3 Paper Lab: Representations

Designing Data Representations

In the following exercises, you will translate a high level description of a machine learning problem into a suited representational format. This type of assignment is very common for a Data Scientist in a real world job. Make sure you motivate your choices, highlight possible shortcomings, etc. In short, demonstrate your insights.

Example 1: Fouling the Joker

In Tim Burtons 1989 Batman movie, the Joker (played by Jack Nicholson) terrorises Gotham by poisoning the city's hygiene products. People who use the wrong (combination of) shampoo, deodorant, toothpaste and/or other hygiene products suffer a violent facial transformation quickly followed by death.

While it is possible to observe what products past victims have used by checking their bathrooms, and one can find out which ingredients these products have and where they were produced, it is impossible to see which ingredients caused the transformation directly, even through biopsy of the victims.

Batman wants to use machine learning to find out which combination of ingredients is causing the deaths and since there is no Lucius Fox in the 1989 movie, he contacts you as a knowledge engineering master (in the movie sense of the word) to recommend him a representational format. Which representation are you going to recommend?

1. Propose a representation of the learning examples suited to this problem.
2. Place your representation in the hierarchy we discussed in class and discuss what information is lost by using a representation lower in the hierarchy if possible, and why a more powerful format is not needed.
3. Clearly describe (or define) the hypothesis space. Give an example hypothesis that illustrates the representational format you chose.

Example 2: Figuring out the Paris

Paris Hiltons real-estate agent was asked to find a Feng-Shui house for Paris in 25 cities around the world. After showing her wealthy client a long list of houses, and being yelled at because the stairs were at the wrong end of the house or the rooms weren't placed in the right configuration given their functionality, the real estate agent finally figures out that Paris' idea of Feng-Shui is not the same as the definition found in books.

She contacts you as a knowledge engineer to help her using this promising technology that she heard about called "machine learning". She wants you to build a system that can (learn to) recognise if a house fits miss Hiltons preferences from its layout, i.e. floor plan and orientation.

1. Propose a representation of the learning examples suited to this problem.
2. Place your representation in the hierarchy we discussed in class and discuss what information is lost by using a representation lower in the hierarchy if possible, and why a more powerful format is not needed.
3. Clearly describe (or define) the hypothesis space. Give an example hypothesis that illustrates the representational format you chose.

Your turn!

Pick 2 of the following learning problems. Classify them as boolean, attribute-value, multi-instance or relational learning tasks. Discuss why the learning problem belongs to the chosen category, and discuss which information would be lost when using a lower level representational format (if available) and why a higher level representation is not necessary (if available). Devise a "language" to represent learning examples and list a (small) number of learning examples to illustrate this language. Devise a hypothesis language and write an (again small) number of hypotheses to illustrate the language.

1. The owner of a movie-theatre with a limited capacity (i.e. number of viewing theatres) wants you to help him predict how much a movie will make during the opening weekend to help him select the most profitable movies. He proposes to use IMDB-like data (who is the director, who starred in the movie as actors, etc ...) to train your machine learning algorithm.

2. You are building a Poker AI (multi-player Texas Hold'em) and want to equip it with opponent modelling, i.e. a module that will try to predict the actions and/or card-strength of each opponent.

3. You want a learning mail program to classify emails as spam or ham.

4. A busy hospital contacts you to build a system that can learn to predict how long patients will have to stay in intensive care after surgery. On the intensive care ward, the patients are constantly monitored, which means that several indicators such as blood pressure, heart rate, temperature, etc. are recorded often.

5. You want to build a system that predicts who will win the "Tour the France" next year. Don't forget that cycling IS a team sport.

6. You plan to get filthy rich using machine learning techniques to predict which stocks values are more likely to rise or drop based on twitter messages.

7. You want to help your teacher by building a model that predicts the grade of a student for the course Advanced Concepts in Machine Learning, based on the students past academic results.

8. You want to help your fashion-incompetent (and presumably colourblind) uncle to learn to dress correctly. He has be keeping notes on the combinations he has been wearing and the approving or disapproving looks he has been getting in his workplace.

Handing In

Work in pairs, but each of you should upload your pdf solution (unzipped!) through the student portal. Mention the name of your partner in the comment box.