Machine Learning

Exercise Sheet 1

Winter Term 2023/2024
Prof. Dr. Niels Landwehr
Dr. Ujjwal

Available: 03.11.2023
Hand in until: 09.11.2023 11:59am
Exercise sessions: 13.11.2023/15.11.2023

How to submit your solutions

Please submit your solution to this exercise sheet until 09.11.2023 at 11:59am (strict deadline). Solutions can be submitted only in electronic format through the tutorial Learnweb. You will find upload links for the five different tutorial groups there, please make sure you submit your solution to the correct tutorial group to which you have been assigned.

Task 1 – Problem Settings in Machine Learning [15 points]

Consider the following example application domains for machine learning:

1. Infer whether a text-based review of a product is negative, neutral, or positive
2. In a computer network, detect unusual traffic that may be indicative of an attack against the network
3. Predict the selling price of an apartment or house based on features describing the property and its location
4. Learn to automatically set the bitrate for a video stream viewed by a user at any point in time based on information about the current connection quality, state of the video buffer, location of the user etc.
5. Predict whether or not an advertisement shown on a web page will be clicked by a user, based on information about the user, ad, and web page.
6. Learn to generate photorealistic images of human faces
7. As an email service provider, detect groups of similar email messages that may constitute a spam campaign
8. Learn to play a real-time strategy computer game such as *StarCraft*

For each domain, determine which kind of machine learning problem setting it corresponds to and explain why. Use the list of problem settings from the slide 'Summary: Machine Learning Problems' in the introductory lecture. If an example could be formalized as different problem settings depending on the kind of data that is available, please list and explain all possible choices.

Task 2 – Feature Representations [15 points]

For each of the example application domains for machine learning listed in Task 1, briefly discuss which kind of features could be used to represent instances (for supervised and

unsupervised learning) or to represent the state of the environment (for reinforcement learning).

Task 3 – The IID Assumption in Machine Learning [10 points]

In supervised machine learning, we usually assume that training examples are drawn independently from a fixed distribution over inputs and outputs, $(\mathbf{x}_n, y_n) \sim p(\mathbf{x}, y)$. Crucially, we also assume that the data the model will encounter at application time, that is, after it has been deployed, also comes from the same distribution, $(\mathbf{x}_{new}, y_{new}) \sim p(\mathbf{x}, y)$.
For each of the following example application scenarios for machine learning, comment on why the latter assumption might not be fully justified. That is, in how far could the distribution of data at application time deviate from the distribution of the training data?

1. We are using machine learning to build a spam filter. Training data is obtained by collecting spam and non-spam emails from various sources over a period of one year.

2. We use machine learning to build a simple speech recognition system that can recognize spoken commands (in English). We train the model using audio data recorded from a large set of speakers recruited through an ad in the local newspaper. The model will be deployed worldwide.

3. We use machine learning to build a model that will predict future stock prices of companies based on features describing the company and historical data about stock price movements.

4. We use machine learning to build a system that can detect roads, buildings, cars and pedestrians from images of a forward-facing camera in a car with the goal of developing an autonomous vehicle. We collect training data through a three-month period by driving thousands of kilometers through real-world traffic in two different states in the US.