathbf{x^T z} = < \mathbf{x}, \mathbf{z} >$ where $< -, - >$ is the standard inner product. Thus, using its bilinearity property:

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j < \mathbf{x_i}, \mathbf{x_j} > \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} < a_i \mathbf{x_i}, a_j \mathbf{x_j} > \\ &= < \sum_{i=1}^{n} a_i \mathbf{x_i}, \sum_{j=1}^{n} a_j \mathbf{x_j} > \\ &= < \mathbf{x}, \mathbf{x} >\end{aligned}$$

where $\mathbf{x} = \sum_{i=1}^{n} a_i \mathbf{x_i}$. As the standard inner product is positive definite, we must have that $< \mathbf{z}, \mathbf{z} > \geqslant 0$ for all $\mathbf{z} \in \mathbb{R}^n$ $\therefore$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j}) \geqslant 0 \tag{1}$$

Now, consider $c(\sum_{i=1}^{n} a_i)^2$. Note that $c \geqslant 0$ and, for all real numbers $a \in \mathbb{R}$, $a^2 \geqslant 0$ $\therefore ca^2 \geqslant 0$. As $\sum_{i=1}^{n} a_i \in \mathbb{R}$:

$$c(\sum_{i=1}^{n} a_i)^2 \geqslant 0 \tag{2}$$

Using (1) and (2), we get:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) = c(\sum_{i=1}^{n} a_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j}) \geqslant 0$$

Therefore, $c \geqslant 0 \implies K_c$ is PSD.

($\implies$) Suppose that K_c is PSD. Then, we must have that:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) \geqslant 0$$

holds for all $\mathbf{x_1}, ..., \mathbf{x_n} \in \mathbb{R}^n, a_1, ..., a_n \in \mathbb{R}$ and $n \in \mathbb{N}$. Assume $c < 0$, then if we choose:

$$\mathbf{x_i} = \sqrt{\frac{c'}{n}} \cdot \mathbf{1} \quad \text{for all } i \in \{1, ..., n\}$$

where $c' < |c| = -c$ and $\mathbf{1} = (1, ..., 1)^T \in \mathbb{R}^n$. Then,

$$\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) &= c(\sum_{i=1}^{n} a_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (\mathbf{x_i^T x_j}) \\
&= c(\sum_{i=1}^{n} a_i)^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (c') \\
&= c(\sum_{i=1}^{n} a_i)^2 + c'(\sum_{i=1}^{n} a_i)^2 \\
&= (c + c')(\sum_{i=1}^{n} a_i)^2
\end{aligned}$$

Since $c' < -c$, we must have $(c + c') < 0$. Also, as for all real numbers $a_1, ..., a_n \in \mathbb{R}$, $(\sum_{i=1}^{n} a_i)^2 \geqslant 0$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) = (c + c')(\sum_{i=1}^{n} a_i)^2 < 0$$

This is a contradiction, as K_c is PSD ($\because \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j K_c(\mathbf{x_i}, \mathbf{x_j}) \geqslant 0$). Thus the assumption that $c < 0$ must be false ($c \geqslant 0$ must hold). Therefore, K_c is PSD $\implies c \geqslant 0$.

■

(b) Suppose we use K_c as a kernel function with linear regression (least squares). Explain how c influences the solution.

Consider the feature map $\phi : (x_1, ..., x_n)^T \mapsto (x_1, ..., x_n, \sqrt{c})^T$ (only possible as $c \geqslant 0$). Then, given $\mathbf{x}, \mathbf{z} \in \mathbb{R}^n$ where $\mathbf{x} = (x_1, ..., x_n)^T$ and $\mathbf{z} = (z_1, ..., z_n)^T$:

$$\begin{aligned}
\phi(\mathbf{x})^T \phi(\mathbf{z}) &= \begin{bmatrix} x_1 & ... & x_n & \sqrt{c} \end{bmatrix} \begin{bmatrix} z_1 \\ \vdots \\ z_n \\ \sqrt{c} \end{bmatrix} \\
&= c + \sum_{i=1}^{n} x_i z_i \\
&= K_c(\mathbf{x}, \mathbf{z})
\end{aligned}$$

Thus, ϕ is a feature map for the kernel K_c. Note that when $c = 0$, then ϕ is the identity map and linear regression with K_c is equivalent to OLS linear regression without a bias term. However when $c > 0$, ϕ adds a constant bias component to the input vector. Then linear regression with K_c is equivalent to OLS linear regression with a bias term (the actual value of the bias component $\sqrt{c}$ is irrelevant since the bias weight can be scaled to keep the bias term the same).

In conclusion, linear regression with K_c is OLS linear regression with a bias term if $c > 0$ and without a bias term if $c = 0$.

