Name: ____

Matriculation number: ______________________________

Mockup Exam Machine Learning

University of Freiburg, Department of Computer Science

Dr. Josif Grabocka, Dr. Michael Tangermann

Question	Points
1 Short Questions	/10
2 Multiple Choice	/20
3 Linear and Logistic Regression	/10
4 Linear Discriminant Analysis	/5
5 Support Vector Machines	/5
6 Tree-based Methods	/5
7 Algorithm Independent Principles	/5
Total	/60

Short Questions	a	b	c	d	e
Points (out of 2 each)					

Multiple Choice	a	b	c	d	e	f	g	h	i	j
Points (out of 2 each)										

This mockup exam reflects the structure of the real exam. Please note, that exam questions can be about all topics covered by the course.

In the real exam you can earn a maximum of **60 points** and you have **90 minutes** to answer the questions either in **English or German**.

The exam is closed book: **you are not allowed to use notes, calculators or other devices**.

Please answer questions in the space provided, or on the back of the same sheet if necessary. Put your name and matriculation number on every sheet in case pages get separated.

Your name: ____

Your matriculation number: ______________________________

Your signature: ____

Good luck!

Name: ____

Matriculation number: ____________________________________

Part 1 – Short Questions, Short Answers (10P)

(a) **(2P)** Mark for each of the following methods, if the optimization problem faced during training can be solved analytically or not. (Write "yes" or "no" behind each method.)

- support vector machine (for classification)
- linear discriminant analysis
- logistic regression
- independent component analysis

(b) **(2P)** List exactly two purposes for which principal component analysis (PCA) is widely used:

-
-

(c) **(2P)** Carefully draw a scatter plot of 2D-data, where the two dimensions x_1 and x_2 have a Pearson's correlation of zero, but the two dimensions are still dependent on each other.
(Hint: use a sufficient number of data points in your plot to avoid misinterpretations during scoring)

(d) **(2P)** Algorithm-independent principles: Name **two hyperparameters** and a machine learning model in which each of them plays a role. If a hyperparameter is categorial then provide at least three values. If it is numerical, define the largest possible space of valid values by providing upper/lower bounds.

Hyperparameter 1	Hyperparameter 2
1. Name:	1. Name:
2. Model:	2. Model:
3. Space / Set:	3. Space / Set:

(e) **(2P) Explain** what sample size disparity means in the context of fair and unbiased machine learning. **Name two real-life ML problems** where fair and unbiased machine learning is specifically important.

Name: ____

Matriculation number: ____________________________________

Part 2 – Multiple-choice Questions (20P)

Please mark correct answers of a multiple-choice question with an ×. Note that none, some or all of the answers can be correct. Per question, you will receive points only, if **all correct answers** and **none of the wrong answers** are marked.

If you realize you marked a box with a wrong answer, strike through **all** boxes of the task and draw new ones to the right side of it.

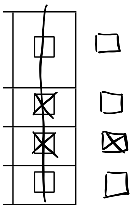

Figure 1: Example of how to correct a wrong answer.

(a) **(2P)** You have a dataset with 10^6 normally distributed and i.i.d. data points, each with $D = 1192$ dimensions and want to calculate a principal component analysis. Which of the following statements is/are true?

Projecting the data to the eigenvector with the smallest eigenvalue retains most of the variance in the data when using only a single principal component.	□
The calculated sample covariance matrix on this dataset has full rank.	□
PCA should not be used because the data is normally distributed.	□
PCA can be used to reduce the number of data points.	□

(b) **(2P)** You want to train a logistic regression to classify a dataset. Which of these functions is/are a sigmoid function?

$S(x) = \frac{e^x}{e^x + 1}$	□
$S(x) = \frac{e^{-x}}{e^x + 1}$	□
$S(x) = \frac{e^x}{e^{-x} + 1}$	□
$S(x) = \frac{1}{e^{-x} + 1}$	□

(c) **(2P)** You have a dataset with 10^3 entries, each entry belonging to one of **five** different classes. Unfortunately, the class labels have been lost. Which of the following statements is/are true?

We can use DBSCAN or k-means clustering to recover the lost class labels.	□
Applying k-means clustering yields a Voronoi tesselation of the dataset.	□
Using k-means clustering with $k = 5$ will provide better results than DBSCAN, because the number of clusters is known.	□
If the underlying (unknown) classes have unequal variance, k-means clustering should be avoided.	□

Name: ____

Matriculation number: ____________________________________

(d) **(2P)** Linear regression can be realized using different regularization approaches. Which of the following statements is/are true?

