Problem 1 (Classification) **(10 points)**
Submit your answers to "ML Exam - Problem 1"

CCARD is a company that facilitates electronic funds transfers throughout the world using credit card. Recently, CCARD has received several complaints from clients saying they have been charged for items that they did not purchase. CCARD approached us and asked us to develop a binary classifier $f_\theta(x)$ to predict if a transaction $\mathbf{x} \in \mathbb{R}^{100}$ is fraudulent or genuine. They provide us with a dataset $D = \{\mathbf{x}_i, y_i\}_{i=1}^N$ with $N = 30000$ transactions out of which only 500 are fraudulent, i.e. $y = 1$, and the rest are genuine, i.e., $y = 0$. This means the percentage of fraud is only 1.67% of all transactions. We have chosen to model our classifier with a neural network with 3 fully connected layers and ReLU as the activation function. Our goal is to find the optimal parameters θ^*:

$$\theta^* = \arg\min_\theta \mathcal{L}(D;\theta), \tag{1.1}$$

where the loss function is the binary cross entropy, i.e.,

$$\mathcal{L}(D;\theta) = -\sum_{(\mathbf{x},y)\in D} y \log\left(P_\theta(\hat{y} \mid \mathbf{x})\right) + (1-y)\log\left(1 - P_\theta(\hat{y} \mid \mathbf{x})\right), \tag{1.2}$$

with $P_\theta(\hat{y} = 1 \mid \mathbf{x}) = \frac{1}{1+e^{-f_\theta(\mathbf{x})}}$.

(a) Is it possible to find a closed form solution for the optimal model parameters θ^*? If not, explain which optimization algorithm you would use, explaining your decision and stating if it has any parameters. If applicable, write down the update rule of the optimization algorithm. (1 points)

(b) We split our dataset in training, validation and test sets. Assume, we have trained our classifier for 20 epochs. Figure 1 shows the evolution in epoch of the loss function for the different splits. Describe in your own words what you observe in the figure and indicate whether overfitting or underfitting is observed. If it occurs, we propose to use early stopping to avoid this issue. Do you think it is a good strategy? Why? Which is the optimal epoch to stop the training? Why? (2 points)

(c) We have trained four different classifiers with different configurations of the hyperparameters (e.g., number of neurons per layer, learning rate, batch size). Table 2 shows the results for the four different configurations. CCARD told us that it is important to miss as few frauds as possible To that end, which of the configurations in Table 2 would you choose? Explain your decision, and indicate which evaluation metric(s) you used to make the decision. (2 points)

(d) CCARD is now using our algorithm. However, they tell us that our algorithm predicts too many valid transactions as frauds, and ask us to solve this issue. We propose to modify the loss function and instead use a cost-sensitive cross entropy loss with the costs shown in Table 3.

(i) Derive the expression for the cost-sensitive cross-entropy loss.
Hint: Note that in the standard Cross-entropy loss above, we are minimizing the negative log-likelihood (not the error probability). In the likelihood computation, one can encourage the correct predictions using the cost-sensitive weights, for example, by encouraging true positive predictions with a factor of $-C_1$. (2 points)

(ii) To improve precision, which cost should be higher? C_1 or C_2? Explain your answer. (1 points)

(e) CCARD has provided us with a test set with 8 samples, as shown in Table 4. Compute the accuracy, F2-score, precision and recall on this test set. **Important!** *We are asking for the F2-score not the F1-score.* (2 points)

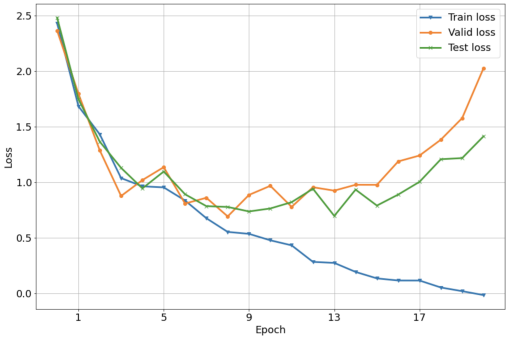


Figure 1: Evolution of the loss function $\mathcal{L}(D;\theta)$ with respect to the epochs for the train (blue), validation (orange) and test (green) sets.

Config.	Accuracy	Precision	Recall	F1 Score
1	98.07	95.24	3.34	6.45
2	81.33	6.73	86.96	12.50
3	95.00	34.44	47.52	39.94
4	67.67	3.39	20.55	5.83

Table 2: Evaluation metrics evaluated in the validation set for four different configurations.

	Pos. prediction	Neg. prediction
Pos. label	0	C_1
Neg. label	C_2	0

Table 3: Cost matrix.