Question 10

Suppose we perform linear regression (OLS) with a Gaussian kernel $K_\beta(\mathbf{x}, \mathbf{t}) = \exp(-\beta||\mathbf{x} - \mathbf{t}||^2)$ to train a classifier on a dataset $(\mathbf{x_1}, y_1), ..., (\mathbf{x_m}, y_m) \in \mathbb{R}^n \times \{-1, 1\}$. Thus obtaining a function $f : \mathbb{R}^n \to \mathbb{R}$ which is of the form $f(\mathbf{t}) = \sum_{i=1}^m \alpha_i K_\beta(\mathbf{x_i}, \mathbf{t})$. The corresponding classifier is then $\text{sign}(f(\mathbf{t}))$. This classifier depends on the parameter β selected for the kernel. In what scenario will chosen β enable the trained linear classifier to simulate a 1-Nearest Neighbour classifier trained on the same dataset?

Assumptions. (The following assumptions are important for the proof to work!)

- For a given value of $\mathbf{t} \in \mathbb{R}^n$, we will assume that there is a unique nearest point in the dataset. Formally, let $L = \arg\min_{i \in \{1,...,m\}} ||\mathbf{x_i} - \mathbf{t}||^2$ then for all $j \in \{1, ..., m\} \setminus \{L\} : ||\mathbf{x_L} - \mathbf{t}||^2 < ||\mathbf{x_j} - \mathbf{t}||^2$. This assumption is sensible as a 1-Nearest Neighbour classifier is undefined if the nearest point (in the training dataset) is not unique and has different labels.
- Assume that all points in the dataset are unique, for all $i, j \in \{1, ..., m\} : \mathbf{x_i} \neq \mathbf{x_j}$. This is a less sensible assumption as this is a possibility in normal k-NN, but is necessary for inverse of the Gaussian kernel Gram Matrix to exist (for large values of β).

From the question, we know/define the following:

- $K_\beta(\mathbf{x}, \mathbf{t}) = \exp(-\beta||\mathbf{x} - \mathbf{t}||^2) \therefore$ for all $i \in \{1, ..., m\}$: $K_\beta(\mathbf{x_i}, \mathbf{t}) = \exp(-\beta||\mathbf{x_i} - \mathbf{t}||^2)$ where $\beta > 0$
- K_β is the Gram Matrix of the Gaussian kernel using the dataset $\{\mathbf{x_1}, ..., \mathbf{x_m}\}$ (**Note.** K_β without brackets and inputs denotes the Gram Matrix).
- Given $\alpha = (\alpha_1, ..., \alpha_m)^T$, then $\alpha = K_\beta^{-1}\mathbf{y}$ where $\mathbf{y} = (y_1, ..., y_m)^T$.
- For any $\mathbf{t} \in \mathbb{R}^n$, the classifier is $\text{sign}(f(\mathbf{t}))$, where:

$$f(\mathbf{t}) = \sum_{i=1}^m \alpha_i K_\beta(\mathbf{x_i}, \mathbf{t}) \tag{1}$$

Consider K_β (the Gram Matrix of the Gaussian kernel):

$$(K_\beta)_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \exp(-\beta||\mathbf{x_i} - \mathbf{x_j}||^2) & \text{otherwise} \end{cases}$$

Since $\mathbf{x_i} \neq \mathbf{x_j}$ (see assumptions), $||\mathbf{x_i} - \mathbf{x_j}|| > 0 \therefore$

$$\lim_{\beta\to\infty} (K_\beta)_{i,j} = \lim_{\beta\to\infty} \exp(-\beta||\mathbf{x_i} - \mathbf{x_j}||^2) = 0 \quad \text{for } i \neq j \quad \textbf{consequently} \quad \lim_{\beta\to\infty} (K_\beta)_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

Define $\tilde{K}_\beta = K_\beta - I$ ($\therefore K_\beta = \tilde{K}_\beta + I$), then $\lim_{\beta\to\infty} \tilde{K}_\beta = 0$. Formally, we can express this as:

$$\forall \epsilon > 0 \;\; \exists M \in \mathbb{R}_+ \;\; \forall \beta > M : \;\; ||\tilde{K}_\beta||_{op} < \epsilon$$

Let $\beta_0 \in \mathbb{R}_+$ be a real number such that $\forall \beta > \beta_0 : ||\tilde{K}_\beta||_{op} < 1$. Using $K_\beta = \tilde{K}_\beta + I$ and Neumann series,

$$K_\beta^{-1} = (\tilde{K}_\beta + I)^{-1} = \sum_{n=0}^{\infty} (-1)^n (\tilde{K}_\beta)^n$$

This is guaranteed to converge when $||\tilde{K}_\beta||_{op} < 1$ (can be enforced by asserting $\beta > \beta_0$). Also, when $||\tilde{K}_\beta||_{op} < 1$, we can express the result above as:

$$K_\beta^{-1} = I + O(||\tilde{K}_\beta||_{op}) \textbf{ and so } \alpha = (I + O(||\tilde{K}_\beta||_{op}))\mathbf{y} \tag{2}$$

Consequently,

$$\lim_{\beta\to\infty} \tilde{K}_\beta = 0 \iff \lim_{\beta\to\infty} ||\tilde{K}_\beta|| = 0 \implies \lim_{\beta\to\infty} K_\beta^{-1} = I$$