Lasso can be solved analytically.	□
Ridge regression requires an iterative solution.	□
Lasso is more prone to overfitting compared to ordinary least squares regression.	□
Ridge regression delivers sparse models.	□

(e) **(2P)** Which of the following methods/problems is/are unsupervised?

Independent component analyis	□
Neural network training to discriminate cancer cell from other cells	□
Linear regression	□
k-means clustering	□
Linear discriminant analysis	□
Principal component analysis	□

(f) **(2P)** Deep learning is very popular these days. Which of the following statement(s) is/are true?

Deep Neural networks are a silver bullet and always perform best.	□
Deep neural networks are built of multiple layers of computation.	□
Binary cross entropy is commonly used as a loss function to train neural networks for regression.	□
Neural networks are typically trained using backpropagation.	□

(g) **(2P)** Which of the following method(s)/approach(es) is/are used to improve the generalization error and avoid overfitting?

Using no regularization	□
Using ensembles of models	□
Using dropout for neural network models	□
Collecting more data	□

(h) **(2P)** Hyperparameter optimization is crucial to achieve peak performance.
Which of the following statement(s) is/are true?

Random search can be trivially parallelized.	□
Random search can get stuck in a local minimum.	□
Successive halving successively halves the budget allocated to good configurations.	□
Bayesian optimization uses a model to guide the search.	□

(i) **(2P)** Tree-based methods are popular in machine learning.
Which of the following statement(s) is/are true?

Decision trees can directly handle categorical data.	□
Decision trees are high-variance models.	□
Gradient boosting uses stochastic gradient descent.	□
Fitting a random forest can be parallelized.	□

(j) **(2P)** Feature preprocessing can help to improve performance.
Which of the following method(s) can be used to remove irrelevant features?

Using forward selection	□
Using backward elimination	□
Removing highly correlated features	□
Removing features that correlate with the labels	□

Name: ____

Matriculation number: ____________________________________

Part 3 – Linear and Logistic Regression (10P)

To improve the machine learning course, we want to study the relationship between the scores in the assignments and the exam grades. For each student i of the overall 20 students, our data set contains the average score s_i earned per assignment together with the grade g_i earned in the final exam. This data is visualized in Figure 2.

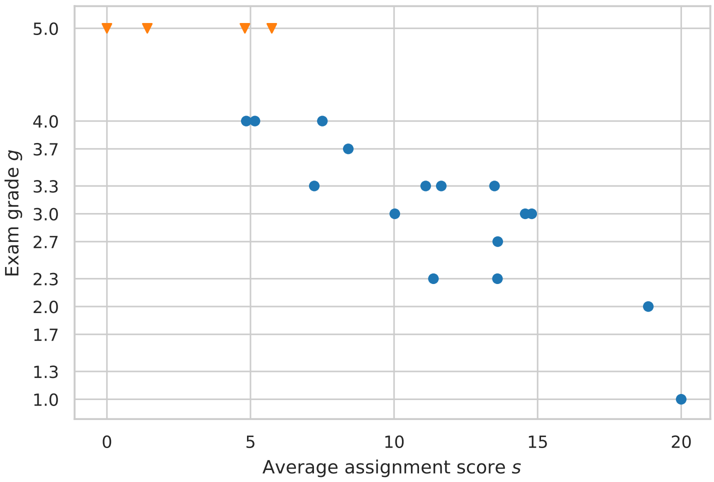

Figure 2: Assignment scores of 20 students vs. exam grade.

(a) **(2P)** Write down the formula of the optimization problem in non-augmented form of a linear regression model minimizing the quadratic loss (also known as L_2-loss) for the above problem. If you use variables that are not defined above, define them.

(b) **(1P)** Assume that you know the parameters b and w of the linear regression model after training. Now an unseen student with an assignment score s_{test} shows up. Write down the formula for predicting whether this student will pass the exam (grade ≤ 4.0) or not according to the linear regression model.

Name: ____

Matriculation number: ____________________________________

(c) **(2P)** Passing or failing is actually a classification problem. Instead of using a linear regression model, a logistic regression can be used to predict if a student will pass or fail and it might even be favorable. Name exactly two advantages of logistic regression compared to linear regression in this context.

-

-

(d) **(1P)** For logistic regression, we propose to use the following model

$$h(s) = \frac{1}{1 + exp(-b - w \cdot s)} \tag{1}$$

where b and w are the free parameters. To find these parameters, name an optimization strategy and the property of the optimal solution.

