Machine Learning Assignment 1

Even Wanvik, NTNU 28/08/2019

1 Theory

1.1 What is *concept learning*? + explain with an example

> "The problem of searching through a predefined space of potential hyptheses for the hypothesis that best fits the training examples" *Tom Michell*

When a human being is learning something, much of it is based on generalized concepts gained from past experiences. For instance, if a human were to identify if a certain type of car, we differentiate the betwen the cars based on a set of features. This bundle of features can be called a concept.

Similarly, we can provide a machine with a training sample of a given signal or dataset from which it can learn the correct concepts needed to identify wether new data or objects belong to a specific category. These generalized concepts is commonly referred to as a hypothesis.

An example: Let's say we want to identify reptiles in a dataset containing all types of animals. We extract a random subset for training the model, in which we have a set of features; *scales, coldBlooded, legs, eggLaying*. To start with we have a random sample from the training set as the starting hypothesis. This hypothesis will constantly evolve as we challenge the current hypothesis against the training data. This will go on until the hypothesis remains unchanged, and we have the best possible concept needed to differentiate reptiles from other animals.

1.2 What is function approximation and why do we need them?

Function approximation is the process of adjusting the given model, or function, to most likely represent the true target function. As for the evolution of hypothesis explained in the previous question, we need these function approximations to actively determine the vital parameters and their weight.

1.3 What is inductive bias in the context of machine learning, and why is it so important? Decision tree learning and the candidate elimination algorithm are two different learning algorithms. What can you say about the inductive bias for each of them?

> “An inductive bias of a learner is the set of additional assumptions sufficient to justify its inductive inerference as deductive inference” *Tom Michell*

Inductive bias is a set of assumptions used to predict a given output if it encounters a new input. Without this bias, the algorithm wouldn’t have learned anything except how to handle distinct key-value pairs, for instance, if a car encounters a cat, but it is trained to avoid dogs, it might not with high enough certainty know what to do.

When using a decision tree learning algorithm, we use a bias called a search bias which is greedy and keeps the most relevant searches higher up in the tree to make it as short as possible. The candidate elimination algorithm, however, uses a representational bias because it cannot represent all hypothesis. So instead of greedily choosing which part of the whole hypothesis space to search, it assumes that the solution to the problem can be expressed as a conjunction of concepts.

1.4 What is *overfitting*, and how does it differ from *underfitting*? Briefly explain what a validation set is. How can cross-validation be used to mitigate overfitting?

Overfitting refers to a model that models the training data too well. Overfitting occurs when the model learns both the valuable data and noise in the training data, which will be applied to new datasets and negatively impact the model’s ability to generalize. Underfitting, on the other hand, refers to a model that neither has learned the training data nor infer from new data.

The validation set makes up about 20 percent of the bulk of data used (training set 60 %) when training the model. The validation set is used for choosing the best of the models found by the training data and optimizing it. During the validation phase, overfitting is checked and avoided.

Cross-validation uses the initial training data to generate *n* different mini train-test subsets and used to generate *n* different hypothesis, which allows us to tune the hyperparameters with only our original training set. This way of repeating the expoeriment multiple times gives a more accurate indication of how well the model generalizes to unseen data. Cross-validation does not prevent overfitting in itself, but it may help in identifying a case of overfitting.

1.5 See the problems in seperate pdf

Listing 1: Sample Python code – Fibonacci sequence calculated analytically.

```python
from math import *

```

```python
# define function
def analytic_fibonacci(n):
  sqrt_5 = sqrt(5);
  p = (1 + sqrt_5) / 2;
  q = 1/p;
  return int( (p**n + q**n) / sqrt_5 + 0.5 )

# define range
for i in range(1,31):
  print analytic_fibonacci(i)
```

Following Listing 1... Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non ident, sunt in culpa qui officia deserunt mollit anim id est laborum.

blem 2

Listing 2: Sample Bash code.

```bash
#! /bin/bash
python stage1.py
echo "Stage I done!"
python stage2.py
echo "Stage II done!"
python stage3.py
echo "Stage III done!"
```

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non ident, sunt in culpa qui officia deserunt mollit anim id est laborum.