Now consider the classifier $\text{sign}(f(\mathbf{t}))$. Using the fact that $\exp(x) > 0$ for all $x \in \mathbb{R}$:

$$\text{sign}(f(\mathbf{t})) = \text{sign}(f(\mathbf{t}) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2)) \tag{3}$$

where $L = \arg\min_{i\in\{1,...,m\}} ||\mathbf{x_i} - \mathbf{t}||^2$. Using (1) and definition of $K_\beta(\mathbf{x}, \mathbf{t})$:

$$\begin{aligned} f(\mathbf{t}) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) &= \sum_{i=1}^{m} \alpha_i K_\beta(\mathbf{x_i}, \mathbf{t}) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) \\ &= \sum_{i=1}^{m} \alpha_i \exp(-\beta||\mathbf{x_i} - \mathbf{t}||^2) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) \\ &= \sum_{i=1}^{m} \alpha_i \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \end{aligned}$$

Using (2), we know $\alpha_i = y_i + O(||\tilde{K}_\beta||_{op}) \therefore$

$$\begin{aligned} f(\mathbf{t}) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) = &\sum_{i=1}^{m} y_i \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \\ &+ O(||\tilde{K}_\beta||_{op}) \sum_{i=1}^{m} \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \end{aligned}$$

Note that for all $i \in \{1, ..., m\}$: $||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2 \geqslant 0$ (by definition of L) $\therefore$ as $\beta > 0$, $\exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \leqslant 1$. So $\sum_{i=1}^{m} \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2) \leqslant m \therefore$

$$\begin{aligned} f(\mathbf{t}) \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) &= \sum_{i=1}^{m} y_i \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) + O(||\tilde{K}_\beta||_{op}) \\ &= y_L + \sum_{i=1, i\neq L}^{m} y_i \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) + O(||\tilde{K}_\beta||_{op}) \end{aligned} \tag{4}$$

Using the assumptions, we know for all $i \in \{1, ..., m\}\backslash\{L\}$: $||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2 \geqslant \min_{j\in\{1,...,m\}\backslash\{L\}} ||\mathbf{x_j} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2 > 0$. As $\beta > 0$, for $i \in \{1, ..., m\} \setminus \{L\}$:

$$\exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \leqslant \exp(-\beta \cdot \min_{j\in\{1,...,m\}\backslash\{L\}} (||\mathbf{x_j} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) \tag{5}$$

and

$$\lim_{\beta\to\infty} \exp(-\beta \cdot \min_{j\in\{1,...,m\}\backslash\{L\}} (||\mathbf{x_j} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) = 0$$

Formally, this limit can be expressed as:

$$\forall\epsilon > 0 \;\; \exists M \in \mathbb{R}_+ \;\; \forall\beta > M : \; |\exp(-\beta \cdot \min_{j\in\{1,...,m\}\backslash\{L\}} (||\mathbf{x_j} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2))| < \epsilon$$

Let $\beta_1 \in \mathbb{R}_+$ be a real number such that $\forall\beta > \beta_1$: $|\exp(-\beta \cdot \min_{j\in\{1,...,m\}\backslash\{L\}} (||\mathbf{x_j} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2))| < \frac{1}{m}$. Thus, using this result, (5) and the fact that for all $j \in \{1, ..., m\}$ $-1 \leqslant y_j \leqslant 1$:

$$-\frac{1}{m} < y_i \exp(-\beta(||\mathbf{x_i} - \mathbf{t}||^2 - ||\mathbf{x_L} - \mathbf{t}||^2)) < \frac{1}{m} \qquad \text{for } \beta > \beta_1 \textbf{ and } i \in \{1, ..., m\} \setminus \{L\} \tag{6}$$

Note that $\lim_{\beta\to\infty} ||\tilde{K}_\beta||_{op} = 0 \therefore \lim_{\beta\to\infty} O(||\tilde{K}_\beta||_{op}) = 0$. By a similar argument as above, there is a real number $\beta_2 \in \mathbb{R}_+$ such that:

$$-\frac{1}{m} < O(||\tilde{K}_\beta||_{op}) < \frac{1}{m} \qquad \text{for } \beta > \beta_2 \tag{7}$$

Let $\beta' = max\{\beta_0, \beta_1, \beta_2\}$. Using (4), (6) and (7):

$$y_L - 1 < f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) < y_L + 1 \qquad \text{for } \beta > \beta'$$

Notice, that (for $\beta > \beta'$):

$$\begin{aligned} y_L = -1 &\implies f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) < 0 \implies \text{sign}(f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2)) = -1 \\ y_L = 1 &\implies f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2) > 0 \implies \text{sign}(f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2)) = 1 \end{aligned}$$

Thus, for $\beta > \beta'$, $\text{sign}(f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2)) = y_L$. Using (3), we obtain the result:

$$\text{sign}(f(\mathbf{t})) = \text{sign}(f(\mathbf{t})\ \exp(\beta||\mathbf{x_L} - \mathbf{t}||^2)) = y_L \qquad \text{for } \beta > \beta'$$

where y_L is the label of the closest point (in the dataset) $\mathbf{x_L}$ to $\mathbf{t}$. This is equivalent to a 1-Nearest Neighbour classifier!

