Project in Practical Machine Learning

Johannes Verwijnen

Department of Computer Science
University of Helsinki

Spring 2016

Outline

Project? in Practical? Machine Learning

- Welcome to the second iteration of this new project/lab course
- I'm your lecturer, Johannes Verwijnen (a mouthful - I know). If you want to talk to me, you can
  - visit me in B333 (very unlikely I'm there)
  - visit me at Ekahau offices in Salmisaari (more likely I'm there, better reserve time beforehand)
  - email me at jverwijn@cs.helsinki.fi
  - find me on IRC as duvin
  - call/SMS me on 0505731020
  - book a time using doodle `https://doodle.com/duvin` (better book several alternative times)

Project in Practical? Machine Learning

- This course counts as advanced studies in the Algorithms, data analytics and machine learning subprogram
- The idea of this course is to introduce you to a more "realistic" setting of doing machine learning than what we're currently offering in other courses
- Realism here refers to problematics with
  - live data
  - choice & parametrization of ML method
  - running a system in the networked world
- Prerequisites: Intro to ML, Scientific Writing (or similar knowledge), programming knowledge in chosen environment

How?

- You will
  - find a result that you wish to predict periodically
  - find the data that you wish to use for prediction
  - choose a suitable ML technique
  - implement and run an online system that will create periodic predictions and follow their accuracy
  - write a report of all that with reflection

  in a group of ~~1/1/4~~ 1-2 students
- There will be two general lectures (today and next week) with common content for all students
- Later, each group will have 2 formal meetings with the lecturer about their project to ensure mutual understanding of the tasks
- Peer support is available on IRC channel #tkt-ppml2016

Why?

- It's fun!
- Credit points (2-6)
  - Each credit point should represent $\sim$27 hours of work
  - 4 hours of lectures
  - 4 hours of meetings with lecturer
  - Project work (needs to be documented)
- Grading (0-5)
  - Based on report & presentation
  - Weight on reflection and result presentation rather than prediction accuracy
  - Report is needed for a pass (1) grade

Lectures

- 2 lectures, one with visiting guest lecturer:
  - Wed 20.1. 16-18 CK110
    - Course lecture on administrative issues
  - Wed 27.1. 16-18 C222
    - Guest lecturer: Janne Sinkkonen, PhD, Senior Data Scientist at Reaktor
    - Course lecture on data sources, dirtiness and context, existing tools & libraries and expected outcomes
- guest lectures are "motivational" in nature, giving context and ideas around usage of ML in the industry
- we'll start with the guest lecture, having a break after it for networking
- attendance is voluntary, although course lecture content is expected to be known to all students (slides available on course page)

Group meetings

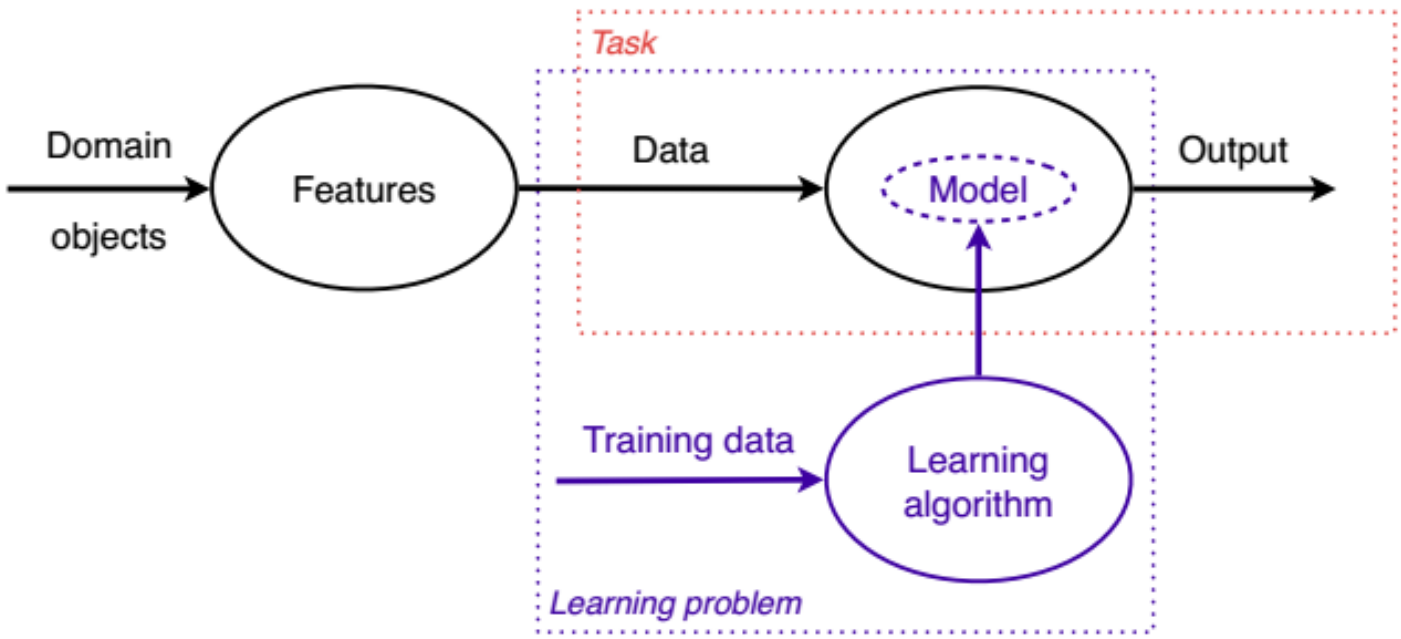
- 2 group meetings with the lecturer:
- First meeting starting next week
  - You should have
    - your target variable (what to predict)
    - data source
    - programming environment

