UNIVERSITY OF FREIBURG

Department of Computer Science

Dr. Joschka Bödecker, Dr. Frank Hutter, Dr. Michael Tangermann

Mockup Exam Machine Learning, Summer Term 2017

Question	Points
1. Short Questions	/20
2. Principal Component Analysis	/10
3. Backpropagation	/10
4. Algorithm-Independent Principles	/10
5. Decision Trees	/10
6. Linear Regression	/10
Total	/70

Short answer question	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. The questions can cover all topics discussed in the course, and you should by no means focus on the topics given here.

In the real exam there will be a maximum of **70 points** as well and you have **90 minutes** to answer the questions. You can write your answers either in **English or German**.

This examination is closed book: **you are not allowed to use any notes or calculators**. Please answer questions in the space provided, and if necessary continue your answers on the back of the same sheet. Please also put your name and matriculation number on every page, in case pages get separated.

Your name: _____

Your matriculation number: _____

Your signature: _____

Good luck!

Part 1 – Short questions, short answers

(a) **(2P)** Name two possible reasons to use linear regression in contrast to a non-linear fit.

(b) **(2P)** Consider the sample covariance matrix (covariance matrix estimated on data) in LDA. What do you commonly observe for the largest and smallest eigenvalues of the matrix in a high-dimensional space with few data points?

(c) **(2P)** Consider the following statement:

"The function $\mathbf{k} : \mathbb{R}^N \text{ x } \mathbb{R}^N \rightarrow \mathbb{R}$ with $\mathbf{k}(x, y) = \frac{x^T y}{||x|| \cdot ||y||}$ defines a positive definite kernel function." Is this true or false? Discuss why.

(d) **(2P)** What is the purpose of the VC-dimension in the Vapnik-Chervonenkis inequality?

(e) **(2P)** Describe the main idea behind Backpropagation Through Time (BPTT) for training recurrent neural networks.

(f) **(2P)** Consider applying valid convolutions to an RGB image with 24 x 24 pixels. You use 6 filters of size 5x5 with a stride of 3. What are the output dimensions (width, height, depth)?

(g) **(2P)** Explain two advantages and two disadvantages of Ensembles of Classifiers (e.g., Bagging, Boosting algorithms).

(h) **(2P)** Is there an algorithm that performs best on all datasets? If yes, which one is it. If no, explain why not.

(i) **(2P)** When estimating the performance of a classifier, one can revert to methods such as k-fold cross-validation and holdout set. Name two advantages of both methods over the other.

(j) **(2P)** Explain in words what the Leave-One-Out procedure does. When do we use it?

Part 2 – Dimensionality reduction with PCA for classification (10P)

Your task is to write pseudocode for a classification problem. The script should output the estimated classification accuracy on unseen data. Provided are the matrix $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, containing the dataset (N number of samples and d number of features, standardized with zero mean and unit variance), and vector $\boldsymbol{y} \in \{0,1\}^N$, containing the corresponding labels. Several functions are available for its use in the pseudo-script:

`computePCA`
Input:
$\boldsymbol{C_x} \in \mathbb{R}^{d \times d}$, covariance matrix of the original data.
Output:
$\boldsymbol{\lambda} \in \mathbb{R}^d$, vector containing the eigenvalues of $\boldsymbol{W}$, sorted in descending order
$\boldsymbol{W} \in \mathbb{R}^{d \times d}$, matrix containing the corresponding eigenvectors of $\boldsymbol{C_x}$.

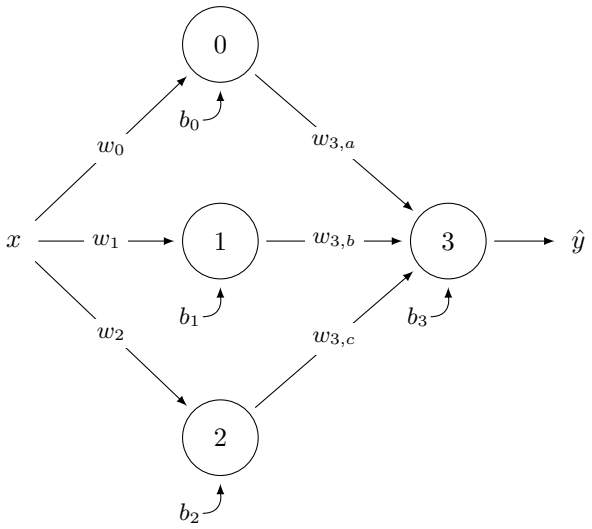
`trainClassifier`
Input:
$\boldsymbol{X} \in \mathbb{R}^{N \times d}$, matrix containing the dataset to train the classifier (training set).
$\boldsymbol{y} \in \{0,1\}^N$, vector containing the corresponding labels for each row of $\boldsymbol{X}$.
Output:
Model: Trained classifier.

