# IML: Exercises Weeks 4

September 27, 2021

## 1 Introduction

This document contains the exercises for week 4 of the IML course. The exercises are split into three parts: revision, homework and practicum. The only questions you have to hand in are the homework questions. You do this every 3 weeks. Note that we will select a subset of the questions to be graded. If anything is unclear or you have any questions, please feel free to use the forum on Brightspace or send an email to the TAs.

## 2 Supervised Learning 2 - week 4

### 2.1 Revision questions

1. What is the goal of classification within the context of our lecture?
2. What will a linear predictor look like in a 3-dimensional cube and a 4-dimensional space, respectively?
3. What is a linear program?
4. What is the update rule for weight $w$ of a perceptron at iteration $t$ given with a misleading sample $x_i$ and its corresponding true label $y_i$?
5. How does the Iterative Dichotomizer 3 algorithm (ID3) decide which feature to focus when generating new nodes?
6. Can you name a feasible solution to handle missing values for a categorical feature in a decision tree?
7. Can decision trees exploit linear structures within data?

### 2.2 Homework questions

#### 2.2.1 Exercise 1

To transform multiclass regression into multiple binary-class tasks, we identified two main methods: one vs one and one vs all. How many binary classifiers need

| Index | Transportation | Group Size | Insurance | Impression (labels) |
|---|---|---|---|---|
| 1 | Train | Medium | No | Good |
| 2 | Car | Large | No | Bad |
| 3 | Train | Large | No | Bad |
| 4 | Car | Medium | Yes | Bad |
| 5 | Train | Small | Yes | Good |
| 6 | Plane | Large | No | Bad |
| 7 | Train | Medium | Yes | Good |
| 8 | Plane | Small | Yes | Good |

Table 1: Data for Exercise 2

to be trained for each of these two methods, assuming $n$ original classes? Does this mean that one of these methods will always be faster than the other? Clearly motivate your answer.
**Hint**: how many samples are in your training set for training the classifiers in these two modes?

### 2.2.2 Exercise 2

Recall the famous Iterative Dichotomizer 3 algorithm (ID3) that was discussed in the lecture. The ID3 algorithm is designed to select the feature that can maximize the information gain at current stage as its next splitting criteria for a decision tree.
**To read**: the information entropy for a dataset $D$ can be defined as:

$$IE(D) = -\sum_{k=1}^{N} p_k \log_2(p_k),$$

where $p_k$ is the portion of data that belongs to the $k$th class (category) to the entire data and $N$ is the number of class in the data.

Assume dataset $D$ is divided based on values on a specific categorical feature $f$. This results in multiple non-overlapped subsets of D, namely, $D^f$. $D^{f_v}$ is one of these subsets which only contains data that values $f_v$ on feature $f$. Thus, the information gain after splitting can be defined as follows:

$$IG(D, f) = IE(D) - \sum_{v=1}^{V} \frac{|D^{f_v}|}{|D|} IE\left(D^{f_v}\right),$$

where $V$ is the number of distinct values on feature $f$ in dataset $D$, $|D|$ and $|D_v^f|$ represent the number of samples in $D$ and $D_v^f$, respectively.

**To answer**: in this exercise, we would like you to demonstrate the selection of feature step by step on a piece of dummy data, which is shown in **table 1**. The task is to use historical reviews to help a travel agency determine whether customers will enjoy their new travel plans or not. Suppose you are required to apply ID3 on the dataset to build a decision tree, please answer the following questions:

1. What is the information entropy for the dataset based on its label?
2. What are the information entropies for the dataset if we divide it based on means of transportation?

   **Hint**: the splitting operation will result in three sub-nodes (by cars, by trains and by planes) and each of them shall have an entropy accordingly.
3. What is the information gain if we divide the dataset based on means of transportation?

   **Hint**: $IG(D, Transportation)$
4. What are the information gains if we divide the dataset based on group size and whether the agency provides insurance, respectively?

   **Hint**: $IG(D, Group\ Size)$, $IG(D, Insurance)$
5. Which feature (among transportation, group size and insurance) shall we choose to be the next (first) splitting criteria if we follow the rules of ID3?

### 2.2.3 Exercise 3

In the lecture we have mentioned that a binary decision tree is equivalent to a many-child decision. Prove this by creating a transformation method between the two types of trees.

## 2.3 Practicum questions

### 2.3.1 Exercise 4

Recall the Iris dataset that we introduced in week 2's exercises, where you are asked to use logistic regression to classify different plants. The data contains records of 150 iris plants. Each record features 4 numerical attributes (real numbers). More information about the data can be found in the user guide of **load_iris** API on Scikit-learn official site. Or on its **UCI source**. In this exercise, you will get familiar with the decision tree classifier provided by scikit-learn through applying it on the Iris data.

- Can you separate training and test data from the original data set?
- Do you need to process the data before applying logistic regression and decision tree on it?
- Can you build a logistic regression classifier to classify Iris plants on training data?
- Can you build a decision tree classifier to classify Iris plants on training data?

- What are the performances of these two algorithms on test data, respectively? (accuracy, precision and recall)
- Which one of the two algorithms do you think suit the task better?

#### 2.3.2 Exercise 5

Handwritten image classification is a famous historical topic within the field of image processing. In this exercise, please try to load the *digits* dataset using scikit-learn API and apply decision tree classifier on this dataset.

The data set contains 1797 images of handwritten digits from 10 classes (0,1,2...9). Each sample in the dataset is a 8x8 image of only one digit. You can load the dataset easily via **sklearn.datasets.load_digits**. More information about the data can be found in user guide of **load_digits** on scikit-learn.

[**Optional**] Decision tree also works (but not perfectly) for another famous handwritten digits dataset, the MNIST. You can easily and freely access the MNIST data on Internet. Please try to find and (down)load the data and apply decision tree on it. A simple decision tree can indeed produce solid results.

#### 2.3.3 Exercise 6

In our previous lectures, we have mentioned loss (cost) functions for several times. When we train a machine learning model (classifier or regressor), we firstly set up a loss function and minimize the loss through learning algorithms. Theoretically, this process is a kind of empirical risk minimization.

However, the paradigm for ID3 to train a decision tree is different. A typical loss function is absent throughout the whole training stage. Do you think the ID3 algorithm for decision tree classifier still follows the empirical risk minimization rule? Please motivate your answer.

**Hint**: As a starting point, you can think of the definition of empirical risk and the reason why learners of decision tree iteratively split the dataset.