

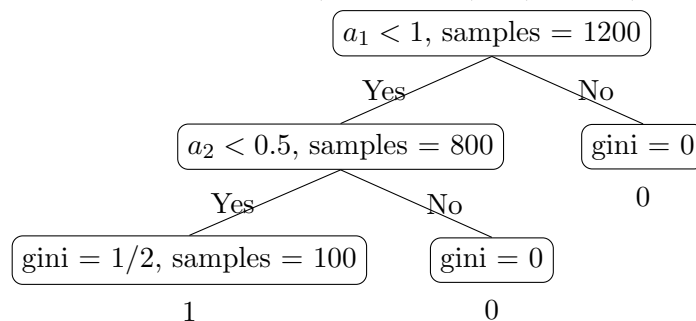
Problem 1: PCA, decision trees, logistic regression. [16 points]

As an engineer focused on advancing electric vehicle design, you are tasked with evaluating the suitability of various materials for use as chassis of the vehicle. Given a dataset of mechanical properties from different materials you aim to develop a predictive model that determines whether a material is suitable for the chassis¹. The dataset has 1500 samples and 15 features. The labels are 1: appropriate, 0: inappropriate.

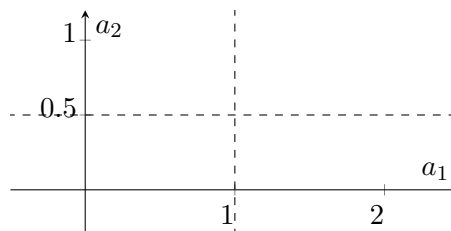
1. You notice that several samples are missing values for features 2, 5, 10. State two approaches to address the missing values. (2 points)
2. You suspect some of the features are correlated. Hence, you perform a dimensionality reduction. Let $X \in \mathbb{R}^{1500 \times 15}$ denote the normalized data, prior to dimensionality reduction.
 - (a) How do you determine the first 2 principal components, $v_1, v_2 \in \mathbb{R}^{15}$, of X ? (2 points)
 - (b) Write the equation for projecting the data on the space spanned by v_1, v_2 . (2 points)

Now, we consider the reduced dimensional dataset $A \in \mathbb{R}^{1500 \times 2}$. Denote the features by a_1, a_2 . We aim to develop a decision tree for determining the suitability of a given material as a chassis. Let us use 1200 samples for training and the remainder for test. From this 1200 samples, 50 had label 1.

3. Suppose the second feature has values in the set $\{-3, -2, -1, 0, 1, 2, 3\}$. To compute the best value for making a split in a node using this feature, for which candidate values you need to compute and compare the gini indices? (1 point)
4. The optimized tree is as follows. What is the false positive rate, that is, the ratio between false positives and the total number of negatives (true label 0) ? (2 points)



5. For the regions shown in the graph below, put the label, $\{0, 1\}$ based on the prediction of the above decision tree. (2 points)



6. Next, rather than using a decision tree, you decide to do the classification with logistic regression. The optimization of logistic loss results in the line $a_2 = -\frac{1}{2}a_1 + 1$. Draw the line on the graph and determine the labels according to the logistic regression predictor. (3 points)
7. The training and test accuracy of the two methods are given below. What would be the probability of error on an unseen data point for each of the two methods? (2 points)

	training set accuracy	test set accuracy
decision tree	0.96	0.92
logistic regression	0.93	0.94

¹The problem is inspired by the following paper.

Problem 2: k-means, Naive Bayes and linear regression. [16 points]

The starting class of mechanical engineering in university X is very large, and being an international university, many students have moved from abroad. To make friends, the students have found *app-F*. They log into this app through their Instagram account, and the app uses their historical data to categorize them and then recommends potential friends as students in the same category.

1. Suppose the app has used historical data of its past $N = 4000$ users, extracting 5 features from the samples, and then has used k -means to categorize people. For $k = 3$, let $\mu_i \in \mathbb{R}^5$, $i = 1, 2, 3$, be the means. For a new student with feature vector $x^i \in \mathbb{R}^5$ how could we use the clusters to design a 1-nearest-neighbor (1-NN) classifier with Euclidean distance metric? (1 point)
2. The above nearest neighbor classifier can act poorly. So, the app decides to design a Naive Bayes classifier. The number of data points in the three categories A, B, C , were 1500, 300, 2200, respectively. What is the prior probability of category A ? (1 point)
3. For data points in category A the mean of feature 1 is 11.5 and its standard deviation is 2.34. Write the likelihood of feature 1 given category A , assuming Gaussian distribution. (2 points)
4. Feature 2 takes values in $\{0, 1\}$. In category A , 1000 samples have value of 1 for their feature 2. Write the probability of feature 2 being equal to 1 given category A . (2 points)
5. Using the first two features and the Naive Bayes assumption, complete the expression for the probability of a student with feature $x^{\text{test}} = (13, 0)^\top$ to belong to category A , using Parts 3 and 4 results. (2 points).

$$P(y^{\text{test}} = A | x^{\text{test}}) = \frac{\dots\dots\dots}{p(x^{\text{test}})}$$

In the campus orientation week, the students listened to a talk from the recent Nobel laureate highlighting the dangers of AI in reducing diversity in our societies². Several students wisely deleted their app, and opted for a new approach to find friends. Namely, by attending their classes, and sitting next to a random student each time. Luckily, the teacher finally has a reasonable number of students to study the potential correlation between class attendance and grades.