Note. In the argument above, $\beta_0, \beta_1, \beta_2$ and (by consequence) β' are values dependent on $\mathbf{x_1}, ..., \mathbf{x_m}, \mathbf{t}$. Since, we 'fixed' these values at the beginning, there was no need to mention this. However, for arbitrary values of $\mathbf{x_1}, ..., \mathbf{x_m}, \mathbf{t}$, it is important to note that the threshold value β' is actually a function $(\beta'(\mathbf{x_1}, ..., \mathbf{x_m}, \mathbf{t}))$.

Conclusion. Given a sufficiently large β (β is greater than the threshold $\beta'(\mathbf{x_1}, ..., \mathbf{x_m}, \mathbf{t})$) and the assumptions detailed at the start, the classifier simulates a 1-Nearest Neighbour classifier for $\mathbf{t} \in \mathbb{R}^n$.

■

Question 11

(a) Let $C \subset X$ be a subset of X with cardinality $|C| = 2n$ for some $n \in \mathbb{N}$. Denote with $Y^C = \{f_1, ..., f_T\}$ the set of all possible function $f : C \to Y$ ($\therefore T = |Y^C| = 2^{2n}$). For any $i \in \{1, ..., T\}$, denote with ρ_i the distribution over $C \times Y$ such that:

$$\rho_i(\{(x, y)\}) = \begin{cases} \frac{1}{2n} & \text{if } y = f_i(x) \\ 0 & \text{otherwise} \end{cases} \quad \forall x \in X, y \in Y$$

Show that $\mathcal{E}_{\rho_i}(f_i) = \inf_{f:X\to Y} \mathcal{E}_{\rho_i}(f) = 0$.

proof. By definition of $\mathcal{E}_\rho(f)$ for a function $f : X \to Y$ and distribution ρ, given in the coursework document, we know that

$$\mathcal{E}_{\rho_i}(f_i) = \int_{X\times Y} \mathbf{1}_{\{f_i(x)\neq y\}}\, d\rho_i(x, y) = \rho_i(\{(x, y) \in X \times Y : f_i(x) \neq y\})$$

Consider the set $\{(x, y) \in X \times Y : f_i(x) = y, x \in C\} = \{(x, f_i(x)) : x \in C\}$. Since C is countable/finite, using the countable additivity property of probabilities:

$$\rho_i(\{(x, f_i(x)) : x \in C\}) = \sum_{x\in C} \rho_i(\{(x, f_i(x))\})$$

By definition of ρ_i, $\rho_i(\{(x, f_i(x))\}) = \frac{1}{2n}$ for all $x \in C$ $\therefore$

$$\rho_i(\{(x, f_i(x)) : x \in C\}) = \sum_{x\in C} \frac{1}{2n} = \frac{|C|}{2n}$$

As $|C| = 2n$, we conclude $\rho_i(\{(x, y) \in X \times Y : f_i(x) = y, x \in C\}) = \rho_i(\{(x, f_i(x)) : x \in C\}) = 1$. By the complement rule for probabilities, $\rho_i(\{(x, y) \in X \times Y : f_i(x) = y, x \in C\}^c) = 0$. As $\{(x, y) \in X \times Y : f_i(x) \neq y\} \subset \{(x, y) \in X \times Y : f_i(x) = y, x \in C\}^c$, $\rho_i(\{(x, y) \in X \times Y : f_i(x) \neq y\}) = 0$. Thus,

$$\mathcal{E}_{\rho_i}(f_i) = \int_{X\times Y} \mathbf{1}_{\{f_i(x)\neq y\}}\, d\rho_i(x, y) = 0 \tag{1}$$

Now notice that, for any $f : X \to Y$ and $(x, y) \in (X \times Y)$, $\mathbf{1}_{\{f(x)\neq y\}} \geqslant 0$. Since integrals (Lebesgue and Riemann) preserve weak inequalities, we know:

$$\mathcal{E}_{\rho_i}(f) = \int_{X\times Y} \mathbf{1}_{\{f(x)\neq y\}}\, d\rho_i(x, y) \geqslant \int_{X\times Y} 0\, d\rho_i(x, y) = 0 \tag{2}$$

From the definition of infimums, we must have $\inf_{f:X\to Y}\mathcal{E}_{\rho_i}(f) \leqslant \mathcal{E}_{\rho_i}(f_i)$ $\therefore$ by using (1), $\inf_{f:X\to Y}\mathcal{E}_{\rho_i}(f) \leqslant 0$. As infimums preserve weak inequalities, by using (2), we know $0 \leqslant \inf_{f:X\to Y}\mathcal{E}_{\rho_i}(f)$. Thus, $\inf_{f:X\to Y}\mathcal{E}_{\rho_i}(f) = 0$.

By the argument above, we have shown: $\mathcal{E}_{\rho_i}(f_i) = \inf_{f:X\to Y}\mathcal{E}_{\rho_i}(f) = 0$.