	1	2	3	4	5	6	7	8
True label	0	1	0	0	1	0	1	0
Prediction	0	0	1	0	1	0	0	0

Table 4: True labels and predictions for the test set consisting on eight samples.

Problem 2 (Bayesian Decision Theory and Regression) **(10 points)**
Submit your answers to "ML Exam - Problem 2"

Regression. The elastic net regression problem is defined as

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \lambda_1\|\mathbf{w}\|_1 + \lambda_2\|\mathbf{w}\|_2^2, \qquad \mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^n$$

(a) Briefly explain the role of λ_1 and λ_2 and how the learned weights are influenced by the choices of λ_1 and λ_2. (2 points)

(b) Is the elastic net regression problem convex? Does it have a unique solution $\mathbf{w}^*$? How can we learn a/the optimal weights of the elastic net regression problem? (1 points)

Let us now assume that the true relationship between features $\mathbf{X}$ and target variable $\mathbf{Y}$ is linear, and consider instead a regularized linear regression model of the form

$$\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathbb{R}^d} \|\mathbf{Y} - \mathbf{X}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|_1, \qquad \mathbf{X} \in \mathbb{R}^{n\times d}, \mathbf{Y} \in \mathbb{R}^n$$

(c) How does **increasing** λ affect the bias and the variance of the predictions made by the resulting regressor? Choose the correct answer from the following options: (1 point)

(a) Decrease bias, decrease variance
(b) Decrease bias, increase variance
(c) Increase bias, decrease variance
(d) Increase bias, increase variance
(e) We cannot predict this

(d) Write down a set of basis functions $\{\phi_k(m)\}_k$ that allows you to represent the months of the year (inputted as a number between 1 and 12) such that the Euclidean distance between the resulting embedding for any two consecutive months (e.g., March and February, January and December, and so on) is the same. (2 points)
Important! *Answers that map any two months to the same features are not allowed.*

Bayesian decision theory. Consider now a binary classification problem $\mathcal{Y} = \{-1, 1\}$, with the following distribution on $\mathcal{X} = [0, 1]$,

$$\mathrm{P}(Y = 1 \mid X = x) = \begin{cases} 0.2, & \text{if } 0 \le x \le 0.25 \\ 0.8, & \text{if } 0.25 < x < 0.75 \\ 0.2, & \text{if } 0.75 \le x \le 1 \end{cases}$$

(e) What is the Bayes optimal error of this problem? (2 points)

(f) Report the optimal set(s) of parameters (w^*, b^*) (i.e., those that minimize the error probability), and the resulting error probability for a classifier(s) of the form

$$f_{(w,b)} = sign(wx + b), \qquad w, b \in \mathbb{R}.$$

(2 points)

Hint: *You don't need to give any derivations. If there are more than one optimal set of parameters, provide all possible optimal parameters in set notation.*

Problem 3 (SVMs & Kernel Methods) **(10 points)**
Submit your answers to "ML Exam - Problem 3"

(a) Consider the following *soft margin Support Vector Machine (SVM)*:

$$\min_{\mathbf{w}\in\mathbb{R}^d, \boldsymbol{b}\in\mathbb{R}, \boldsymbol{\xi}\in\mathbb{R}^n} \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\xi_i$$
$$\text{subject to: } y_i\left(\langle \mathbf{w}, \mathbf{x}_i\rangle + b\right) \geq 1 - \xi_i, \quad \forall i = 1, \ldots, n, \quad \xi_i \geq 0$$

Explain in your own words the role of the slack variables here. Describe the influence of λ on the number of training data that can violate the margin. (1 points)

(b) Consider the following *Kernel Ridge Regression* loss function:

$$L = \frac{1}{2}\|y - \mathbf{K}\alpha\|_2^2 + \frac{\lambda}{2}\alpha^T\mathbf{K}\alpha$$

Where in the above expression is the kernel trick used? Explain it in your own words, how the kernel trick is useful here. (2 points)

(c) Consider now a mapping function $\phi(x) : \mathbb{R} \to \mathbb{R}^3$, which takes the form $\phi(x) = (x, x^2, \exp(x))$. Find the kernel $k(x, y)$ associated with $\phi(x)$. (2 point)

(d) Consider a Logistic Regression (LR) model with the following loss function (cross entropy):

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P_\theta(\hat{y} = 1 \mid \mathbf{x}_i) + (1 - y_i)\log\left(1 - P_\theta(\hat{y} = 1 \mid \mathbf{x}_i)\right)\right] \tag{3.1}$$

where $P_\theta(\hat{y} = 1 \mid \mathbf{x}) = \frac{1}{1+\mathrm{e}^{-\theta^T\mathbf{x}}} = \sigma\left(\theta^T\mathbf{x}\right)$ with sigmoid function $\sigma(\cdot) : \mathbb{R} \to [0, 1]$.