Optimization strategy:

Property of optimal solution (formula not required!):

Name: ____

Matriculation number: ____________________________________

(e) **(3P)** We have rephrased the problem as a classification problem and relabeled the data with discrete labels, as shown in Figure 3. Fitting a logistic regression model (eq. 1) to this data, our optimizer found the parameters $w = 0.8$ and $b = -4$. Which average assignment score s^* is necessary, such that the model predicts a 50% chance of passing the exam? Write down your ansatz and the solution path.

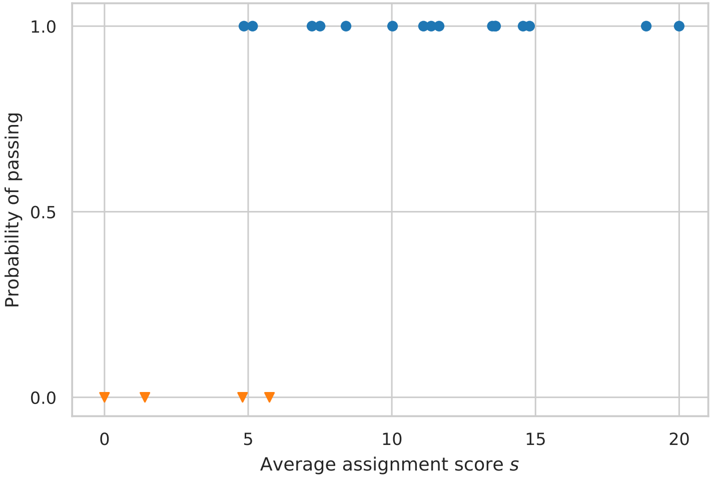

Figure 3: Visualization of the same data, reformulated as a classification problem.

(f) **(1P)** Carefully sketch the logistic decision function for the given parameters $w = 0.8$ and $b = -4$ into Figure 3.

Name: ____

Matriculation number: ______________________________________

Part 4 – Linear Discriminant Analysis (5P)

You are given the following labeled and balanced two class dataset:

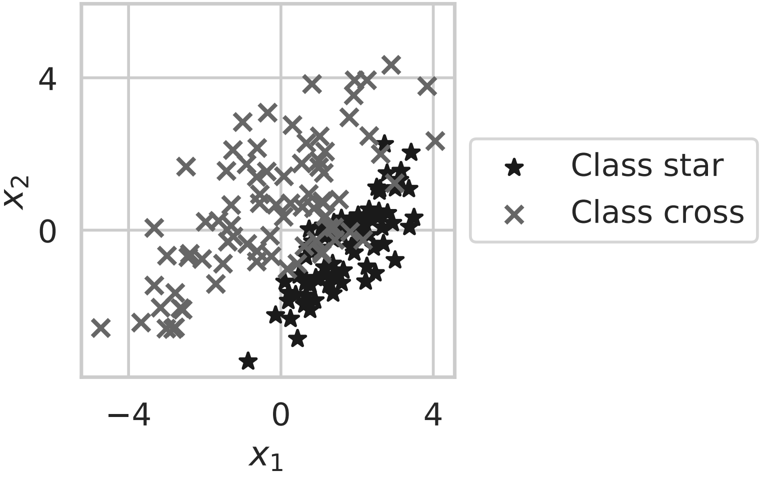

Figure 4: Two-dimensional dataset with two classes.

(a) **(2P)** You intend to train an LDA to classify the dataset shown in Figure 4. Write down the equations how to analytically find the weights $\mathbf{w}$ and bias b of an LDA classifier given the total within-class covariance matrix Σ of the features $\mathbf{x} \in \mathbb{R}^2$ and their respective class means $\boldsymbol{\mu}_{\text{cross}}$ and $\boldsymbol{\mu}_{\text{star}}$.

$\mathbf{w} =$

$b =$

Name: ____

Matriculation number: ______________________________________

(b) **(2P)** An LDA model was trained on the data shown in Figure 4. The model has the following parameters: $\mathbf{w} = [-2.2, 2.0]^T$ and $b = 1.4$.

Now you are given three new data points p_a, p_b and p_c:

Data point	x_1	x_2
p_a	2.5	2
p_b	1	3
p_c	1	-0.5

Calculate the classifier outputs of the LDA (only numeric values, before taking the sign) for the three new data points. Indicate your solution path.

(c) **(1P)** Which assumption of LDA is violated for this dataset? Name **one** alternative classifier that can cope with such small dataset and this violation.

Name: ____

Matriculation number: ________________________________

Part 5 – Support Vector Machines (5P)

In the toy dataset (see Figure 5), linearly separable training data points $\boldsymbol{x} \in \mathbb{R}^2$ are given.

