TCSS 455 Machine Learning

Homework 1

Steve G. Mwangi

TCSS 455 Introduction to Machine Learning
Spring 2020, Homework #1
Due: Monday, Apr 20

Please submit a PDF with your answers to the homework problems below. You can make a scan or a picture of your written answers, or you can type your answers. 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 Associate Director of Student Conduct and Academic Integrity.*

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.

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

(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 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.

TCSS 455 ML

Homework #1

Spring 2020

Due Monday, 04/20/20

① Hypothesis Space and Inductive Bias

Unknown function, f

input, n values: $x_1, x_2, \dots x_n$

output: y (T or F)

each input T, F or U

(a) Hypothesis Space, H, that takes n such 3-valued input arguments and produces a Boolean output.
# of hypotheses

There are 3^n distinct instances for these n, 3-valued input arguments, therefore 2^{3^n} distinct hypothesis.

(b)(i) Is Inductive Bias in H high or low?

Low. (Highly enriched hypothesis space).

(ii) Implications? H will be unable to make inductive leaps beyond the observed training examples.

(c) p different training examples. Each of the form $((x_1, x_2 \dots, x_n), y)$. How many hypothesis in H are consistent with the training examples? 3^n hypothesis.

② Decision Trees 3.4 (textbook)

ID3 searches for just one consistent hypothesis, whereas the CANDIDATE ELIMINATION algorithm finds all consistent hypothesis.

(a) Show the decision tree that would be learned by ID3 assuming it is given the four training examples for the Enjoy Sport?

Attributes for Enjoy Sport

1. Sky	=> Possible Values	(i) Sunny (ii) Cloudy (iii) Rainy
2. Temp	=> "	(i) Warm (ii) Cold
3. Humid	=> "	(i) Normal (ii) High
4. Wind	=> "	(i) Strong (ii) Weak
5. Water	=> "	(i) Warm (ii) Cool
6. Forecast	=> "	(i) Same (ii) Change.

Exam	Sky	Temp	Humid	Wind	Water	Forecast	Enjoy Sport
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm.	Same	Yes
2. 3	Rainy	Cold	High	Strong	Warm.	Change	No X
(c.) 4	Sunny	Warm	High	Strong	Cool	Change.	Yes
(e) 5	Sunny	Warm	Normal	Weak	Warm	Same	No X

$$\text{Entropy}([+3,-1]) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$= 0.4422 + 0.5288$$

$$= \underline{0.9710}$$

Sky

Sunny => [+3, -1] Entropy(Sunny) = 0.8113

Cloudy => [+0, -0] Entropy(Cloudy) = 0

Rainy => [+0, -1] Entropy(Rainy) = 0

$$\text{Gain}(S, \text{Sky}) = 0.9710 - \frac{4}{5}\,0.8113$$

$$= \underline{0.3220}$$

Temp

Warm => [+3, -1] Entropy(Warm) = 0.8113

Cold => [+0, -1] Entropy(Cold) = 0

$$\text{Gain}(S, \text{Temp}) = \underline{0.3220}$$

Humid

Normal => [+1, -1] Entropy(Normal) = 1

High => [+2, -1] Entropy(High) = 0.9183

$$\text{Gain}(S, \text{Temp}) = 0.9710 - \frac{2}{5}(1) - \frac{3}{5}(0.9183)$$

$$= \underline{0.0200}$$

Wind

Strong => [+3, -1] Entropy(Strong) = 0.8113

Weak => [+0, -1] Entropy(weak) = 0

$$\text{Gain}(S, \text{Wind}) = 0.9710 - \left(\frac{4}{5}\right)0.8113$$

$$= \underline{0.3220}$$

Water

Warm => [+2, -2] Entropy(W) = 1

Cool => [+1, 0] Entrop(Cool) = 0

$$\text{Gain}(S, \text{Water}) = \underline{0.171}$$

Forecast

Same => [+2, -1] Entropy(Same) = 0.9183

Change => [+1, -1] Entropy(Change) = 1

$$\text{Gain}(S, \text{Forecast}) = 0.971 - \frac{3}{5}(0.9183) - \left(\frac{2}{5}\right)(1)$$

$$= \underline{0.02002}$$

Highest Gain = Sky

Temp

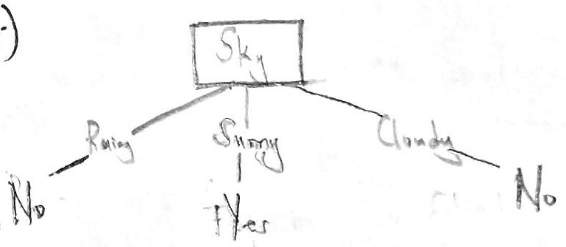
Forecast

Humid

Water

(2) Lowest Gain = Wind

(a.) (A.)



(2.)

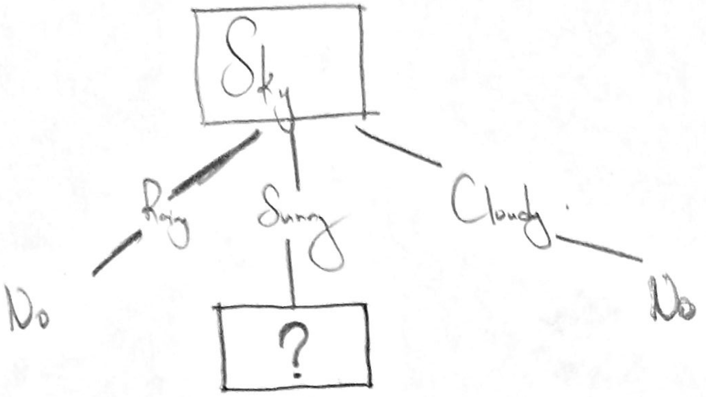
(b.) (B.) Is the learned DT equivalent to one of the members of the version space in (fig 2.3) Chapter 2?

