CS 165B: Homework #2

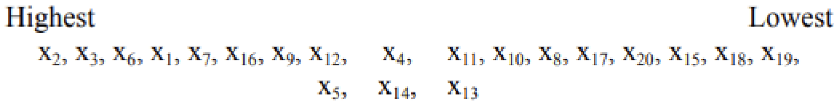
Due on May 4, 2023

Xifeng Yan 07575

Qiyuan Zhuang

Problem 1

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} 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?**

$5 + 3.5 + 2.5 + 2 + 2 = 15$. There are **15** ranking errors.

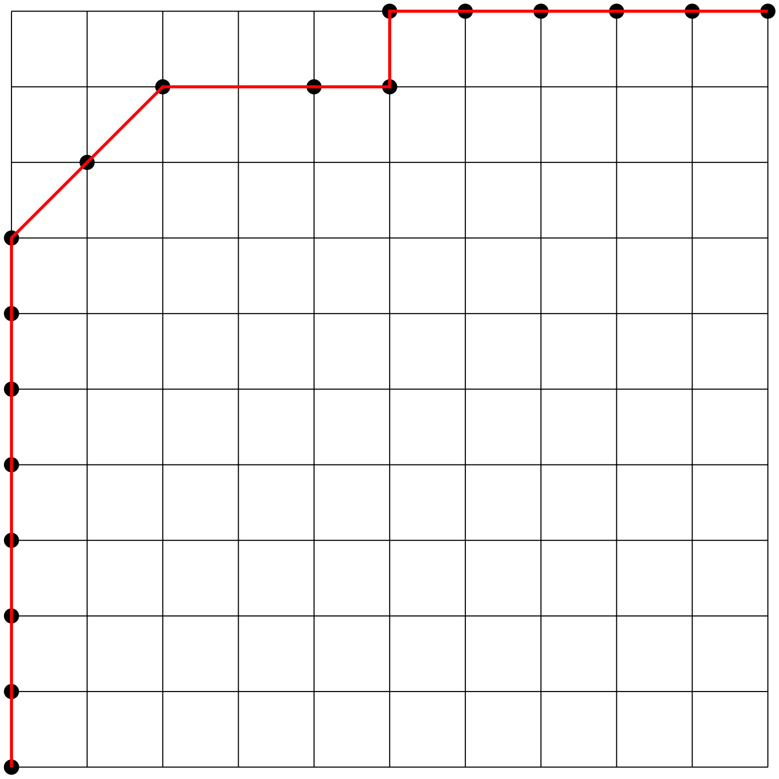
(b) **What is the ranking error rate?**

$$\text{ranking-err} = \frac{err}{PN} = \frac{15}{(10)(10)} = 0.15$$

(c) **What is the ranking accuracy?**

$$\text{ranking-acc} = 1 - \text{ranking-err} = 0.85$$

(d) **Draw the coverage curve for the ranking classifier on this dataset.**



Label of horizontal axis is FP. Label of vertical axis is TP.

Problem 2

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.

Answer: Assume that there is a hypothesis with the form of (a, b) be the LGG. Since it's an open interval, we can always find $\epsilon > 0$ such that $(a, b + \epsilon)$ is a general generalization. That contradicts to the assumption that (a, b) is a LGG.

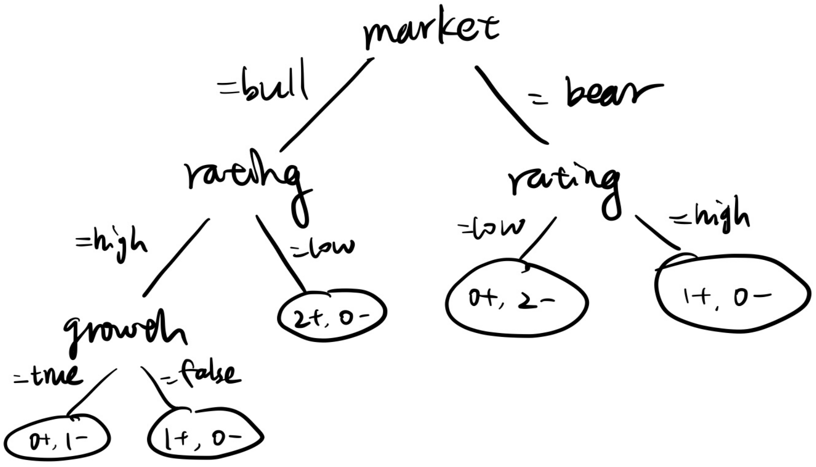
To fix the problem, we can alter the form of hypothesis to a $\leq$ x $\leq$ b, which means each hypothesis will become a closed interval. And thus LGG will be uniquely determined.

