Towards impressive titles

Tobias Axelsson

26 mars 2018

Acknowledgements

I am a student blalsadf

Abstract

asdasd

Sammanfattning

asdasdasd

Contents

Chapter 1

Introduction

The chapter starts with a background describing why road condition monitoring is important and who Trafikverket are, how road condition data is collected today and why the technology behind it needs improvement. An objective for the project is defined followed by its delimitations. Lastly, a thesis structure is presented to simplify navigation through different parts of the project.

1.1 Background

Living in cold areas of the world usually means work for invididual people, municipalities and companies in trying to maintain a non-winter-like infrastructure. This of course, also involves winter road maintenance. Salting and plowing roads is an investment in not only saving lives, but also in lowering socio-economic costs: In two scenarios on a road with 2 cm snow and a daily traffic flow of 2000 vehicles, one with a salted and ploughed road taking four hours to drive, and another scenario on the same road without winter maintenance taking five hours to drive. The total socio-economic costs are 3.5% higher in the non-maintained road, mainly due to increased travel time and thus higher accident costs [1].

Despite the socio-economic savings in performing winter road maintenance, it still represents a notable economic cost. Trafikverket, the agency in charge of road state road maintenance in Sweden, reported that winter road maintenance were roughly 18% of the total road maintenance costs in 2013 [2]. Local contracters are hired to carry out the plowing and salting of state roads, with requirements on both ends regarding when to plow, which roads to prioritize etc. Trafikverket has over 800 Road Weather Information Systems (RWIS)(Fig. 1.1) distributed across state roads in Sweden which are used by contracters to carry out winter road maintenance work [3].

Figure 1.1: RWIS Station at sensor site Myggsjön [4].

Table 1.1 shows

Operation	worst-case cost	time complexity
Insert x into l_i	2	$O(1)$
Update $count_i$	1	$O(1)$

asdasd

1.2 Objective

The objective is to determine if a road surface temperature sensor can be simulated with prediction models based on historic data from road weather information systems.

1.3 Delimitations

The project will focus on supervised learning algorithms. There are also unsupervised learning algorithms that are not covered for the following reasons:

- Performance of unsupervised learning algorithms can be difficult to evaluate [5].
- The data provided for this project by Trafikverket is labeled, which means it is suitable for supervised learning algorithms.

See section 2 for more information on machine learning theory.

1.4 Thesis structure

Chapter 2

Literature Review

The chapter gives both general and specific information on theory used for this project. Mathematical statistics, regression and machine learning are covered in the first three sections, providing a general understanding of the field of study. Specific machine learning models are explained in the final three sections of the chapter.

2.1 Machine learning basics

Machine learning as formally defined by Mitchell [6]: "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P if its performance at tasks in T, as measured by P, improves with experience E". This means that machine learning algorithms are used to solve a set of problems, measure its performance in doing so and ultimately improve in some way from previous experiences. For example, imagine a program designed to determine if a human face is in a photo or not. Since photos are taken at different distances, angles and faces have different characteristics such as eye color, skin color, distance between eyes and nose shape, implementing this "manually" may prove cumbersome. Instead of programming an algorithm to recognize faces, it can be programmed *to learn to recognize faces*. If the algorithm is allowed to analyze a dataset with thousands of photos of human faces, it could learn to distinguish a human face by recognizing parts of the face such as eyes, nose, mouth and where those parts are most likely placed to oneanother.

In essence, machine learning algorithms improve/learn in some way from analyzing a dataset. How they learn can be used to broadly categorize machine learning algorithms as either having supervised or unsupervised learning [7]. Supervised learning algorithms processes a labeled dataset while unsupervised learning tries to make sense of an unlabeled dataset [5]. This project does not concern algorithms related to unsupervised learning, as motivated in section 1.3.

2.2 Supervised learning

In supervised learning, the learner (algorithm) receives a dataset of labeled observations which is used to predict correct values for unseen data[5]. A database table storing weather-related data could for example have thousands of records (observations) where data in each record belong to certain columns (features) such as wind speed w_s, wind direction w_d and time t. The goal of supervised learning is to build a mapping function (model)

$$y = f_{map}(x) \tag{2.1}$$

such that when new input data is used, f_{map} is able to predict a correct output value [8]. The model is built from a dataset which is typically split into three parts [9]:

- Training dataset: Used to fit the model.
- Validation dataset: Used to give an unbiased evaluation of a model built from the training dataset and potentially update its hyperparameters. Hyperparameters are model parameters that are used in some learning algorithms. They are usually fixed before the training process begins [10].
- Test dataset: Gives an unbiased evaluation of the final model.

How the dataset "should" be split is brought up in section 2.9.

Supervised learning can be thought of as having a teacher supervising the algorithm. The correct answers are in the training data and the algorithm learns from being corrected by the teacher [8]. Going back to the forementioned example of the weather station to give a brief example of how a supervised machine learning algorithm works: Suppose a training, validation and test dataset is provided and one wishes to predict wind speed $y = w_s$ based on wind direction and time $x = [x_1, x_2] = [w_d, t]$. During the training process, a supervised learning algorithm goes through the training dataset to build a model, as seen in Eq. 2.1, and possibly updated when validated against the validation dataset. Suppose the supervised learning algorithm used is multiple linear regression (see section 2.5.1) and a model is built from the training process:

$$w_s = f_{map}([w_d, t]) = \beta_0 + \beta_1 w_d + \beta_2 t = 4 + 0.2w_d + 1.7t \tag{2.2}$$

The model can then be tested with the test dataset to see how it performs on unseen data.