Yes. The learned DT is equivalent to the element <Sunny, ?, ?, ?, ?, ?> in the general boundary of the version space. The relationship is that the learned DT is a member of the version space, or one of the hypotheses.

Continue · Hwk 1.

Q. 2

(c) Highest Info Gain of Sky, Temp, Wind, Water, Humid and Forecast.



$$\text{Entropy} = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = \underline{\underline{0.8113}}$$

[3, -1]
[0, -0]

$$\text{Gain}(S, \text{Temp}) = 0.8113 - (0.8113)(1)$$

$$= \underline{\underline{0}}$$

$$\text{Gain}(S, \text{Wind}) = \underline{\underline{0.8113}}$$

$$\text{Gain}(S, \text{Water}) = 0.8113 - \frac{3}{4}(0.9183)$$

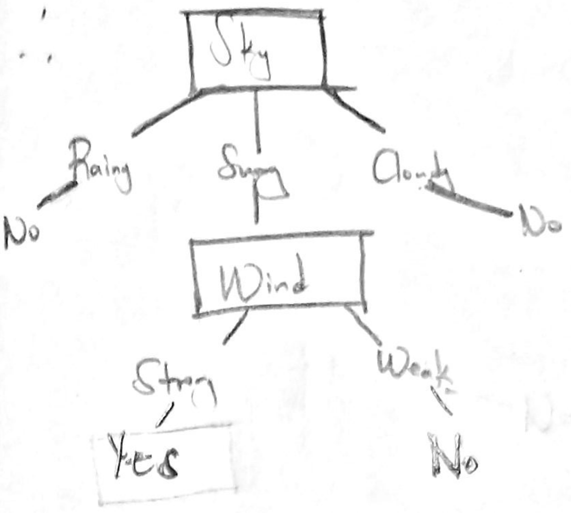
$$= \underline{\underline{0.1226}}$$

$$\text{Gain}(S, \text{Humid}) = 0.8113 - \frac{3}{4}(1)$$

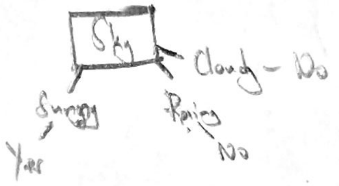
$$= \underline{\underline{0.3113}}$$

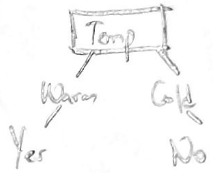
$$\text{Gain}(S, \text{Forecast}) = 0.8113 - \frac{3}{4}(0.9813)$$

$$= \underline{\underline{0.1226}}$$

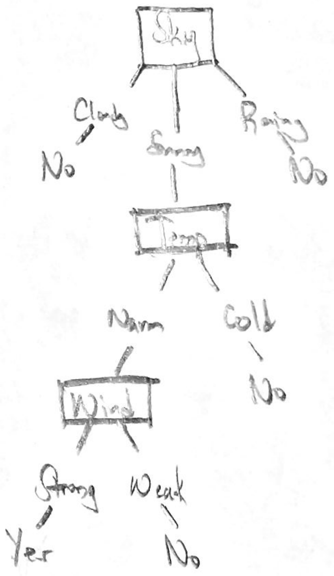


(d) (i) General : {<Sunny, ?, ?, ?, ?, ?>,
<?, Warm, ?, ?, ?, ?>}





Specific : {<Sunny, Warm, ?, Strong, ?, ?>}

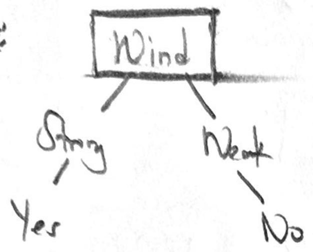


(ii) 2nd training example, makes the General set more specific by adding to the set

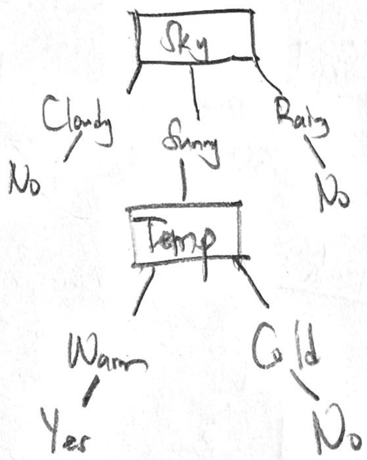
General : {<?, ?, ?, Strong, ?, ?>,
<Sunny, ?, ?, ?, ?, ?>,
<?, Warm, ?, ?, ?, ?>}

[Sunny, Warm, ?, ?, ?, ?]

:



:



Overfitting the model. ~~The~~ The DT
l be biased to ensure tree is
attributes for use by Candidate
gorithm will be less, in which case
ll perform better on the training sample
w data.

Question 3.

a. Lines added:

```
import numpy as np  # 1 of 4 Added this line
from sklearn.model_selection import cross_val_score # 2 of 4 Added this line

results = cross_val_score(clf, X, y, cv=10, scoring='accuracy') # 3 of  4 Added this line
print(np.mean(results)) # 4 of 4 Added this line
```

b. ***Training Accuracy* = 0.96**

10-Fold Cross validation values:

[1, 0.93333333, 1, 0.93333333, 0.93333333, 0.86666667, 0.86666667, 1, 1, 1]

***Mean of Values* = 0.9533333333333334**

Comment: The accuracy computed after cross-validation is less (0.9533 < 0.96) compared to the initial accuracy (0.96). The 10-fold cross-validation gives a better estimate for the model when it comes to new testing data, as opposed to just the training data accuracy, which can be overestimated.