Student's Name: _____

This is a document with examples of tasks that will be in Final Exam. The exam will contain 3 sections.

You will have **1.5 hours** to complete this exam.

In **section 1** you need to mark the only correct option. This will be mostly questions like "true/false". This section will cost **30%** of Exam.

In **section 2** you need to mark the correct option or options. You need to **get the answer exactly correct** for **the full grade**: you get **zero** points for the entire question if you have irrelevant or missing options in your answer. This section will cost **30%** of Exam.

In **section 3** you are mostly required to give a full answer and demonstrate all relevant formulae and calculations. You could use the other side of the exam paper to write down your solutions. **Absent or illegible** answers will be graded zero. This section will cost **40%** of Exam.

EXAMINATION RULES

- Students are required to follow all instructions given by the examiners.
- Talking is NOT allowed under any circumstances.
- Students should use the exam paper as a draft and a paper for solutions. The usage of any additional paper is prohibited.
- **Mobile phones are strictly prohibited** in the examination hall. Students MAY NOT bring any electronic device into the examination hall.
- Students may raise their hand to ask the examiner a question. The examiner may decide not to answer the question: the students are expected to know the required terminology and understand the examination questions.
- Once a student has seen the examination paper, the student is assumed to be in good health at the time of examination.
- Students can complete the exam ahead of time. Early completion is not allowed within 20 minutes before the end of the exam.

I have read and understood the examination rules.
I will not cheat, copy from other students, or use unauthorized materials or devices, and I have not brought such materials or devices into the examination hall.

Signature: _____

Section 1

This section will cost **30%** of Exam. The number of questions here will be different in Final Exam.

(1) 1. The logistic regression model for binary classification can be interpreted as a simple neural network and sigmoid activation.

A. True B. False

Solution: The formula for logistic regression is $\sigma(x^T w)$, which is exactly one layer perceptron with weights w and sigmoid activation.

(1) 2. Can backpropagation containing partial differentiating be used in Neural Network modelling if the target is not a continuous variable?

A. Yes B. No

Solution: Yes, it depends on loss function, for example BCE.

(1) 3. Under which transform $f(y)$ it is true that $MSE(f(y), f(\hat{y}))$ equivalent to $MSLE(y, \hat{y})$

A. $f(y) = \log(1 + y)$ B. $f(y) = \log(y)$

Solution:

$$MSE(f(y), f(\hat{y})) = \frac{1}{n}\sum_{i=1}^{n}(f(\hat{y}_i) - f(y_i))^2$$

$$MSLE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\log(1 + \hat{y}_i) - \log(1 + y_i))^2$$

Section 2

This section will cost **30%** of Exam. The number of questions here will be different in Final Exam.

(2) 1. Which of the following hyperparameter(s), when increased, may cause decision tree to overfit the data?

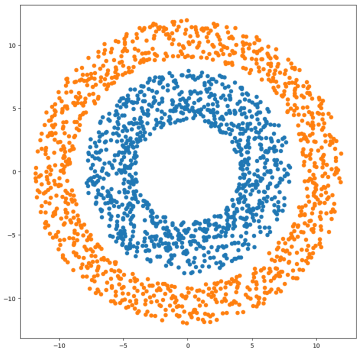
◯ The minimum number of samples required to split

◯ The minimum number of samples required to be at a leaf node

√ **The maximum depth of the tree**

Solution: The minimum number of samples required to split will prevent to have small number of samples at a leaf node, so it is closely related to second answer. When increasing, it may have the effect of smoothing the model. We can treat it as regularization. On the other hand, the maximum depth of the tree, when increased, make model more complex so that it could recognize all training data.

(2) 2. Which of these methods of clustering can handle such clustering problem?



◯ K-means

√ **Spectral Clustering**

√ **DBSCAN**

◯ Gaussian Mixture

Solution: K-means and Gaussian Mixture can provide only convex clusters. On the other hand, Spectral Clustering and DBSCAN can handle such a problem.