6. Let x denote the number of class hours attended, and y denote the grade of a student. For $N = 200$ students, the covariance matrix D of the above features and the mean vector μ is given below. Write the formula for the correlation between x and y and compute it. (2 points)

$$D = \begin{pmatrix} 9 & 5 \\ 5 & 4 \end{pmatrix}, \quad \mu = \begin{pmatrix} 24 \\ 63 \end{pmatrix}$$

The teacher subtracts the mean of the features from each student data point (x^i, y^i) , and then releases the resulting N data pairs, anonymized (hiding identities of students).

7. Student S wants to use the above to fit a linear regression and then decide how many hours to attend lecture to get a desired grade. To this end, she considers fitting a line without an offset, $y = wx$, to the released data. Write the mean-square-loss she needs to optimize. (1 point)
8. Derive the gradient of the loss function with respect to the line slope $w \in \mathbb{R}$. (1 point)
9. Compute the w that optimizes the loss based on the information in matrix D . (2 points)
10. Given the feature means μ in part 6, use the above model to determine how many lecture hours student S should attend to have an expected grade of 83. (2 points)

Note: ideally, the student should have divided the data to a training and test sets. Furthermore, we don't recommend using this approach as it can only give an "expected" grade.

²Listen to the talk here.

Problem 3, Convolutional neural networks, train and test performance. [18 points]

Landmines are responsible for death of thousands of people per year, most of them civilians, and there are still over 100 million landmines buried in earth³. Committed to use your engineering degree for a meaningful cause, you have joined a startup that uses drones equipped with infrared camera to locate buried landmines⁴. Now you will use your AI course to locate landmines from camera images.

Let us consider a grayscale image of size 256×256 .

1. For training data, you analyze 1200 images obtained from landmines with known locations. You have labeled an image 1 if there indeed exists a landmine in the region captured by the image and 0 otherwise. Of these 1200 images, 40 had a landmine. What would be the accuracy of any classifier that labels all images as 0? Should one be satisfied with this classifier? (2 points)

We aim to design an automated method for detecting landmines from stream of camera images.

2. Our aim is to first reduce the noise in the image and then detect the edges. To this end, which of the two filters below should be applied first to the image? Justify your answer. (2 points)

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 0.055 & 0.11 & 0.055 \\ 0.11 & 0.23 & 0.11 \\ 0.055 & 0.11 & 0.055 \end{pmatrix}.$$

3. After the above convolutions (without padding), you use 1000 of data points for training a logistic regression classifier. To this point, you flatten the images, and use them as input to the classifier. How many parameters need to be determined to define this classifier? (2 points)
4. You observe that the false negative rates are higher than the false positive rates of your classifier. To try to remedy this, would you choose $\tau \in (0, 1)$ or $\tau > 1$ below? Justify. (2 points)
 $L(w, b) = \frac{1}{N} \sum_{i=1}^N y^i \log(1 + e^{-z_i}) + \boxed{\tau}(1 - y^i) \log(1 + e^{z_i})$.
5. Now, your classifier seems to overfit to the training data. Propose a modification in the cost above that has the potential to reduce this overfit. (1 point)

Instead of following the steps 2-5 above, you decide to design a convolutional neural network (CNN).

6. You apply 3 convolution filters in the first layer of your CNN (without padding), with each filter having a size of 5×5 . How many weights and biases need to be determined to define these filters? (2 points)
7. Next, you pass the 3 matrices obtained after the convolution step above through a hyperbolic tangent (tanh) nonlinearity. What would be the range of values of each entry of the resulting 3 matrices? (1 point)
8. In the third step, you do a max pooling with stride of 4. What is the size of each of the 3 output matrices? (1 point)
9. Finally, you flatten the image and use it as an input to a logistic regression classifier. How many weights and biases need to be determined for this last layer? (2 points)
10. You find that the training accuracy is 99.9% but the test accuracy is 90%. Is the approach overfitting or underfitting? (1 point)
11. Your colleague thinks that your neural network is too simple. Is it likely that making the network deeper fix the above discrepancy? Justify your answer. (1 point)
12. Your other colleague thinks that stochastic gradient descent has not been able to find the optimal set of parameters. Is this likely based on the above training and test accuracies? (1 point)

³See the United Nations report here.

⁴Inspired by the following article.