Estimation of continous output variables, such as wind speed in the example presented above, is a regression problem. In supervised learning there are also algorithms associated with the problem of classification; how to categorize data [11].

2.2.1 Classification predictive modeling

In a classification problem, the computer is asked to place a new observation into one of k categories, $k \geq 2$ [7]. The problem of categorizing new email as spam or not spam is an example of a classification problem. Google claims that their machine learning models can detect spam and phishing messages with 99.9% accuracy in their widely used Gmail application [9].

Another example of a classification problem, one that may well be the first that machine learning novices encounter, is classification of the Iris flower dataset. The dataset consists of 50 observations with four features: length and width of the sepals and petals, in centimeters. Based on this information, the problem is to classify to which of the following categories each observation belongs to [12]:

- Setosa
- Versicolour
- Virginica

How the classification is carried out depends on the algorithm used to build the model. These kind of algorithms are commonly known as classifiers. There are several classifiers that can be used for the Iris dataset, but their performance in doing so may differ. Performance of classifiers are typically measured in accuracy, which is the amount of correct predictions divided by the number of observations in the test dataset [7].

$$\text{accuracy} = \frac{\#\text{correct predictions}}{\#\text{observations}} \tag{2.3}$$

2.2.2 Regression predictive modeling

In contrast to classification problems, such as categorizing incoming email as spam or not spam, regression problems are about predicting continous quantaties. Regression models can have either real-valued or discrete input variables [11]. The model in eq. 2.2 is an example of a regression model since the goal is to predict a numerical value for wind speed. The problem could be translated into a classification problem by, for example stating that for given numerical intervals, the wind speed is categorized as being low, medium or high. This kind of conversion is known as discretization but even if the conversion is useful, it can result in surprising and/or poor performance [11]. This is why both classification and regression modeling are covered in this project because they are useful for different kind of predictions.

Performance of regression models can be measured by computing the mean squared error (MSE) of the model on the test dataset.

$$MSE_{test} = \frac{1}{n}\sum_{i}^{n}(y'_{test} - y_{test})_i^2 \tag{2.4}$$

where y'_{test_i} are predictions on the test and y_{test_i} are are actual values [7]. It's a measurement of how close each prediction was to the target value on average.

2.3 Difference between linear and non-linear models

2.4 Linear classification algorithms

2.4.1 Decision trees

2.5 Linear Regression algorithms

2.5.1 Multiple linear regression

2.6 Non-linear classification and regression algorithms

2.6.1 Neural network

2.7 Overfitting, capacity, hypothesis space

2.8 Multicollinearity

2.9 Dataset split

2.10 Data pre-processing

Chapter 3

Method

The chapter covers strategies and methods used to achieve the objective of the project. Reasons for each choice of method or strategy are motivated and described in the sections, which are ordered chronologically.

3.1 Overview

3.2 Research approach

3.3 Research strategy

3.4 Tools

Chapter 4

Implementation and results

Describe the process of collecting data, training and implementing machine learning algorithms with different methods.

4.1 Data collection

4.2 Neural network

4.2.1 First iteration

4.2.2 Second iteration

Chapter 5

Analysis

Analyze data from the implementation with respect to the objective of the study.

5.1 Neural network

Chapter 6

Conclusions and recommendations

6.1 Conclusions

6.2 Recommendations

Chapter 7

Discussion

7.1 Thesis process

7.2 Validity and reliability

Validity and reliability of the conclusions. Needed?

7.3 Future work

Bibliography

[1] A. Arvidsson, "The Winter Model – A new way to calculate socio-economic costs depending on winter maintenance strategy", *Cold Regions Science and Technology Volume 136, April 2017, Pages 30-36*, 2017.

[2] Trafikverket, *Trafikverkets årsredovisning 2015*, 2016.

[3] ——, (2017). Vinterväghållning, [Online]. Available: `https://www.trafikverket.se/resa-och-trafik/underhall-av-vag-och-jarnvag/Sa-skoter-vi-vagar/Vintervaghallning/` (visited on 02/07/2018).

[4] Pelpet. (2010). Rwis Station at sensor site Myggsjön, [Online]. Available: `https://commons.wikimedia.org/wiki/File:Rwis_station_Myggsjon_01.JPG` (visited on 02/06/2018).

[5] M. Mehryar, R. Afshin, and T. Ameet, *Foundations of machine learning.* Ser. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012, ISBN: 978-0-262-01825-8. [Online]. Available: `http://proxy.lib.ltu.se/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=msn&AN=MR3057769&lang=sv&site=eds-live&scope=site`.

[6] T. Mitchell, *Machine Learning.* McGraw Hill, 1997.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning.* MIT Press, 2016.

[8] J. Brownlee. (2016). Supervised and Unsupervised Machine Learning Algorithms, [Online]. Available: `https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/` (visited on 02/28/2018).

[9] F. Lardinois. (2017). Google says its machine learning tech now blocks 99.9% of Gmail spam and phishing messages, [Online]. Available: `https://techcrunch.com/2017/05/31/google-says-its-machine-learning-tech-now-blocks-99-9-of-gmail-spam-and-phishing-messages/` (visited on 03/19/2018).

[10] X. Amatriain. (). What are hyperparameters in machine learning?, [Online]. Available: `https://www.quora.com/What-are-hyperparameters-in-machine-learning` (visited on 03/20/2018).

[11] J. Brownlee. (). Difference Between Classification and Regression in Machine Learning, [Online]. Available: `https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/` (visited on 03/22/2018).

[12] R. Fisher. (). Iris Data Set, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/iris (visited on 03/19/2018).