Problem 3

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:

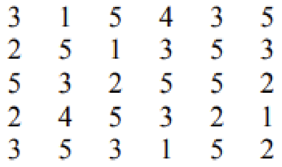
Buy: (rating=high, market=bull, growth=false)
Buy: (rating=low, market=bull, growth=false)
Buy: (rating=high, market=bear, growth=true)
Buy: (rating=low, market=bull, growth=true)
Sell: (rating=low, market=bear, growth=true)
Sell: (rating=low, market=bear, growth=false)
Sell: (rating=high, market=bull,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

Set **Buy** as positive.
rating: Imp $= \frac{3(4/9)+4(0.5)}{7} = \frac{10}{21} = 0.476$
market: Imp $= \frac{4(0.375)+3(4/9)}{7} = \frac{17}{42} = 0.405$
growth: Imp $= \frac{3(4/9)+4(0.5)}{7} = \frac{10}{21} = 0.476$



Problem 4

In a 5-class classification problem, 30 training examples are supplied that have the following class labels:



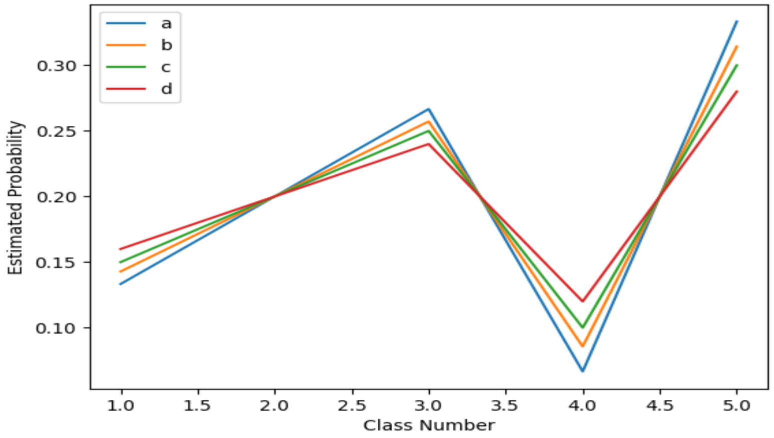
3	1	5	4	3	5
2	5	1	3	5	3
5	3	2	5	5	2
2	4	5	3	2	1
3	5	3	1	5	2

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.

Class	a	b	c	d
1	4/30	5/35	6/40	8/50
2	6/30	7/35	8/40	10/50
3	8/30	9/35	10/40	12/50
4	2/30	3/35	4/40	6/50
5	10/30	11/35	12/40	14/50

$$\text{m-estimate}: \dot{p}_i(S) = \frac{n_i+0.2m}{30+m}$$

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?

From (a) to (b) to (c) to (d), plot becomes more and more smooth(flat). That means estimate become more uniform as number of pseudocounts increasing. Thus we'll obtain a more balanced estimates and the model will be less likely to overfit.

Problem 5

(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.**

Proof. We prove by induction.
Base Case: Consider dataset containing only one example. We can construct a decision tree containing only one leaf simply.
Inductive step: Assume that for a dataset containing n examples, we can construct a decision tree that is consistent with the dataset.
Then we can add an example to the dataset. We can always find an feature value of the example that differs from other examples' since there's no conflict. So a new decision tree containing $n+1$ examples can be built with the former tree and added example.
Conclusion: Then we can conclude that a decision tree can also be built consistently for a dataset containing n+1 examples. The training error is definitely zero because label of each example is predicted correctly. Hence, we've shown that or any dataset without conflict samples (samples with the same features but different labels), there exists a decision tree that is consistent with the dataset. □

(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).**

It may lead to overfitting. The reason is that model will be extremely strict to each feature value and corresponding data including noises and mistakes when "minimal training loss" is used. That means model will not perform good on test data because the criterion reduces the training loss too much and thus errors will be taken.

(c) **What are some ways to avoid the disadvantage in (b)?**

(1) We can set a maximum depth of tree. Then the number of splits will be limited.

(2) We can use Impurity Function like Gini index, which can balance the Bestsplit. The homogeneity of labels within split will be maximized and overfitting will be efficiently avoided.

(3) We can select some specific feature which is important in the model and make the model less complex.