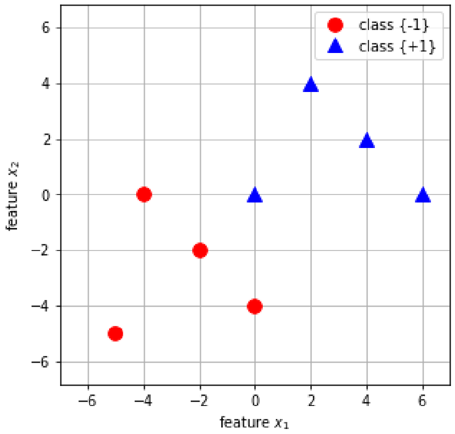

Figure 5: Toy dataset for SVM.

(a) **(1P)** Highlight all support vectors for the optimal linear SVM solution in Fig. 5.

(b) **(2P)** What is the leave-one-out cross-validation error (provide the number!) that you would get on the given data set in Fig. 5? Which data point(s) cause this error? Why?

(c) **(1P)** SVM is an instance-based model. What does this mean? Give exactly one example of a non-instance based model class.

(d) **(1P)** Considering the dual formulation of a trained SVM model, that shall be deployed on a device with very little storage: Which parameters of this formulation determine the minimal set of training data points that *must* be stored?

Name: ______________________________

Matriculation number: ______________________________

Part 6 – Tree-based Methods (5P)

(a) **(1P)** For a random variable X with K possible values $v_1, \dots, v_K$ taken with respective probabilities $p_1, \dots, p_K$, give the equation to compute the entropy $H(X)$.

(b) **(1P)** Consider a split in a random forest at a node with N points into two child nodes with N_l points and N_r points, respectively. Let X denote the random variable following the empirical class distribution of the N data points in the node, and X_l and X_r the respective class distributions in the two leaves. Give the equation for the information gain I achieved by the split. You can use $H(A)$ to denote the entropy of a random variable A.

(c) **(2P)** Consider the following dataset, where the features are *color* and *shape*, and the labels are defined by the *buy* attribute. You want to construct a decision tree to classify whether to buy an item or not.

color	shape	buy
yellow	oval	Yes
blue	oval	Yes
yellow	oval	No
blue	round	No

x	$\frac{1}{5}$	$\frac{1}{4}$	$\frac{1}{3}$	$\frac{2}{5}$	$\frac{1}{2}$	$\frac{3}{5}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{4}{5}$	1
$log_2(x)$	$-\frac{9}{4}$	-2	$-\frac{3}{2}$	$-\frac{4}{3}$	-1	$-\frac{3}{4}$	$-\frac{1}{2}$	$-\frac{2}{5}$	$-\frac{1}{3}$	0

- **Calculate** the initial entropy and information gain for splitting on the feature *shape*. Use the rounded logarithmic values provided in the table.
- **Name** one other split criterion that can be used instead of information gain to construct a classification tree?

(d) **(1P)** An ensemble of estimators can further improve performance, e.g. combining randomized decision trees. Name **two** ways of how to randomize decision trees.

(a)

(b)

Name: ____

Matriculation number: ____________________________________

Part 7 – Algorithm Independent Principles (5P)

Your friend works at a company that develops spam detection systems for emails. A classification algorithm is trained on features extracted from messages to predict whether this message is spam or not. Now, your friend wants to evaluate which of two methods yields a lower generalization error. To allow for a fair comparison, your friend optimizes the hyperparameters for each method. This is the pseudo-code of the implementation:

```
[...]
X, y = loadData()
models = (ma, mb)
loss_function = rmse

# Preprocess data (replace NANs, select informative features
# and normalize data)
preprocessor.fit(X, y)
X, y = preprocessor.transform(X, y)

# Split data
X_train, y_train, X_test, y_test = splitData(X, y)

results = []
for m in models:
    # opt_hypers returns configuration with lowest value of
    # loss_function on y_test
    opt_conf = opt_hypers(m, X_train, y_train,
                          X_test, y_test, loss_function)
    y_pred = m.predict(X_test, model_conf=opt_conf)
    loss_test = loss_function(y_test, y_pred)

    # log results
    results.append((m, loss_test))

# report model with lowest estimated generalization error (loss_test)
[...]
```

(a) **(4P)** The implementation contains several problematic decisions in the context of this task. Find **two** conceptual issues (excluding syntax errors and undefined variables) and explain briefly what should be changed.

1.

2.

(b) **(1P)** Your friend's company develops a new system that estimates if images on a website contain advertisements. For that they want to build an end-to-end learning pipeline. Which model class would be suitable for this task?