■

Alternative proof (for $\mathcal{E}_{\rho_i}(f_i) = 0$)**.** Note that, by its definition, ρ_i is a discrete probability measure on the space $\{(x, f_i(x)) : x \in C\}$. This is because all singletons are measurable ($\rho_i(\{(x, f_i(x))\}) = \frac{1}{2n}$), $\rho_i(\{(x, f_i(x)) : x \in C\}) = 1$ (shown in the main proof) and C is countable/finite. Thus,

$$\mathcal{E}_{\rho_i}(f) = \int_{X\times Y} \mathbf{1}_{\{f(x)\neq y\}}\, d\rho_i(x, y) = \sum_{x\in C} \mathbf{1}_{\{f(x)\neq f_i(x)\}} \cdot \rho_i(\{(x, f_i(x))\})$$

Using the fact that $\rho_i(\{(x, f_i(x))\}) = \frac{1}{2n}$ for all $x \in C$:

$$\mathcal{E}_{\rho_i}(f) = \frac{1}{2n} \sum_{x \in C} \mathbf{1}_{\{f(x) \neq f_i(x)\}} \tag{3}$$

Consequently,

$$\mathcal{E}_{\rho_i}(f_i) = \frac{1}{2n} \sum_{x \in C} \mathbf{1}_{\{f_i(x) \neq f_i(x)\}} = 0$$

■

(b) Let C^n be the set of all possible *input* datasets (of size n) of points in C. Then $C^n = \{S_1, ..., S_k\}$ with $k = |C^n| = (2n)^n$. For every $j \in \{1, ..., k\}$ and input set $S_j = (x_1, ..., x_n) \in C^n$, denote $S_j^i = \{(x_1, f_i(x_1)), ..., (x_n, f_i(x_n))\}$ the corresponding input-output dataset for every $i \in \{1, ..., T\}$. Show that

$$\max_{i=1,...,T} \mathbb{E}_{S \sim \rho_i^n} \mathcal{E}_{\rho_i}(A(S)) \geqslant \min_{j=1,...,k} \frac{1}{T} \sum_{i=1}^{T} \mathcal{E}_{\rho_i}(A(S_j^i))$$

proof. Consider an input-output dataset $(x_m, y_m)_{m=1}^n$. In the coursework document, we are told that each pair (x_m, y_m) is independently sampled from any $\rho \therefore$

$$\rho_i^n(\{(x_m, y_m)_{m=1}^n\}) = \prod_{m=1}^{n} \rho_i(\{(x_m, y_m)\})$$

As $\rho_i(\{(x_m, y_m)\}) = \frac{1}{2n}$ when $y_m = f_i(x_m)$ and $\rho_i(\{(x_m, y_m)\}) = 0$ otherwise,

$$\rho_i^n(\{(x_m, y_m)_{m=1}^n\}) = \begin{cases} \frac{1}{(2n)^n} & \text{if } y_m = f_i(x_m) \text{ for all } m \in \{1, ..., n\} \\ 0 & \text{otherwise} \end{cases}$$

Note that $C^n = \{S_1, ..., S_k\}$ (where $k = (2n)^n$) $\therefore$ for any $(x_m)_{m=1}^n \in C^n$ there exists some $j \in \{1, ..., k\}$ such that $S_j = (x_m)_{m=1}^n$. Then, by definition, $S_j^i = (x_m, f_i(x_m))_{m=1}^n$ and so the result above is equivalent to

$$\rho_i^n(\{(x_m, y_m)_{m=1}^n\}) = \begin{cases} \frac{1}{k} & \text{if } S_j^i = (x_m, y_m)_{m=1}^n \text{ for some } j \in \{1, ..., k\} \\ 0 & \text{otherwise} \end{cases}$$

Thus, for distribution ρ_i^n, the possible input-output datasets are $S_1^i, ..., S_k^i$ (with equal probability $\frac{1}{k}$) $\therefore$

$$\begin{aligned} \mathbb{E}_{S \sim \rho_i^n} \mathcal{E}_{\rho_i}(A(S)) &= \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i)) \rho_i^n(\{S_j^i\}) \\ &= \frac{1}{k} \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i)) \end{aligned} \tag{4}$$

Now, consider the sum $\frac{1}{kT} \sum_{i=1}^{T} \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i))$. Using the fact that for any set of scalars $\{\alpha_1, ..., \alpha_L\}$: $\min_l \alpha_l \leqslant \frac{1}{L} \sum_{l=1}^{L} \alpha_l \leqslant \max_l \alpha_l$, we obtain the following inequalities:

$$\frac{1}{kT} \sum_{i=1}^{T} \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i)) = \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i)) \leqslant \max_{i=1,...,T} \frac{1}{k} \sum_{j=1}^{k} \mathcal{E}_{\rho_i}(A(S_j^i))$$

$$\frac{1}{kT}\sum_{i=1}^{T}\sum_{j=1}^{k}\mathcal{E}_{\rho_i}(A(S_j^i)) = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \min_{j=1,...,k}\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i))$$

Thus,

$$\max_{i=1,...,T}\frac{1}{k}\sum_{j=1}^{k}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \min_{j=1,...,k}\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i))$$

Using (4),

$$\max_{i=1,...,T}\mathbb{E}_{S\sim\rho_i^n}\,\mathcal{E}_{\rho_i}(A(S)) \geqslant \min_{j=1,...,k}\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i))$$

