Tutorial 2: Introduction to machine learning; probability theory

Machine learning concepts

1. Why might we want to give an AI agent the ability to learn from experience, rather than just hard-coding the agent function ourselves?
2. Explain how a **supervised learning algorithm** learns an unknown mathematical function. What's the input of the algorithm? What assumptions does it make? What's the goal of the learning algorithm?
3. **Quiz question:** What happens when a model/hypothesis **overfits** the training data? What is the potential problem of overtraining and what can we do to detect it? What can be done to reduce possibility of overtraining?

A quick introduction to SciPy

We're going to be using a Python toolbox called **SciPy** (packaged in an app called **Canopy**) for some of our machine learning experiments. It's a nice environment for implementing and testing machine learning algorithms, and visualising their performance.

4. Start Canopy, click on 'Editor', then open `intro.py` (from the Tutorial 2 website). This is a simple example of how you define a set of points, a linear function of those points and graph it in a figure.
   (a) Click 'Run' (the green arrow) to see what it does.
   (b) The lower pane of the editor is for interaction. Type some variable names into this pane, and Python will show you their values: this will give you an idea of the data structures that the code created.
   (c) Modify the code to create and display some different functions.
5. Open `learning.py`. This code defines a quadratic function, creates a data sample (sampled from the function with noise), and then attempts to fit two different hypotheses about the function, learning the parameters from the training set.
   (a) Run it, and work out what it does.
   (b) Experiment with different amounts of noise, different split of training vs. testing set, and other hypotheses functions.

Probability Theory

1. What are the meanings of the following terms:
   (a) **Sample space**
   (b) **Random variable**
   (c) **Event**
   (d) The **conditional probability** of an event E_1, given some other event E_2
   (e) **Probability distribution** of a random variable
2. Here's a toy database that records information about 12 people.

Owns_jandals	Loves_pork_pies	Goes_whitebaiting	Can_cook_pavlova	Gloomy	Nationality
0	0	1	0	1	english
1	0	0	1	0	english
1	1	0	0	1	english
1	1	0	0	0	english
0	1	0	0	1	english
0	0	0	1	0	english
1	0	0	1	0	kiwi
1	1	0	0	1	kiwi
1	1	1	1	0	kiwi
1	1	0	1	0	kiwi
1	1	0	1	0	kiwi
1	0	1	1	0	kiwi

   (a) **Quiz question:** Using relative frequencies in this database:
      i. Estimate the probability that a person is gloomy.
      ii. Estimate the conditional probability that a person is gloomy given that they are English.
      iii. Estimate the conditional probability that a person is Kiwi, given that they own jandals, go whitebaiting and can cook pavlova.
      iv. Estimate the variance of the probability distribution of Gloomy.

      (It's fine to express your answers as fractions.)
   (b) Let's say Dunedin's temperature for month of April is normally distributed with mean of 10.8°C and standard deviation of 5.4°C. Assuming temperature on any given day is independent of temperature on any other days, how many days in April would you expect to have temperature over 15°C? (Hint: If you're not sure how to compute probabilities for normally distributed random variables, take a look at this link for help)