`applyClassifier`
Input:
Model, classifier model as returned by `trainClassifier`.
$\boldsymbol{X} \in \mathbb{R}^{N \times d}$, dataset on which the labels will be predicted (test set).
Output:
$\boldsymbol{y} \in \{0,1\}^N$, predicted labels.

(a) **(9P)** On the next page, write pseudo-code solving the scenario described above. Include a step of dimensionality reduction using PCA. Choose the abstraction level of your pseudo-code such that your script contains a maximum number of 15 lines.

(b) **(1P)** Explain your strategy to select the number of PCA components to perform the dimensionality reduction step.

Part 3 – Backpropagation (10P)



Consider the neural network above with weights w_i and biases b_i pictured above. Determine the symbolic expressions for the forward pass **(3P)** and the gradients $\frac{\partial L}{\partial w_0}$, $\frac{\partial L}{\partial w_1}$, $\frac{\partial L}{\partial b_0}$, $\frac{\partial L}{\partial b_1}$ with the backward pass **(7P)**!

Use the squared error loss function $L(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2$. You can reuse previously defined expressions, e.g. you do not need to expand $\frac{\partial L}{\partial \hat{y}}$ again, if you evaluated it previously.

Units 0, 1 and 2 have the rectified linear activation function h_{relu}, while unit 3 is activated linearly with h_{linear}. You are not required to evaluate their derivates and can denote them as h'_{relu} and h'_{linear} respectively. Please use a_i to denote the activation of unit i before applying the activation function, and z_i for its result.

Part 4 – Algorithm Independent Principles (10P)

Your friend works at a company that develops weather prediction systems. There is one data point per day, and predictions shall be made for the next days. Now your friend wants to try out different types of models and evaluate their performance using cross-validation.

1. **(2P)** Are the individual data points in your friend's data independently distributed? Why or why not?

2. **(2P)** What problems would occur when using standard k-fold cross-validation for this data?

3. **(6P)** Explain which changes are necessary to the following regular cross validation code (words are enough, you don't need to correct the code) to suit your friend's time series data.

```
num_points = len(X)
num_folds = 10
# Shuffle data to make sure there are no sorting artifacts
indices = np.shuffle(range(num_points))
X = X[indices]
y = y[indices]

fold_indicators = [i % num_folds for i in range (len(X))]
fold_indicators.sort()

for fold in range(num_folds):
    X_train = X[fold_indicator != fold]
    y_train = y[fold_indicator != fold]
    X_valid = X[fold_indicator == fold]
    y_valid = y[fold_indicator == fold]

    model.fit(X_train, y_train)
    scores.append(model.score(X_valid, y_valid)
```

Part 5 – Decision Trees (10P)

(a) **(2P)** Explain the decision tree algorithm in words. Give the definition of entropy, expressed in terms of $p(v_k)$ where v_k is the number of data points that have value k for attribute v. Think of what you would do in the following three distinct cases that can happen when deciding on a split point in a node:

- All instances have the same class label, but various attribute values
- All instances have the same value for all attributes, but various class labels
- The instances have various class labels and various attribute values

(b) **(1P)** Consider the following dataset. Explain why the ID column should not be considered while building your model.

ID	PK	speed	color	buy
1	High	Fast	Blue	Yes
2	High	Slow	Blue	No
3	Low	Fast	Red	Yes
4	Low	Fast	Blue	No

(c) **(5P)** Execute the decision tree algorithm on the given dataset, where the labels are defined by the *buy* attribute. You should ignore the ID attribute. Calculate the entropy and information gain for the split in the root node. Use the rounded logarithmic values provided in the table. For all other nodes, you don't need to explicitly calculate the entropy and information gain, as you can infer the best possible split according to this criterion.

$log(1)$	0
$log(3/4)$	-0.4
$log(2/3)$	-0.6
$log(1/2)$	-1
$log(1/4)$	-2
$log(1/3)$	-1.6

(d) **(1P)** Is there a decision tree that has a smaller depth than this one? Draw this one. Why did the decision tree algorithm as executed in the previous question not generate this tree?

(e) **(1P)** Name two differences between Bagging Decision Trees and Random Forest

Part 6 – Linear Regression (10P)

Farmer Huber has cattle near the Feldberg (Black Forest). He is interested in predicting the weather to know in advance when to lead his animals down to the valley for the winter season. From the German weather service, he downloaded the data containing the daily maximal temperature of the last 500 days. In his notation, the time is called x_i and the temperatures are called y_i for $i = 1 \dots 500$. He has read about linear regression as a model to predict the continuous outcomes for future time points. He also knows that a simple one-dimensional linear regression model could look like $f(x) = w \cdot x + b + \epsilon$ where ϵ is an error term.