■

(c) Fix $j \in \{1, ..., k\}$ with $S_j = \{x_1, ..., x_n\}$ and denote $R_j = \{v_1, ..., v_p\}$ the subset of points of C that do not belong to S_j. Show that

$$\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{2}\,\min_{v\in R_j}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}}$$

proof. Using equation (3) [see (a) - Alternative proof], we know

$$\mathcal{E}_{\rho_i}(A(S_j^i)) = \frac{1}{2n}\sum_{x\in C}\mathbf{1}_{\{A(S_j^i)(x)\neq f_i(x)\}}$$

Given that $R_j \subset C$ ($R_j = C \setminus S_j$) and $\mathbf{1}_{\{A(S_j^i)(x)\neq f_i(x)\}} \geqslant 0$ for all $x \in C$,

$$\mathcal{E}_{\rho_i}(A(S_j^i)) = \frac{1}{2n}\sum_{x\in C}\mathbf{1}_{\{A(S_j^i)(x)\neq f_i(x)\}} \geqslant \frac{1}{2n}\sum_{v\in R_j}\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}}$$

Notice that $S_j = \{x_1, ..., x_n\} \subset C \therefore |S_j| \leqslant n$ ($x_1, ..., x_n$ are not necessarily distinct) and $|R_j| = |C \setminus S_j| = |C| - |S_j|$. Combining these results, we get $|C| - n \leqslant |C| - |S_j| = |R_j|$. As $|C| = 2n$ and $|R_j| = p$ ($R_j = \{v_1, ..., v_p\}$ which are distinct), $n \leqslant p \therefore$

$$\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{2p}\sum_{v\in R_j}\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}}$$

This can be rewritten as

$$\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{2p}\sum_{m=1}^{p}\mathbf{1}_{\{A(S_j^i)(v_m)\neq f_i(v_m)\}}$$

Thus,

$$\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{m=1}^{p}\mathbf{1}_{\{A(S_j^i)(v_m)\neq f_i(v_m)\}} = \frac{1}{2}\cdot\frac{1}{p}\sum_{m=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v_m)\neq f_i(v_m)\}}$$

Using the fact, that for any set of scalars $\{\alpha_1, ..., \alpha_L\}$: $\min_l \alpha_l \leqslant \frac{1}{L}\sum_{l=1}^{L}\alpha_l$,

$$\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{2}\cdot\frac{1}{p}\sum_{m=1}^{p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v_m)\neq f_i(v_m)\}} \geqslant \frac{1}{2}\cdot\min_{m=1,...,p}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v_m)\neq f_i(v_m)\}}$$

Note that this inequality can be written as

$$\frac{1}{T}\sum_{i=1}^{T}\mathcal{E}_{\rho_i}(A(S_j^i)) \geqslant \frac{1}{2}\cdot\min_{v\in R_j}\frac{1}{T}\sum_{i=1}^{T}\mathbf{1}_{\{A(S_j^i)(v)\neq f_i(v)\}}$$

■

(d) With the same assumption as above, show that for any $v \in R_j$

$$\frac{1}{T}\sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} = \frac{1}{2}$$

proof. Let $v \in R_j$. For any function $f_i \in Y^C$, we define a function $f_{i'} : C \to Y$ as

$$f_{i'}(x) = \begin{cases} f_i(x) & \text{if } x \in C \setminus \{v\} \\ -f_i(x) & \text{if } x = v \end{cases}$$

As $Y = \{-1, 1\}$, $f_i(x) \neq f_{i'}(x)$ if and only if $x = v$.

Notice that, $f_{i'} \in Y^C$ (as Y^C is the set of all functions $f : C \to Y$) $\therefore F = \{\{f_i, f_{i'}\} : i = 1, ..., T\}$ is a covering of Y^C. Also,

$$f_{(i')'}(x) = \begin{cases} f_{i'}(x) = f_i(x) & \text{if } x \in C \setminus \{v\} \\ -f_{i'}(x) = f_i(x) & \text{if } x = v \end{cases}$$

$$\therefore \quad f_{(i')'} = f_i \tag{5}$$

This shows that $\{f_i, f_{i'}\} = \{f_{i'}, f_{(i')'}\}$ and can be used to show that F is pairwise disjoint.

Let $\{f_i, f_{i'}\}, \{f_j, f_{j'}\}$ be distinct pairs in F and assume that $\{f_i, f_{i'}\} \cap \{f_j, f_{j'}\} \neq \emptyset$. *WLOG* we can assume that $f_i = f_j$ (as, by (5), $\{f_i, f_{i'}\} = \{f_{i'}, f_{(i')'}\}$ we can swap $f_i \leftrightarrow f_{i'}$ and/or $f_j \leftrightarrow f_{j'}$ to assert that this condition holds) $\therefore f_{i'} = f_{j'}$. But this means that $\{f_i, f_{i'}\} = \{f_j, f_{j'}\}$ (contradiction). Therefore, $\{f_i, f_{i'}\} \cap \{f_j, f_{j'}\} = \emptyset$ $\therefore F$ is pairwise disjoint.

