# CS 165B – Machine Learning, Spring 2023

## Assignment #2
## Due Thursday, May 4 by 11:59pm

---

**Notes:**

- *This assignment is to be done individually. You may discuss the problems at a general level with others in the class (e.g., about the concepts underlying the question, or what lecture or reading material may be relevant), but the work you turn in must be solely your own.*
- Be sure to re-read the "Policy on Academic Integrity" on the course syllabus.
- Be aware of the late policy in the course syllabus – i.e., *late submissions will be accepted but 20% deduction per day.*
- Justify every answer you give – **show the work** that achieves the answer or **explain** your response.
- Any updates or corrections will be posted on Piazza, so check there occasionally.
- **All assignments must be clear and legible.** It is recommended to type the solutions on this Microsoft Word file directly and save it as PDF. If you'll be submitting a handwritten assignment, please ensure that it's readable and neat. If your writing is not easily readable, your solution is not easy to follow on the page, or your PDF is not of very high quality, your assignment will not be graded. DO NOT submit picture of your written work. (If you must scan your written work in, use a high-quality scanner. Plan in advance.)
- To turn in your assignment:
  - Use entry code **2K4D2G** to register as a student for **CS 165B Spring 2023 Machine Learning** on Gradescope.
  - Submit a PDF version to Gradescope

---

**Problem #1 [8 points]**
A ranking classifier ranks 20 training examples $\{x_i\}$, from highest to lowest rank, in the following order:

Highest Lowest

$x_2, x_3, x_6, x_1, x_7, x_{16}, x_9, x_{12}, x_4, x_{11}, x_{10}, x_8, x_{17}, x_{20}, x_{15}, x_{18}, x_{19}, x_5, x_{14}, x_{13}$

Examples $x_4$ and $x_{14}$, $x_5$ and $x_{12}$, $x_{11}$ and $x_{13}$ all have the same rank.
Examples $x_1$ through $x_{10}$ are in the positive class (which should be ranked higher); examples $x_{11}$ through $x_{20}$ are in the negative class (which should be ranked lower).

(a) How many ranking errors are there?

(b) What is the ranking error rate?
(c) What is the ranking accuracy?
(d) Draw the coverage curve for the ranking classifier on this dataset.

**Problem #2 [6 points]**
Consider a learning problem where real numbers are used as instances, and intervals over the real numbers are used as hypotheses. Each hypothesis in this scenario takes the form of $a < x < b$, where x represents the instance, and a and b are real constants. For instance, $4.5 < x < 6.1$ is a hypothesis that categorizes instances between 4.5 and 6.1 as positive, while others are considered negative. Provide an informal explanation for why there cannot be a hypothesis that is the least general generalization. Additionally, propose a slight alteration to the hypothesis representation that would allow for a least general generalization.

**Problem #3 [16 points]**
Build the best Classification and Regression Tree (CART) from the following training samples that use three features, each of which has two values. You should use gini gain (i.e., use gini index as impurity function) as criterion when choosing features:

Buy: (rating=high, market=bull, growth=false)
Buy: (rating=low, market=bull, growth=false)
Buy: (rating=high, market=bear, growth=true)
Buy: (rating=low, market=bull, growth=true)
Sell: (rating=low, market=bear, growth=true)
Sell: (rating=low, market=bear, growth=false)
Sell: (rating=high, market=bull, growth=true)

**Problem #4 [10 points]**
In a 5-class classification problem, 30 training examples are supplied that have the following class labels:

| | | | | | |
|---|---|---|---|---|---|
| 3 | 1 | 5 | 4 | 3 | 5 |
| 2 | 5 | 1 | 3 | 5 | 3 |
| 5 | 3 | 2 | 5 | 5 | 2 |
| 2 | 4 | 5 | 3 | 2 | 1 |
| 3 | 5 | 3 | 1 | 5 | 2 |

From the training data, we wish to estimate the probabilities of each class. Empirically estimate each class probability using (a) relative frequency, (b) Laplace correction, (c) m-estimate with m=10 and an even distribution of the pseudocounts, and (d) m-estimate with m=20 and an even distribution of the pseudocounts.

Plot these empirical probabilities, with class number on the x-axis and estimated probability on the y-axis – i.e., four plots (one for each approach), each of which consists of five connected points (the estimated class probabilities). Plot all four distributions on a single graph and label them.

Describe the trend as we go from (a) to (b) to (c) to (d) – that is, what does increasing the number of pseudocounts do (in general) to the probability distribution?

**Problem #5 [10 points]**

(a) Prove that for any dataset without conflict samples (samples with the same features but different labels), there exists a decision tree that is consistent with the dataset (the training error is zero). You can assume the tree uses entropy as impurity function and the features are categorical.
(b) Explain the disadvantage of using "minimal training loss" as the criterion for choosing BestSplit (i.e., choose the feature that decrease the training loss most for each node).
(c) What are some ways to avoid the disadvantage in (b)?

**Problem #6 [10 points]**

For the following set of points: {(-2 , -1) , (1 , 1) , (3 , 2)}
a) Find the least square regression line for the given data points. Show all work.

b) Plot the given points and the regression line in the same rectangular system of axes.

**Problem #7 [10 points]**

We have collected the following dataset for predicting heart attacks:

HEART ATTACK: (CHEST PAIN=yes,MALE=yes,SMOKES=no,EXERCISES=yes)
HEART ATTACK: (CHEST PAIN=yes,MALE=yes,SMOKES=yes,EXERCISES=no)
HEART ATTACK: (CHEST PAIN=no,MALE=no,SMOKES=yes,EXERCISES=no)
HEALTHY: (CHEST PAIN=no,MALE=yes,SMOKES=no,EXERCISES=yes)
HEART ATTACK: (CHEST PAIN=yes,MALE=no,SMOKES=yes,EXERCISES=yes)
HEALTHY: (CHEST PAIN=no,MALE=yes,SMOKES=yes,EXERCISES=yes)

a) Build a decision tree from this data by minimizing entropy.
b) Explain what information gain is, and why it is equivalent to minimizing entropy. Finally, explain what impurity measures are and how this connects to minimizing entropy.