(2) 3. Adding an extra feature to a linear regression model may:

√ **Not decrease train coefficient of determination** R^2

◯ Not increase train coefficient of determination R^2

√ **Nothing definitive can be said about the test coefficient of determination** R^2

Solution: The formula for coefficient of determination

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

where y_i is observed data, $\bar{y}$ is an average of all y_i, and $\hat{y}_i$ is our predicted value.

When we add extra feature, $\sum(y_i - \hat{y}_i)^2$ will not change if it is completely uncorrelated with observed data, or it will decrease if it is somehow correlated. Hence, train R^2 will be the same or increase. However, we cannot say this about test R^2

Section 3

This section will cost **40%** of Exam.

1. You were given a task to predict if a person replies "**yes**". You fit a binary classifier to the training dataset and got this confusion matrix on the validation dataset. Supposing that "**yes**" is a positive class, **write a formula and calculate** the following values:

		predicted no	predicted **yes**
actual	no	50	5
	yes	10	100

The confusion matrix for problem 1.

(1) (a) Accuracy

(1) (b) True Positive Rate

(1) (c) F_1 score

Solution: The accuracy score is

$$\text{accuracy} = \frac{\text{tp} + \text{tn}}{n} = \frac{50 + 100}{165} = \frac{10}{11}$$

$$\text{tpr} = \frac{\text{tp}}{\text{tp} + \text{fn}} = \frac{100}{100 + 10} = \frac{10}{11}$$

Now precision and recall are

$$\text{recall} = \mathbb{P}(h(x) = \text{yes} \mid C_x = \text{yes}) \approx \frac{\text{tp}}{\text{tp} + \text{fn}} = \frac{100}{100 + 10} = \frac{10}{11}$$

$$\text{precision} = \mathbb{P}(C_x = \text{yes} \mid h(x) = \text{yes}) \approx \frac{\text{tp}}{\text{tp} + \text{fp}} = \frac{100}{100 + 5} = \frac{20}{21}$$

Formula for F_1:

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = \frac{2}{\frac{11}{10} + \frac{21}{20}} = \frac{40}{43}$$

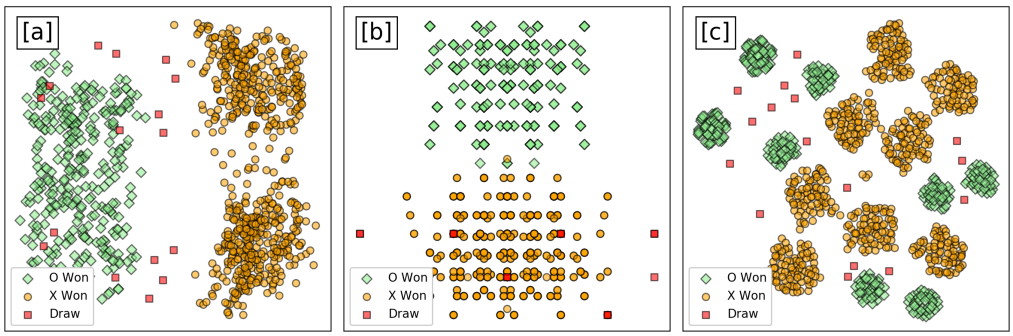
2. Consider a dataset of all endgame positions of the standard 3×3 **Tic-Tac-Toe**. Each state is represented by a vector in $\mathbb{R}^9$ with "**x**" being mapped to $+1$, "**o**" to -1 and empty cells to 0. Match each dimensionality reduction method with the picture of its resulting extracted features for this dataset (bullet – "**o**" won, diamond – "**x**" won, square - draw).

	o	**o**
x	**x**	**x**
o	**x**	

(1) (a) Some pair of components of a linear PCA

(1) (b) 2D t-Distributed Stochastic Neighbor Embedding (t-SNE)

(1) (c) 2D Isometric Mapping (ISOMAP)



Solution: The original 9-dimensional data has grid-like structure because is the subset of $\{-1, 0, 1\}^9$. Since PCA is linear, it preserves some grid-like structure, thus, **b** corresponds to PCA. t-SNE, significantly takes into account the local density (exp) and tries to separate locally dense classes. Thus, plot **c** corresponds to t-SNE. In IsoMap, nearest neighbors will mainly be from the same class, thus IsoMap is expected to provide dense clusters from the same class, i.e. **a** is IsoMap.

