Master's thesis

Daniel Kari

Contents

Chapter 1

Introduction

The use of statistical methods in sports is a continuosly evolving field of study.

Chapter 2

Machine learning

2.1 Introduction

Machine learning can be broadly defined as a process in which a machine changes its structure to improve expected future performance based on past experience [1]. It is commonly used when a problem cannot be solved explicitly or when creating such solution is not feasible. This means the solutions provided with machine learning are never perfect, except for trivial cases. In all other occasions one needs to settle for 'good enough'.

When machine learning is applied in practice, the first step is to specify the task one wants to solve. This might for example be detecting whether or not an image contains a dog. After that suitable data is gathered and possibly labeled manually. We might also want to determine which variables in the data are most relevant for the task in question. This phase might include formal process of *feature selection*, which can be considered as a field of its own. Quality and quantity of data are essential to gain meaningful results [2].

The learning structure, also called learner, then receives subset of the data for *training*. In *supervised learning* the data consists of input-output pairs. The goal is to predict the output, which might integer or real-valued depending on the task, based on the input for unseen data points, also called *test set*. When the output is an integer and number of output values is in some sense limited, the task is called *classification*. Detecting dogs in images is a good of example of simple classification, since the image either has a dog in it or it does not. The two alternatives can be encoded as zeroes and ones and thought as a binary classification problem. By contrast, in *regression* problems the output is a continuous numerical value. The difference between the tasks affects for example how the quality of prediction is measured.

The most fundamental objective in machine learning is *generalization* [2]. In supervised setting, the learner is trained with finite sample of labeled data, utilizing some *function* in prediction. After training, the learner should be able to predict the output for separate test data as accurately as possible, that is to say generalize well.

If the function is complex, it might be able to predict the labels for training sample perfectly of with very little error. However, this does not guarantee the prediction will be accurate for different sample of the data. The function is *overfitting*, if it models random noise in the training sample, resulting in decreased prediction accuracy for test data. Correspondingly the function can also be too simple and *underfit*, if it fails to detect patterns present in the data.

In order to avoid underfitting and overfitting, we need to establish some sort of model evaluation criteria.

2.2 Formal definition

train test split

Free lunch

2.3 Model evaluation

2.4 Cost function

2.5 Multilayer Perceptron

Chapter 3

Prediction

3.1 Background

Predicting the outcome of an NHL game has been generally considered most difficult out of the four major North American leagues.

3.2 Variables affecting prediction

Bibliography

[1] Nils J. Nilsson: Introduction to Machine Learning, Stanford University, 1998.

[2] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar: Foundations of Machine Learning, MIT Press, Second Edition, 2018.