DSA5105 Homework 2

Due: 25/09/2021

Name: _____

Student Number: _____

This homework tests your overall understanding of the material covered in the class so far. Please use it as an gauge of your knowledge, and also to prepare for the mid-term test, which will be of a similar format. There are two parts:

- In the first part, there are 5 multiple choice questions. For each question, please mark one and only one choice. Each question is worth 1 point.
- The second part consists of 4 written problems. The first two problems are mandatory, **the last two are optional**. They will not be graded but solutions will be provided. Each problem is worth 5 points. Please write your solutions to each problem legibly on separate pieces of paper and attach them to this booklet. If you prefer, you may also typeset your solutions.

Please scan and upload your solutions to the Luminus submission folder.

Question:	1	2	3	Total
Points:	5	5	5	15
Score:				

1. (5 points) **Multiple Choice Questions.**

   (a) Error on the training dataset is a correct measure of model performance on unseen data.
   ◯ True ◯ False

   (b) Least squares on linear basis models must have a unique solution.
   ◯ True ◯ False

   (c) Any 3 distinct points in $\mathbb{R}^2$ each belonging to one of two possible classes are linearly separable.
   ◯ True ◯ False

   (d) When tuning hyper-parameters such as number of leaves and nodes for a decision tree ensemble, it is appropriate to tune them using the test set to minimize the test error.
   ◯ True ◯ False

   (e) In a trained support vector machine for linearly separable data, only the data points corresponding to the support vectors are required to make predictions on new data.
   ◯ True ◯ False

2. Recall that in regression using linear basis models on a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ ($x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$), we minimize the empirical risk

$$R_{\text{emp}}(w) = \frac{1}{2N}\|\Phi w - y\|^2 = \frac{1}{2N}\sum_{i=1}^{N}(w^\top \phi(x_i) - y_i)^2, \tag{1}$$

   where $\phi(x) = (\phi_0(x), \ldots, \phi_{M-1}(x))$ denotes the feature vector and $\Phi_{ij} = \phi_j(x_i)$ is the design matrix. $y = (y_1, \ldots, y_N)$ is the vector of outputs.

   (a) (2 points) Derive the least squares solution $\widehat{w} = (\Phi^\top\Phi)^{-1}\Phi^\top y$ (assume $\Phi^\top\Phi$ is invertible).

   (b) (2 points) Suppose now that we weight the training samples unevenly, so that we have the weighted empirical risk

$$R_{\text{emp}}(w, a) = \sum_{i=1}^{N} a_i (w^\top \phi(x_i) - y_i)^2 \tag{2}$$

   where each $a_i > 0$ represents the uneven weighting. Derive the least square solution in this case under similar invertibility conditions as in (a).

   (c) (1 point) Discuss the application scenario in which having uneven weights $\{a_i\}$ could be useful.

3. Like decision trees, the *nearest-neighbour* classifier is another class of piece-wise constant hypothesis space for classification. In this case, given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ and a new sample x, we simply assign to x the label of the nearest sample to x in $\mathcal{D}$, i.e.

$$f(x) = y_{i(x)} \text{ such that } i(x) = \arg\min_{j=1,\ldots,N}\|x - x_j\|^2. \tag{3}$$

   Here, $\|\cdot\|$ denotes the usual Euclidean norm.

   (a) (3 points) Show that in the nearest neighbour classifier, the index $i(x)$ can be written purely as an optimization problem over inner/dot products.

   (b) (2 points) Formulate the kernel version of the nearest-neighbour classification (3).

4. **[Optional]** Consider a one dimensional linear regression problem on the dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$ ($x_i \in \mathbb{R}, y_i \in \mathbb{R}$). Our hypothesis consists of affine functions of the form

$$f(x) = w_0 + w_1 x \tag{4}$$

Now, suppose that we have some noise ϵ_i injected to each data point x_i. Each ϵ_i is an independent scalar random variable with mean 0 and variance σ^2. Our goal is now to minimize the expected risk under this noise:

$$\min_{w_0, w_1} \mathbb{E} \frac{1}{2N} \sum_{i=1}^{N} [w_0 + w_1(x_i + \epsilon_i) - y_i]^2 \tag{5}$$

(a) Show that minimizing (5) is equivalent to the original simple least squares, but with an additional regularizer on w_1 (which you should identify).

(b) Now suppose there is also a bias to the noise so that $\mathbb{E}\epsilon_i = b \neq 0$. Does the minimization problem change? Why or why not?

5. **[Optional]** Recall that the Gaussian or Radial Basis Function (RBF) kernel is given by

$$k(x, x') = \exp\left(-\|x - x'\|^2 / 2s^2\right) \qquad s > 0. \tag{6}$$

Show that k is a SPD kernel (See Definition 2.16 in the notes). You may assume the results in Proposition 2.18 in the notes without proof. You may also find hints in Exercise 2.19 in the notes useful. *Note: here $x, x' \in \mathbb{R}^d$ are vectors, so the direct derivation given in the lectures for $d = 1$ does not apply.*