Problem 1. (Train, validation, test [15 points])

You are addressing a regression problem with $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. You have tried five different approaches: A, B, C, D, E. Each approach gives you a predictor. So, your set of predictors are $f_A, f_B, f_C, f_D, f_E : \mathbb{R}^d \to \mathbb{R}$. You divided the dataset into a training and test set. You further performed 5-fold cross-validation on the training set. You obtained the following average train error and validation error (averaged over the 5-folds) for each model.

model	train error	validation error
A	1.355	1.423
B	9.760	9.165
C	5.033	0.889
D	0.211	5.072
E	0.633	0.634

Answer the following questions. Note that full marks will be given only if you justify your answer.

1. Let $I_1, I_2, \ldots, I_5$ denote the indices of the datasets in the validation set for each fold. For model A, write the formula for the average and the standard deviation of the mean-square error of the 5 folds.

2. Which model(s) seems to be overfitting?

3. Which model(s) seems to be underfitting?

4. Your colleague tells you that he does not believe the result of one of your models and he thinks you might have made a mistake in coding one of the models. Based on the train and validation errors above, which model is more likely to have a mistake in its coding?

5. Suppose that the variance of the error across the folds for model E is very high and for model B is much lower. Which model is likely to have a test error on unseen data more similar to the validation error reported above?

6. Your friend suggests that you average model A and model B outputs for the regression task. Explain how you could check the correlation of the errors of the models to determine whether this averaging could potentially give you a better result on unseen data?

7. Suppose that model A corresponds to a decision-tree with depth 2 whereas Model E corresponds to a neural network with 5 hidden layers. Which of the two models would you choose for the regression problem and why? You might consider that the regression task impacts humans and interpretability is helpful.

Problem 2. (Naive Bayes classifier [22 points])

We consider a cancer diagnosis problem. In this problem, our features are $x \in \mathbb{R}^2$, with x_1 denoting the average radius and x_2 denoting the average texture of a tumor. The classification is whether the tumor is *malignant* or *benign* based on the feature x[1]. For N number of patients a medical expert has labeled the data x^i with $y^i = 0$ for benign and $y^i = 1$ for malignant tumor. Our goal is to design an algorithm that learns to do the labelling for a new patient by measuring the x_1, x_2 of its tumor. For this, we want to use a Naive Bayes Classifier.

1. Given a data point x write the Bayes rule for determining the conditional probability of class 0 and class 1.

2. Suppose our dataset contained feature and diagnosis for $N = 1000$ patients, from which 50 had a malignant tumor. What is the empirical estimate of prior probability $P(y = k)$ for $k \in \{0, 1\}$ based on this data?

3. Suppose we use a Gaussian Naive Bayes Classifier in the rest of this exercise. Write the assumptions underlying this model.

4. Explain how you would compute the conditional distribution of average radius of a tuomr, conditioned on the tumor being malignant.

5. After fitting the parameters of the Gaussian distribution for each feature and conditioned on each type of tumor, you found that your classifier classifies only 40 of the 50 malignant tumors as malignant and it classifies 945 of the benign tumors as benign. Let the malignant tumor correspond to the "positive" class. What is the number of false positives, the number of false negatives and the error rate of your classifier?

6. Given a new patient's feature data x, write the formula for determining whether the patient has a malignant or a benign tumor based on your trained classifier.

7. Suppose you now use your classifier to analyse data from new patients. The medical doctor asks you to bring the cases for which the machine learning has less confidence regarding the diagnosis directly to him so that he can use his expert knowledge. How would you use the information obtained from the probabilistic prediction of the Naive Bayes filter to decide which cases need additional attention?

[1] A tumor is an abnormal collection of cells. It forms when cells multiply more than they should or when cells don't die when they should. A tumor can be malignant (cancerous) or benign (not cancerous)source

Problem 3. (Neural networks [15 points])

We have an audio signal from a piece of music and we want to classify the music according to its genre[2]. For this, we use a training dataset which consists of a library of audio signals, labelled according to their genre. Let $x \in \mathbb{R}^d$ denote an audio signal. Here, $x = (x_1, x_2, \ldots, x_d)$ with x_i denoting the acoustic pressure measured at time step i. For training, we use a dataset consisting of 1,000 4-second music excerpts evenly distributed into nine classes: rock, reggae, blues, classical, disco, country, hip-hop, jazz, and pop. We consider audio samples with 5,000 samples per second for 4-second. Therefore, the input to our classifier is a $5,000 \times 4 = 20,000$-dimensional vector.

1. Suppose we have a neural network with two hidden layers and an output layer for the 9 classes. Each hidden layer has 10 nodes. How many weights and biases need to be determined for each layer? Show your work.