Given a training dataset $\{\mathbf{x}_i, y_i\}_{i=1}^N$, where $\mathbf{x}_i \in \mathbb{R}^L$ and $y_i \in \{0, 1\}$, assume that each of the input vectors can be transformed as $\phi(\mathbf{x}) : \mathbb{R}^L \to \mathbb{R}^D$ with $D \geq L$. Derive step by step the dual loss function for the Kernel Logistic Regression for binary classification. State whether and, if so, where you used the kernel trick and/or the representer theorem. Make sure to explicitly state your assumptions and explain every step in your own words. (4 points)

(e) Assume a *Radial Basis Function* (RBF) kernel, which we use to *kernelize* Logistic Regression. How can we avoid that the resulting *kernelized Logistic Regression* overfits? Explain your response, indicating if there is any theoretical result you based your decision on. (1 point)

Problem 4 (Unsupervised Learning) **(10 points)**

Submit your answers to "ML Exam - Problem 4"

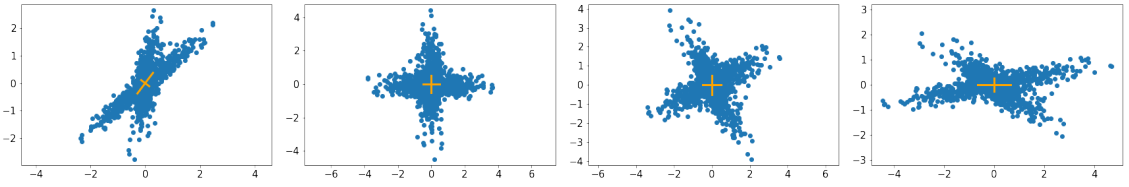
(a) Given the following set of points:

$$\mathbf{x}_1 = (7,5,2); \quad \mathbf{x}_2 = (2,3,4); \quad \mathbf{x}_3 = (6,6,5);$$
$$\mathbf{x}_4 = (9,2,3); \quad \mathbf{x}_5 = (7,5,8); \quad \mathbf{x}_6 = (5,5,7);$$

Apply (that is, make the computations and explain the steps you follow) agglomerative hierarchical clustering using single-linkage with respect to the dissimilarity measure $d(\boldsymbol{a}, \boldsymbol{b}) = \sum_{i=1}^{3} |a_i - b_i|$, and draw the resulting dendrogram. Recall that in a dendogram the y-axis is the dissimilarity measure and the x-axis represents the given points. (3 points)

(b) Show that, if $\mathbf{X} = [\mathbf{x}_1 \mathbf{x}_2 \cdots \mathbf{x}_N]^\top \in \mathbb{R}^{N \times D}$ is your dataset (N instances, D features) which is feature-wise centred ($\frac{1}{N}\sum_n x_{nd} = 0$ for all d), then the transformation $\mathbf{Y} = \mathbf{X}\mathbf{U}\mathbf{\Lambda}^{-1/2}$ yields a data set with covariance equal to the identity matrix—that is, unit variance feature-wise and zero covariance pairwise. Here, $\mathbf{U} = [\mathbf{u}_1 \mathbf{u}_2 \cdots \mathbf{u}_D]$ is the projection matrix given by PCA with D components, and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D)$ the diagonal matrix composed by their associated eigenvalues. (3 points)

(c) We have applied different transformations to a given dataset $\mathbf{X} \in \mathbb{R}^{N \times 2}$. In Figure 2 we show the data (blue dots) as well as the principal directions of their covariance matrix (orange lines whose length is their associated eigenvalue). Which plots correspond to the resulting transformed data after applying PCA and whitening the data? And to ICA (unmixed) data? Justify your answers.



(a) Original data $\mathbf{X}$. (b) Transformed data 1. (c) Transformed data 2. (d) Transformed data 3.

Figure 2: Plots of the same data under different transformations.

(2 points)

(d) Assume a dataset $\mathbf{X} \in \mathbb{R}^{N \times M}$ and a linear transformation of the form $\mathbf{X} = \mathbf{S}\mathbf{A}$ with $\mathbf{S} \in \mathbb{R}^{N \times D}$ and $\mathbf{A} \in \mathbb{R}^{D \times M}$ ($D < M$). What unsupervised approach would you apply if we want $\mathbf{S}$ to explain as much of $\mathbf{X}$ as possible? And if we want to recover the original $\mathbf{S}$, whose rows are independent of each other? For the latter, why is it a problem to assume Gaussianity on row $\mathbf{s} \in \mathbf{S}$? Justify all your answers. (2 points)