Machine Learning
Winter 2022, Homework #1
Due: Tuesday, Jan 25

Staple all of your pages together (and order them according to the order of the problems below) and have your name on each page, just in case the pages get separated. Write legibly (or type) and Organize your answers in a way that is easy to read. Neatness counts!

For each problem, make sure you have acknowledged all persons with whom you worked. Even though you are encouraged to work together on problems, the work you turn in is expected to be your own. When in doubt, invoke the Gilligan's Island rule (see the syllabus) or ask the instructor. *Suspected cheating cases will be reported to the Director of Student Conduct and Advocacy.*

All homeworks are due at the beginning of the lecture on the due date. I will accept one homework up to one lecture late without penalty. You do not need to inform me – I will accept it automatically, no questions asked or documentation required.

---

1. **Hypothesis space and inductive bias.** (4 points) We want to learn an unknown function f that takes n input arguments $x_1, x_2, \ldots, x_n$ and produces one output y. The input variables can take one of 3 different values, i.e. each x_i can be either T (*true*), F (*false*), or U (*unknown*). The output variable y is boolean, i.e. can take on one of 2 different values. An example is a healthcare scenario where each of the x_i corresponds to a symptom (the patient has the symptom, the patient does not have the symptom, or we don't know whether the patient has the symptom), and y corresponds to the diagnosis, e.g. the patient has COVID-19 or not.

   (a) Let's consider the hypothesis space $\mathcal{H}$ consisting of all functions that take n such 3-valued input arguments and produce a boolean output. How many hypotheses are there in $\mathcal{H}$? Briefly explain your answer.

   (b) Is the inductive bias in $\mathcal{H}$ high or low? What are the implications of this for a machine learning algorithm that tries to learn the unknown function f from training data?

   (c) Say that you get a training dataset with p different training examples, each of the form $((x_1, x_2, \ldots, x_n), y)$. How many hypotheses in $\mathcal{H}$ are consistent with these training examples? Briefly explain your answer.

2. **Decision trees.** (4 points) Solve problem 3.4 from the textbook. In problem 3.4(b), if your answer to the second question is "yes", then give that member of the version space. In problem 3.4(c), rebuild the decision tree from scratch.

3. **Machine Learning in Python.** (2 points) The file Files/homeworks/hw1/iris.py on the Canvas course website contains Python code to train a shallow decision tree for the classification of flowers. The input features are the flowers' sepal length, sepal width, petal length, and petal width, and the label corresponds to the species, i.e. "iris setosa", "iris versicolor", or "iris virginica". The data is provided in the file iris.csv[1]. The code computes the *training accuracy*, i.e. the accuracy obtained when classifying all instances from the training dataset that was used to build the classifier in the first place.

   (a) Add a few lines of code to compute the accuracy in a 10-fold cross-validation set up. Use functions or methods from sklearn for k-fold cross-validation instead of implementing

[1] Fisher's iris flower dataset is well known to data scientists, see `https://en.wikipedia.org/wiki/Iris_flower_data_set`

your own. This will save you time and help you to get more familiar with sklearn. Include your code in your homework solution. You shouldn't make changes to the code that was provided, so there is no need to include any other code in your homework solution than the few lines that you added.

(b) What is the training accuracy? What is the accuracy obtained using 10-fold cross-validation? Briefly comment on which one is the lowest, and why that does (or does not) agree with your expectations.