TCSS 455 Machine Learning
Spring 2019, Homework #1
Due: Tuesday, Apr 23

Please submit a hard copy of your answers to the homework problems below. 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 University Conduct Officer.*

All homeworks are due at the beginning of 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 value of the input and output variables can be T (*true*), F (*false*) or U (*unknown*).

   (a) Let's consider the hypothesis space $\mathcal{H}_1$ consisting of all functions that take n such 3-valued input arguments and produce one 3-valued output. How many hypotheses are there in $\mathcal{H}_1$? Briefly explain your answer.

   (b) Let us restrict ourself to the hypothesis space $\mathcal{H}_2$ consisting of functions that are defined in the following way:

   - if any of $x_1, x_2, \ldots, x_n$ is U then $y =$ U
   - otherwise y is equivalent to a conjunction of any subset[1] of the variables $\{x_1, x_2, \ldots, x_n\}$

   How many hypotheses are there in $\mathcal{H}_2$? Briefly explain your answer.

   (c) Give a set of training examples such that there is no hypothesis in $\mathcal{H}_1$ consistent with these examples, or explain why you can not do that.

   (d) Give a set of training examples such that there is no hypothesis in $\mathcal{H}_2$ consistent with these examples, or explain why you can not do that.

   (e) In which of the two cases ($\mathcal{H}_1$ or $\mathcal{H}_2$) is the inductive bias the highest? What are the implications of the choice of hypothesis space $\mathcal{H}_1$ versus $\mathcal{H}_2$ for a machine learning algorithm that tries to learn the unknown function f from training data?

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,

[1] For instance, "y is equivalent to the conjunction of $\{x_2, x_4, x_7\}$" means that the value of y is the value of $x_2 \wedge x_4 \wedge x_7$; "$y$ is equivalent to the conjunction of $\{x_2\}$" means that the value of y is the value of x_2; "y is equivalent to the conjunction of $\{\}$" (the empty set) means that the value of y is T (*true*).

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[2]. The code computes the *training accuracy*, i.e. the accuracy obtained when labeling 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. 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.

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