3. Consider you binary classification task. For each subtask write a correct formula using x as input, y as true label, p as probability of positive class:

(1) (a) Hyperbolic Tangent activation function _____

(1) (b) Gradient for sigmoid activation function _____

(1) (c) Binary cross-entropy loss for one input _____

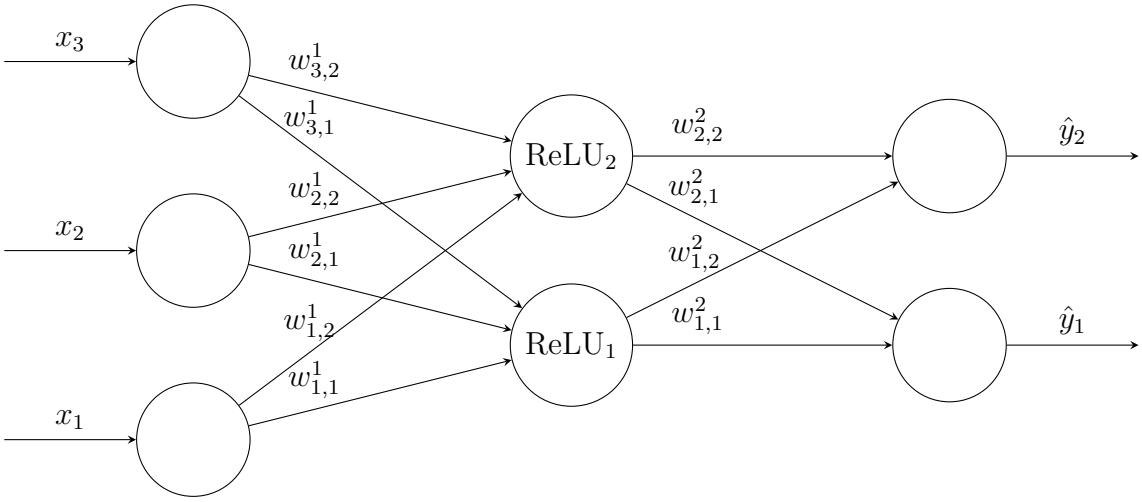
Solution:

$$\tanh x = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

$$l(y, p) = -(y \log p - (1 - y) \log(1 - p))$$

(3) 4. Suppose you have such feed-forward neural network without bias terms. The loss $l(\hat{y}, y) = 1/2(y_1 - \hat{y}_1)^2 + 1/2(y_2 - \hat{y}_2)^2$. Write down the gradient $\dfrac{\partial l(\hat{y}, y)}{\partial w^1_{1,1}}$ only in terms of weights and outputs.



Solution:

$$\begin{aligned}\frac{\partial l(\hat{y},y)}{\partial w^1_{1,1}} &= \frac{\partial l(\hat{y},y)}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial w^1_{1,1}} + \frac{\partial l(\hat{y},y)}{\partial \hat{y}_2}\cdot\frac{\partial \hat{y}_2}{\partial w^1_{1,1}} \\ &= \frac{\partial l(\hat{y},y)}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial \mathrm{ReLU}_1}\cdot\frac{\partial \mathrm{ReLU}_1}{\partial w^1_{1,1}} + \frac{\partial l(\hat{y},y)}{\partial \hat{y}_2}\cdot\frac{\partial \hat{y}_2}{\partial \mathrm{ReLU}_1}\cdot\frac{\partial \mathrm{ReLU}_1}{\partial w^1_{1,1}} \\ &= \frac{\partial \mathrm{ReLU}_1}{\partial w^1_{1,1}}\cdot\left(\frac{\partial l(\hat{y},y)}{\partial \hat{y}_1}\cdot\frac{\partial \hat{y}_1}{\partial \mathrm{ReLU}_1} + \frac{\partial l(\hat{y},y)}{\partial \hat{y}_2}\cdot\frac{\partial \hat{y}_2}{\partial \mathrm{ReLU}_1}\right)\end{aligned}$$