Problem 6

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.**

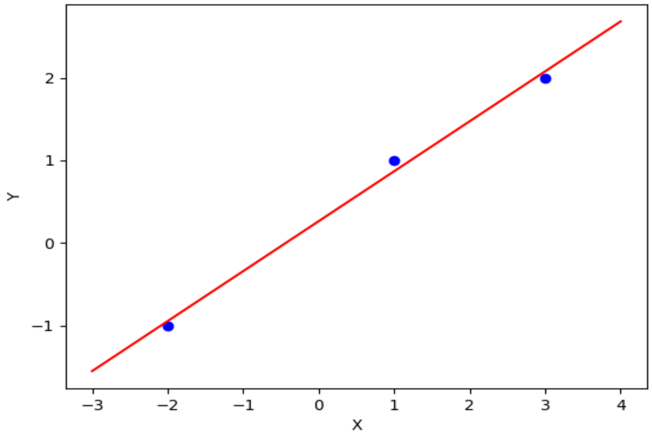
Set $y = a + bx$ and we have $y_i = a + bx_i + \epsilon_i$. We need to find (a, b) that minimizes $S = \sum_i^n [y_i - (a + bx_i)]^2$ where $n = 3$

$$\begin{cases} \frac{\partial S}{\partial a} = -2\sum_i [y_i - (a + bx_i)] = 0 \\ \frac{\partial S}{\partial b} = -2\sum_i [y_i - (a + bx_i)]x_i = 0 \end{cases} \Rightarrow \begin{cases} 3a + 2b = 2 \\ 2a + 14b = 9 \end{cases} \Rightarrow \begin{cases} a = \frac{5}{19} = 0.263 \\ b = \frac{23}{38} = 0.605 \end{cases}$$

Hence, the least squares regression line for the given points is:

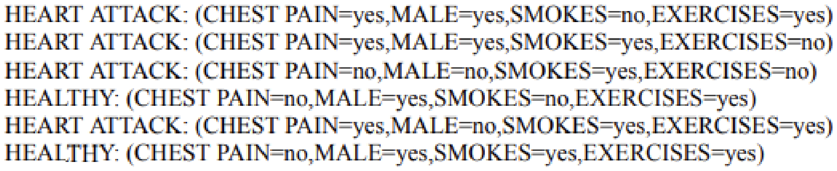
$$y = 0.263 + 0.605x$$

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



Problem 7

We have collected the following dataset for predicting heart attacks:

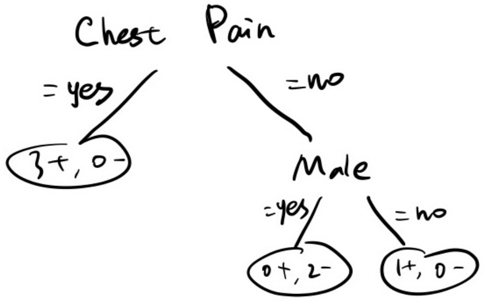


HEART ATTACK: (CHEST PAIN=yes,MALE=yes,SMOKES=no,EXERCISES=yes)
HEART ATTACK: (CHEST PAIN=yes,MALE=yes,SMOKES=yes,EXERCISES=no)
HEART ATTACK: (CHEST PAIN=no,MALE=no,SMOKES=yes,EXERCISES=no)
HEALTHY: (CHEST PAIN=no,MALE=yes,SMOKES=no,EXERCISES=yes)
HEART ATTACK: (CHEST PAIN=yes,MALE=no,SMOKES=yes,EXERCISES=yes)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

(a) **Build a decision tree from this data by minimizing entropy.**

Set **Heart Attacks** as positive: $\dot{p} = \frac{P}{P+N}$, $\text{Imp}(\dot{p}) = -\dot{p}\log\dot{p} - (1-\dot{p})\log(1-\dot{p})$

Chest Pain: $\frac{3(0.0)+3(0.918)}{6} = 0.459$
Male: $\frac{4(1.0)+2(0.0)}{6} = 0.667$
Smokes: $\frac{4(0.811)+2(1.0)}{6} = 0.874$
Exercises: $\frac{4(1.0)+2(0.0)}{6} = 0.667$



(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.**

Information gain is a splitting criterion, which measures the increase in information about the class gained by including the feature, **purity measured by entropy**.
It's equivalent to minimizing entropy because information gain is just defined in terms of entropy. While choosing a split that maximizes information gain, we're also choosing the split that results in child nodes with the lowest entropy.
Narrowly we can say: information gain = 1 − entropy.