    figured out. You should also have looked at
    - what ML & web frameworks to use
    - where to host your system
    - what ML algorithm could work
  - You will get
    - feedback on your choices
    - an idea of what is needed for the amount of credit points you are targeting
- Please book this meeting from my doodle ASAP (remember to give several alternative options, length: 2 hours) `https://doodle.com/duvin`

Group meetings (2)

- Second meeting in beginning of March
  - You should have
    - selected your ML algorithm and parametrized it
    - a working implementation of the whole system
    - an idea on how well you are doing
    - notes on how you selected your tools
    - be ready to "let go" of the system
  - You will get
    - to know what more is needed (if anything) that the system is acceptable
    - discussion around how to measure the "goodness" of your system
    - input on what to include in report and presentation, grading hints
- Please book this meeting from my doodle once you feel you are ready for it!

A Machine Learning System



[1]Graphic from Peter Flach. Machine Learning: The Art and Science of Algorithms That Make Sense of Data. Cambridge University Press, New York, NY, USA, 2012

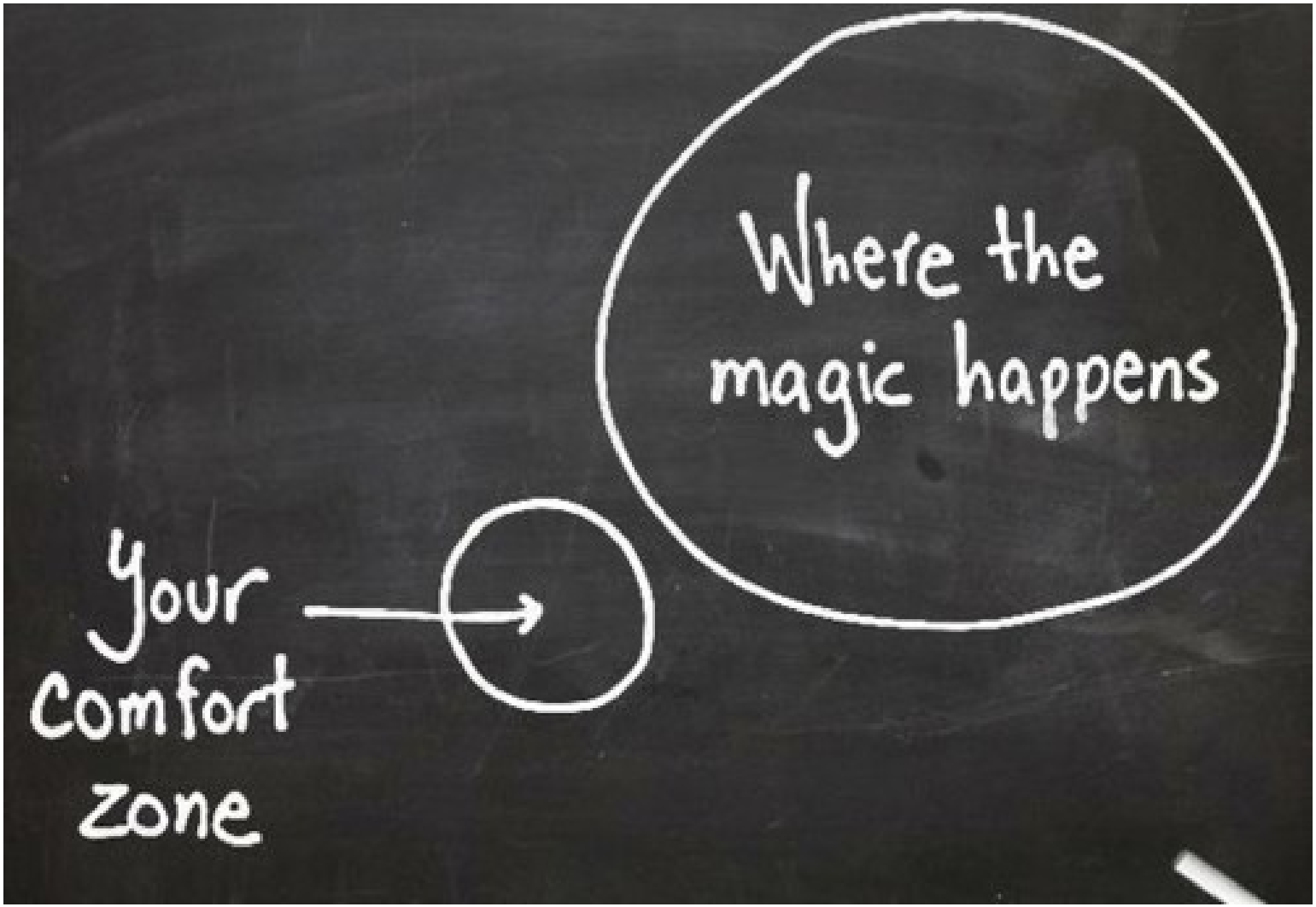
What the product should look like

- Concentrating on integration of a ML technique with periodic data in/output
- Handling live incoming data
- Storing and analyzing predictions
- Not concentrating on
  - Feature selection/extraction
  - Level of accuracy
  - Efficiency of implementation

Examples

- Predict stock markets (or indices or whatever)
  - Training data: old stock value data
  - Input: stock price, calculated features
  - Predict: index/stock up/down, individual stock scores
- Predict traffic data
  - Training data: old weather and traffic data