Problem 4, Reinforcement learning. [20 points]

Lithium-ion batteries are crucial in power systems, electric vehicles, and consumer electronics. However, developing an optimal charging scheme for them is challenging. We will use reinforcement learning to this end⁵. Let $x \in \mathbb{R}^4$ denote the state of the battery dynamics (capturing voltages at different locations and the state of charge of the battery) and $u \in \mathbb{R}$ denote the input, which is the current applied to the battery. The discretized dynamics are given by:

$$x(t+1) = f(x(t), u(t), w(t)),$$

where $w(t)$ is the uncertainty, sampled from the uniform distribution on the interval $[0, 1]$. Furthermore, $f : \mathbb{R}^4 \times \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}^4$ is unknown but we have a simulator that can evaluate it at any x, u, w .

1. Given an initial condition $x(0) \in \mathbb{R}^4$, and an input $u(0), u(1)$ write the steps for generating a sample trajectory $x(1), x(2) \in \mathbb{R}^4$. (2 points)

Our goal is to ensure x_4 , the battery charge, reaches a desired level x_d as fast as possible, while not damaging the battery. To this end, we design a reward, which consists of a unit penalty for every time step taken while $x_4(t)$ has not reached x_d and a term that penalizes the voltage x_1 going beyond a threshold \bar{V} : $\max(x_1(t) - \bar{V}, 0)$ while x_d has not been reached. The discount factor is $\gamma \in (0, 1)$.

2. Fill in the expression for the reward of a trajectory that takes T steps to reach x_d . (1 point)

$$J(x(0), \dots, x(T), u(0), \dots, u(T-1)) = \mathbb{E} \sum_{t=0}^T \underline{\hspace{10em}}$$
3. Write the expression for the discounted cumulative reward of the following trajectory: (2 points)

$$\{x(t)\}_{t=0}^{10}$$
 reaches x_d at time step 10 with x_1 being one unit above \bar{V} at step 5.

Now, we want to design a control policy $\pi_\theta : \mathbb{R}^4 \rightarrow \mathbb{R}$, where θ is a policy parameter, to minimize the cost given in Part 2 of this problem.

4. How many parameters should be determined in each of the following policy classes: (3 points)
 - (a) a linear policy with $u = \theta^T x$.
 - (b) a neural network policy of 2 hidden layers, with 10 neurons in each depth.
5. Consider the linear policy in Part 4(a) above with addition of noise from a Gaussian distribution with zero mean and constant variance. Let us denote the corresponding probability density function of the policy by $\pi_\theta(u|x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\theta^T x)^2}{2\sigma^2}}$. Determine $\nabla_\theta \log \pi_\theta(u|x)$. (2 points)
6. Describe an approach for sampling N trajectories of length T and explain how to estimate $J(\pi_\theta)$ using these sampled trajectories. (2 point)
7. Given the formula for estimating the policy gradient below, how would you use the above N trajectories, Part 5 and Part 6, to compute $\hat{\nabla}_\theta J(\pi_\theta)$? (2 points)

$$\hat{\nabla}_\theta J(\pi_\theta) = \frac{1}{N} \sum_{i=1}^N \left(\sum_{t=0}^T \gamma^t r(x^i(t), u^i(t)) \right) \left(\sum_{t=0}^T \nabla_\theta \log \pi_\theta(u^i(t)|x^i(t)) \right).$$

8. Write the stochastic gradient ascent update for θ with a learning rate of η . (2 points)
9. You initialize θ randomly and after 100 iterations of the stochastic gradient ascent, θ converges to θ_s . With a different initialization, would the algorithm converge to θ_s ? Justify. (1 point)
10. Now, you implement the policy π_{θ_s} on your headphone, which also has a Lithium-ion battery, by plugging and programming a variable resistor on the outlet. Unfortunately, the voltage gets out of the limit and your headphone stops working. Provide three potential reasons for why the implementation outcome is different than the simulation outcome. (3 points)
Note: this motivates the flourishing research field of safe data-driven control.

⁵Inspired by the following article, where we significantly simplified the model (from 61 dimensions to 4 dimensions).

Problem 5, AI Ethics. [2 bonus points]

You are considering working for a new startup whose goal is to bring robots to senior people's homes. There are a number of ethical issues that you reflect upon before deciding to join this company⁶. Categorize each of the 6 issues below in the three categories discussed in AI Ethics videos, namely:

A. Narrative framing; B. Social justice; C. Ethics in technology.

1. The company seeks publicity in the Swiss news and social media channels to highlight benefits of robots in old people's homes.
2. The algorithms to train the robots are based on data obtained from robots interacting with old people in 2 hospitals in Canton Aargau.
3. An estimated number of 1000 nurses would lose their jobs 2030 if the company becomes widely successful in Switzerland.
4. Each robot costs over 30,000 CHF and thus, only a fraction of seniors can afford to have such robots.
5. A select number of seniors in certain homes are consulted on whether such technology is preferred to human nurses.
6. The control algorithm for navigation of robots is based on reinforcement learning and there is no clear way to interpret and understand how robots make decisions.

⁶Inspired by the following article.