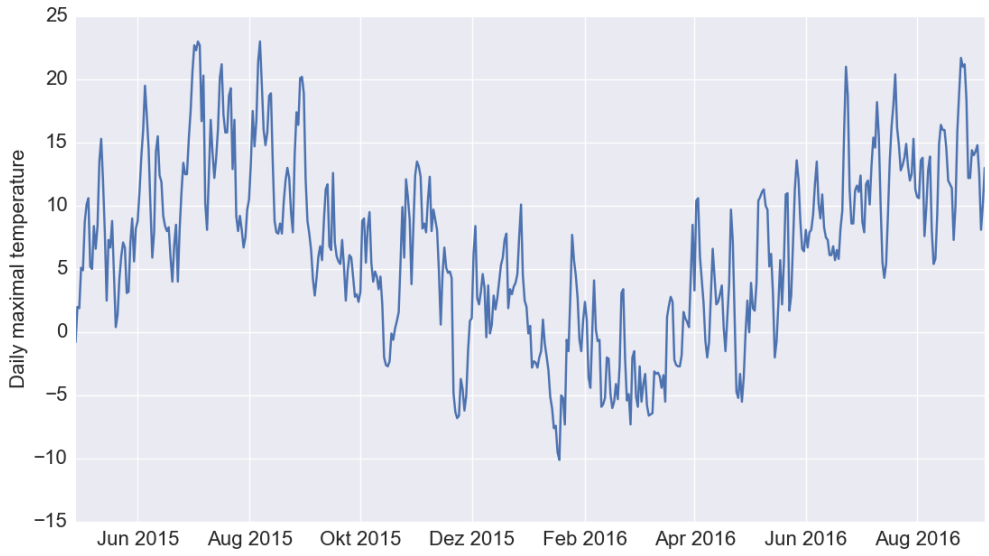


Figure 1: Daily maximum temperature at Feldberg (Black Forest).

(a) **(2P)** Write down the objective function of a linear regression model minimizing the quadratic loss (also known as L_2-loss).

(b) **(2P)** Now, he downloaded a software to fit a linear regression model. The program fits the data and delivers the outputs $b = 5.69$ and $w = 3.48 \cdot 10^{-03}$. Please help farmer Huber by providing an explanation of the temperature development *according to the model* and add the linear fit to Figure 1.

(c) **(2P)** However, you as an expert are not convinced by the results. To further investigate the model, you plot the residuals $f(x_i) - y_i$ for $i = 1 \dots 500$ given in Figure 2. By looking at these

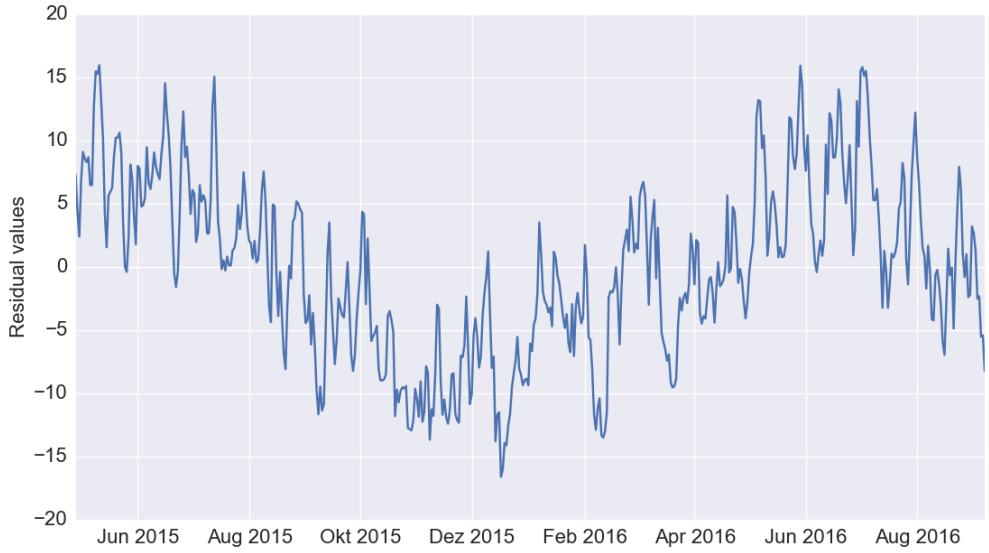


Figure 2: Residual values obtained by the linear fit.

residuals, which assumption of the linear regression model is violated? What is the reason for this violation?

(d) **(2P)** In general, linear regression can be extended by adding a regularization term $\alpha^2(b^2 + w^2)$ to the objective function, where α is a parameter determining the strength of the regularization. What is the effect of this additional term and why might it be beneficial?

(e) **(2P)** Given your background in machine learning, how would you advice Mr. Huber to obtain a more accurate prediction?