  - Input: daily weather measurements, calculated features
  - Predict: percentage of trains running, road traffic problems

Lecture



Data sources

- There are plenty of open data sources available
- Some require you to get an API key
- Mostly free for academic use (yes, this is academic use)
- In case you are struggling with finding a topic/data source, some examples are listed on the course web page

Dirtiness of data

- The data will be dirty compared to academic data sets used in many courses
- You will need to add a pre-processing phase to clean the data (and possibly do other stuff) for it to make sense to your ML technique
- The data most of you will use will be machine-generated, so its dirtiness can be well anticipated (compared to data filled in by humans)
- You need to learn about the context of your topic area and find how different exceptional situations are handled in your data source

Examples of context-sensitive situations

- Stock data
  - Trading suspended for stock (no data, zeroes, previous day's value)
  - Stock split (2:1 split, half the value)
  - Dividend (stock price drops by dividend on day it is applied)
- Weather data
  - No data from station (no data, zeroes, previous day's value)
  - Instrument failure (data not in believable range)
  - Communications failure

Other pre-processing

- Data aggregation
  - You might want to enrich the data by some aggregate measures.
  - As we're using time series data you might want to add rolling window statistics
  - These may be binary, categorical or numeric
- Normalization
  - This is related to data dirtiness as well
  - Variable categorization ("binning")
  - Remove irrelevant data

ML Tools overview

- In academics mostly MATLAB/R-based tools
- Problematic for running your model in real time

Janne Sinkkonen, PhD
Senior Data Scientist
Reaktor