Now we can find separate parts independently:

$$\frac{\partial l(\hat{y},y)}{\partial \hat{y}_1} = \hat{y}_1 - y_1 \qquad \frac{\partial l(\hat{y},y)}{\partial \hat{y}_2} = \hat{y}_2 - y_2$$
$$\frac{\partial \hat{y}_1}{\partial \mathrm{ReLU}_1} = w^2_{1,1} \qquad \frac{\partial \hat{y}_2}{\partial \mathrm{ReLU}_1} = w^2_{1,2}$$

And for the gradient for ReLU:

$$\frac{\partial \mathrm{ReLU}_1}{\partial w^1_{1,1}} = x_1 f(x_1 w^1_{1,1} + x_2 w^1_{2,1} + x_3 w^1_{3,1})$$

where $f(s) = 0$ if $s < 0$ and $f(s) = 1$ if $s \geq 0$

Plugging in all the values we get:

$$\frac{\partial l(\hat{y},y)}{\partial w^1_{1,1}} = x_1 f(x_1 w^1_{1,1} + x_2 w^1_{2,1} + x_3 w^1_{3,1})((\hat{y}_1 - y_1)w^2_{1,1} + (\hat{y}_2 - y_2)w^2_{1,2})$$

5. You are given a dataset from the contest on credit fraud detection. This dataset is of size 284807 samples and 31 columns. The following columns are present:

    1. **Time**
    2. **V1-V28** – are 28 components from PCA decomposition of original large set of features. Due to confidentiality, names or other information on the original features cannot be obtained.
    3. **Amount** – dollar value of transaction
    4. **Class** – binary variable which takes value of 1 when transaction is fradulent and 0 otherwise

    Your task will be to detect fraud. Answer the following questions:

(1) (a) In this case PCA was used for dimensionality reduction. Explain briefly what is dimensionality reduction and write at least two of the other possible dimensionality reduction methods, that might have been applied to get stable components. Outline the main features of each of them and explain your choice.

During exploration of the dataset, you find out that out of 284807 samples only 492 are fradulent.

(1) (b) What kind of classification problem is this? What methods will you use to overcome this issue manipulating the data (write at least 2 and explain)?

(1) (c) What metrics are applicable to measure the quality of classification in this case? (write at least 2 and explain)

(1) (d) Assume you have applied methods above, write down 2 possible classification algorithms you might use, outline specificity of each method and what will you use to determine the best method.

Solution:

(a) Dimensionality reduction is a method used in machine learning to reduce the number of input variables by obtaining a set of principal variables that capture the most significant information from the original dataset. This is used to simplify the model, reduce computation time, and help alleviate the "curse of dimensionality". Two other possible methods are MDS and Isomap, which tries to save the pairwise distance between points in initial and reduced space. MDS uses some defined metric like euclidian. Isomap computes firstly neighbourhood graph and uses geodesic distance.

(b) This is the problem of unbalanced classification, where the most important class to detect is underrepresented in the data. One approach to the problem will be to resample a dataset to soften or remove class imbalance. Firstly, we can randomly over or undersample the data, to increase relative representation of the minor class. These methods are fast and easy to implement, but undersampling can discard potentially important data while oversampling can lead to overfitting due to the replication of minority class data.

Another resampling technique is SMOTE, which generates synthetic data points for underrepresented classes. It helps create a more generalized decision boundary by generating diverse examples, potentially improving model performance on unseen data. However, it can introduce artificial noise to the minority class and can be computationally intense.

(c) F_1 Score: The harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when you need to balance precision and recall.

ROC-AUC: This metric is used to evaluate the performance at all classification thresholds. It is especially useful for imbalanced classes because it measures the ability of a classifier to distinguish between classes, but not if the imbalance is very strong

(d) Random Forest: it is an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the mode of the classes (classification) of the individual trees. It naturally handles class imbalance by the ensemble nature, learning from the imbalance through the construction of trees which inherently perform a form of sampling.

Gradient Boosting Machines (GBM): it builds an additive model in a forward stage-wise fashion and it allows for the optimization of arbitrary differentiable loss functions. It can be equipped with techniques like weighted updates, focusing more on the minority class during the learning process to address class imbalance directly.

To determine the best approach I will apply cross-validation technique using f1 metrics