As $F = \{\{f_i, f_{i'}\} : i = 1, ..., T\}$ is a pairwise disjoint covering of Y^C, F is a partition of Y^C. As each element in F is a pair and $|Y^C| = T$, $|F| = \frac{T}{2}$.

Consider $\mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} + \mathbf{1}_{\{A(S_j^{i'})(v) \neq f_{i'}(v)\}}$. As $v \in R_j$ ($\therefore v \notin S_j$), $Y = \{-1, 1\}$ and $f_i(x) \neq f_{i'}(x)$ if and only if $x = v$:

- $S_j^i = S_j^{i'}$ $\therefore A(S_j^i) = A(S_j^{i'})$
- Either $(f_i(v), f_{i'}(v)) = (-1, 1)$ or $(1, -1)$

Thus,

$$\mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} + \mathbf{1}_{\{A(S_j^{i'})(v) \neq f_{i'}(v)\}} = \mathbf{1}_{\{A(S_j^i)(v) \neq 1\}} + \mathbf{1}_{\{A(S_j^i)(v) \neq -1\}}$$

As $A(S_j^i)(v) = -1$ or 1,
$$\mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} + \mathbf{1}_{\{A(S_j^{i'})(v) \neq f_{i'}(v)\}} = 1 \tag{6}$$

Now consider the sum $\sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}}$. As $Y^C = \{f_1, ..., f_T\}$ and $F = \{\{f_i, f_{i'}\} : i = 1, ..., T\}$ is a partition of Y^C, we have

$$\sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v)) \neq f_i(v)\}} = \sum_{\{f_i, f_{i'}\} \in F} [\mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} + \mathbf{1}_{\{A(S_j^{i'})(v) \neq f_{i'}(v)\}}]$$

Using (6), we get

$$\sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} = \sum_{\{f_i, f_{i'}\} \in F} 1$$

As $|F| = \frac{T}{2}$,

$$\sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} = \frac{T}{2}$$

$$\therefore \quad \frac{1}{T} \sum_{i=1}^{T} \mathbf{1}_{\{A(S_j^i)(v) \neq f_i(v)\}} = \frac{1}{2}$$

■

(e) We will prove an auxiliary result: let Z be a random variable with values in $[0, 1]$ and let $\mathbb{E}[Z] = \mu$. Show that for any $a \in (0, 1)$

$$\mathbb{P}(Z > 1 - a) \geqslant \frac{\mu - (1 - a)}{a}$$

proof. Note that $\mathbb{P}(Z > 1 - a) = 1 - \mathbb{P}(Z \leqslant 1 - a)$ and $\mathbb{P}(Z \leqslant 1 - a) = \mathbb{P}(1 - Z \geqslant a)$. Thus,

$$\mathbb{P}(Z > 1 - a) = 1 - \mathbb{P}(1 - Z \geqslant a) \tag{7}$$

Let the random variable $W = 1 - Z$. Then

- As Z takes values in [0,1], W also takes values in [0,1] ($\therefore W$ is non-negative)
- $\mathbb{E}[W] = \mathbb{E}[1 - Z] = 1 - \mathbb{E}[Z] = 1 - \mu$

As W is non-negative, $a > 0$ and $\mathbb{E}[W] = 1 - \mu$, we can use Markov's Inequality to obtain

$$\begin{aligned} \mathbb{P}(W \geqslant a) &\leqslant \frac{\mathbb{E}[W]}{a} \\ &\leqslant \frac{1 - \mu}{a} \end{aligned}$$

As $W = 1 - Z$, we know $\mathbb{P}(1 - Z \geqslant a) \leqslant \frac{1-\mu}{a}$ and, by using (7),

$$\mathbb{P}(Z > 1 - a) \geqslant 1 - \frac{1 - \mu}{a}$$

$$\therefore \quad \mathbb{P}(Z > 1 - a) \geqslant \frac{\mu - (1 - a)}{a}$$

■

(f) Combine the results from the previous exercises to show that for any algorithm A and for any integer n there exists a distribution ρ over $X \times Y$ such that

$$\mathbb{P}_{S \sim \rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{1}{7}$$

proof. Note that $0 \leqslant \mathbf{1}_{\{f(x) \neq y\}} \leqslant 1$ for all $(x, y) \in X \times Y$. As integrals preserve weak inequalities, for any ρ and $f : X \to Y$,

$$0 = \int_{X \times Y} 0 \; dp(x, y) \leqslant \int_{X \times Y} \mathbf{1}_{\{f(x) \neq y\}} \; dp(x, y) \leqslant \int_{X \times Y} 1 \; dp(x, y) = 1$$

$$\therefore \quad 0 \leqslant \mathcal{E}_p(f) \leqslant 1$$

Then, for an arbitrary distribution ρ, the random variable $\mathcal{E}_p(A(S))$ takes values in $[0, 1]$. By using the result in (e), letting $Z = \mathcal{E}_p(A(S))$ and $a = \frac{7}{8}$, we get

$$\mathbb{P}_{S \sim \rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{\mathbb{E}_{S \sim \rho^n} \; \mathcal{E}_p(A(S)) - \frac{1}{8}}{\frac{7}{8}}$$