2. You observe that the number of training data you have is relatively small compared to the number of parameters. After training the network, you get a very small training error. Hence, you suspect your neural network is possibly overfitting to the training data. What approach could you use to reduce this potential overfit?

3. Your friend who is very musical thinks that music audio signals can be distinguished based on local characteristics of the signal and she suggests you to use a 1-dimensional convolutional neural network. Suppose now for each of the first and second layer of the neural network, you use 128 filters of dimension 5 for each layer. Hence, each filter is given by $w = (w_1, w_2, \ldots, w_5)$, where w_i's needs to be designed. The filter is applied with a stride of 5. It follows that after applying each filter to an audio signal of length d, we will have a signal of length $\frac{d}{5}$. How many parameters need to be determined for the first, the second and the output layer? Show your work.

4. You train the network above using stochastic gradient descent, with a batch size of 100 and for 10 epochs. How many iterations of gradient descent is being run in each epoch?

5. Your friend says that she has used the same network architecture, learning rate and batch size. However, her training and test error are different from yours. What could be some reasons for the difference?

Remark: For more information regarding this approach, you may see this paper, and for the audio dataset you may see here.

[2]Such a problem arises in music streaming or music shopping services.

Problem 4. (Decision trees [18 points])
Consider a classification problem with $x \in \mathbb{R}^2$ and $y \in \{\text{square}, \text{triangle}\}$. The training data is shown in Figure 1 below. There are N_t triangles and N_s squares in the training data, where $N_s = mN_t$ with $m \in (0, 1)$. So, for example, if there are 100 triangles, and $m = .1$. then there are 10 red squares and a total of 110 data points.

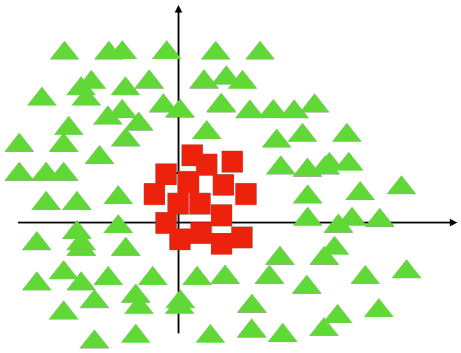

Figure 1: Classification problem training data

1. A so-called null classifier gives the majority label of the training data to any test point $x \in \mathbb{R}^2$. Hence, it considers that x has no effect on the label. Since we have $N_t = (1/m)N_s > N_s$ the majority label is triangle and the null-classifier labels any test point x as a triangle. What is the gini index of this classifier? What is the error rate of this classifier on the training data?

2. Now, consider feature 1 and the threshold at $x_1 = 1$ shown in Figure 2 below as a candidate for forming a split in a first node of a decision tree to be constructed for classification. So, the split criteria is whether $x_1 > 1$. Suppose that a fraction of $c \in (0, 1)$ number of triangles falls to the right of the line at $x_1 = 1$ shown in the figure. In other words, cN_t of triangles have $x_1 > 1$. Hence, $(1 - c)N_t$ are the number of triangles to the left of the line. Write the gini index of the two leaves and of the node according to this split.

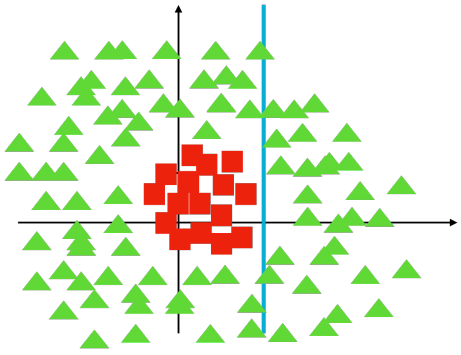

Figure 2: Classification problem with one node of the decision tree

3. Show that the gini index after the split is smaller than the gini index of the null classifier.

4. Observe that anywhere you put a line, the number of "triangles" is more than the number of squares. Thus, show that no matter where you put the blue line, the accuracy of the classifier

does not improve, even though its gini index can improve[3]

5. Draw the boundaries corresponding to a decision tree that could separate the two classes.

6. What is the depth of the decision-tree that separates these two classes?

7. Your friend suggests to you to use a logistic regression for this classification problem. She thinks that it is sufficient to consider two feature as $\Phi_1(x_1, x_2) = x_1^2 + x_2^2$ and $\Phi_2(x_1, x_2) = 1$ for the logistic regression problem. How many parameters you would need to learn for the logistic regression model? What would the decision boundaries look like in this case?

[3]This is one of the motivations of using other criteria than accuracy in defining the decision-trees.