By combining the results from (b), (c) and (d):

$$\max_{i=1,\dots,T} \mathbb{E}_{S\sim\rho_i^n}\ \mathcal{E}_{\rho_i}(A(S)) \geqslant \frac{1}{4}$$

Let $\rho = \rho_I$, where $I = \text{argmax}_{i=1,\dots,T}\ \mathbb{E}_{S\sim\rho_i^n}\ \mathcal{E}_{\rho_i}(A(S))$, then

$$\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{\mathbb{E}_{S\sim\rho^n}\ \mathcal{E}_p(A(S)) - \frac{1}{8}}{\frac{7}{8}}$$

$$\mathbb{E}_{S\sim\rho^n}\ \mathcal{E}_\rho(A(S)) \geqslant \frac{1}{4}$$

By combining these inequalities,

$$\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{\mathbb{E}_{S\sim\rho^n}\ \mathcal{E}_p(A(S)) - \frac{1}{8}}{\frac{7}{8}} \geqslant \frac{\frac{1}{4} - \frac{1}{8}}{\frac{7}{8}}$$

$$\therefore \quad \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{1}{7}$$

■

(g) (i) Summarize the above results in the formal statement of a theorem.

Theorem (No-Free-Lunch). Let A be a binary classification learning algorithm (uses misclassification excess risk) with a domain X of infinite cardinality. Then given that the training set is of cardinality $n \in \mathbb{N}$, there is a probability distribution ρ over $X \times Y$ (where $Y = \{-1, 1\}$) such that:

- There is a function $f_\star : X \to Y$ such that $\mathcal{E}_\rho(f_\star) = 0$ ($\inf_{f:X\to Y}\ \mathcal{E}_\rho(f) = 0$).
- $\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{1}{7}$

Note: The coursework document states that we should assume that X is infinite. An alternative statement can be made including finite domains X.

(ii) Consider the definition of *learnable space of functions:* **Definition** A space $\mathcal{H}$ of functions $f : X \to Y$ is **learnable** if there exists an algorithm A such that, for any $\epsilon, \delta \in (0, 1)$ and any distribution ρ over $X \times Y$ there exists an integer $\overline{n}$ such that for any $n \geqslant \overline{n}$

$$\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) - \inf_{f\in\mathcal{H}} \mathcal{E}_\rho(f) \leqslant \epsilon) \geqslant 1 - \delta$$

Compare this definition with the results above. Does the No-Free-Lunch theorem imply that the space Y^X of all functions $f : X \to Y$ is (or is not) *learnable*? Why?

Discussion of Definition. This definition implies that a learning algorithm A can be constructed that, given a sufficiently large training dataset (n), chooses a function from $\mathcal{H}$ with misclassification excess risk arbitrarily close (within ϵ) to the optimal function in $\mathcal{H}$ (lowest misclassification excess risk) with arbitrary probability $(1 - \delta)$. This, though possibly practically infeasible (for all ϵ, $\delta \in (0, 1)$), provides a theoretical guarantee.

proof (Y^X is not a learnable space of functions (using misclassification excess risk). From the No-Free-Lunch theorem, we know there exists a distribution ρ such that $\inf_{f\in Y^X}\ \mathcal{E}_\rho(f) = \inf_{f:X\to Y}\ \mathcal{E}_\rho(f) = 0$. Now, suppose that Y^X is a learnable space of functions, then by definition: (where n be the size of the training dataset)

$$\forall\epsilon, \delta \in (0, 1)\ \exists\overline{n} \in \mathbb{N}\ \forall n \geqslant \overline{n} : \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) \leqslant \epsilon) \geqslant 1 - \delta$$

Using the fact that $\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \epsilon) = 1 - \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) \leqslant \epsilon)$:

$$\forall \epsilon, \delta \in (0,1)\ \exists \overline{n} \in \mathbb{N}\ \forall n \geqslant \overline{n} : \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \epsilon) \leqslant \delta$$

If we choose $\epsilon = \frac{1}{8}$ and $\delta < \frac{1}{7}$, then:

$$\exists \overline{n} \in \mathbb{N}\ \forall n \geqslant \overline{n} : \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \leqslant \delta < \frac{1}{7}$$

$$\therefore \quad \exists \overline{n} \in \mathbb{N}\ \forall n \geqslant \overline{n} : \mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) < \frac{1}{7}$$

This contradicts the No-Free-Lunch theorem (for all n: $\mathbb{P}_{S\sim\rho^n}(\mathcal{E}_p(A(S)) > \frac{1}{8}) \geqslant \frac{1}{7}$). Thus, Y^X **is not** a learnable space of functions.

■

(h) Comment on the implications of the No-Free-Lunch theorem with respect to the design of a machine learning algorithm.

The No-Free-Lunch theorem states that a single universal optimal learning algorithm (using misclassification excess risk) does not exist for binary classification. This implication can be extended to general machine learning problems. Thus, all learning algorithms must be tailored to fit the specific task and underlying data distribution. In practice, this is accomplished by:

- Making assumptions on the data distribution and building models based on these assumptions.
- Using different loss functions and regularization.
- Experimenting and testing multiple models/parameters.