CS245: Big Data Analytics (Fall 2023) November 20, 2023

Assignment 3: Classification

Due Date: **December 7, 2023**

Instructions

- Submit your answer on Gradescope as a PDF file. Both typed and scanned handwritten answers are acceptable.
- You may write scripts to solve each of the problems unless otherwise specified,
- Late submissions will not be accepted. Exceptionally, each student may request a *one-day* extension for one of the three assignments, provided they contact the instructor and TA *before the deadline*.
- Cite all resources used. Plagiarism will be reported.
- There is an OPTIONAL bonus question at the end of the homework!

Problems

Problem 1: Naïve Bayes and Bayesian Networks (30 points)

A team of researchers is investigating the influence of several factors on the decision to get a dog. They have gathered the following dataset, which captures an individual's education level, work travel habits, and physical activity. You can assume this dataset is an accurate representation of the broader population.

Education (Educated/Uneducated)	Work Travel (Regular/Non-Regular/Never)	Physical Activity (High/Low)	Dog (Yes/No)
Educated	Regular	High	No
Uneducated	Non-Regular	Low	Yes
Educated	Never	High	No
Educated	Regular	Low	Yes
Uneducated	Non-Regular	High	No
Uneducated	Never	Low	Yes
Educated	Non-Regular	Low	Yes
Uneducated	Regular	High	No
Educated	Never	Low	Yes
Uneducated	Regular	High	Yes

1. **(10 points)** Implement the Naïve Bayes method to compute the probability that an individual does not have a dog given that they are educated, travels for work regularly, and has low levels of physical activity.
2. **(10 points)** Assume that (1) work travel and physical activity are dependent on education and (2) having a dog is dependent on work travel and physical activity but not directly on education. Construct a Bayesian network for this dataset, specifying the Conditional Probability Table

(CPT) for each node. Based on your network, compute the probability that an individual does not have a dog given that they are uneducated, never travels for work, and has low levels of exercise.

3. **(5 points)** Compare the results obtained from Naïve Bayes and Bayesian networks. Discuss the assumptions made by each method and why it may or may not be reasonable for this dataset.

4. **(5 points)** Modify the Bayesian network by adding a new variable "Income" with levels "High" and "Low". Discuss how this change might influence the choice to buy a dog.

Problem 2: Decision Trees (30 points)

Consider the following dataset with three attributes: Forecast (Sunny, Overcast, Rainy), Time of Showing (Day, Night), and Traffic (on the 405) (True, False), and whether the person will buy tickets to "Killers of the Flower Moon" (Yes, No).

Forecast	Time of Showing	Traffic	Buys Tickets
Sunny	Day	False	No
Overcast	Day	True	Yes
Rainy	Night	False	Yes
Sunny	Night	True	Yes
Overcast	Night	False	Yes
Rainy	Day	False	No
Sunny	Day	True	No
Rainy	Night	True	No

1. **(10 points)** Construct a decision tree using the ID3 algorithm. Show your work in computing information gain for each attribute at each decision point.

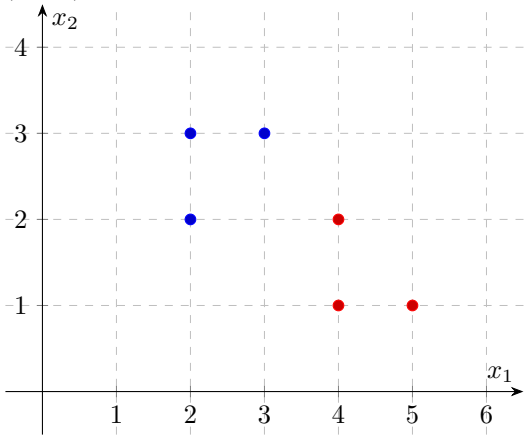
2. **(10 points)** Based on your decision tree, what conditions will most likely result in "Yes" for buying movie tickets?

3. **(10 points)** How does changing the decision rule (e.g., from information gain to Gini index) impact the structure of the decision tree?

Problem 3: Support Vector Machines (30 points)

Consider the following 2-dimensional dataset of two classes, each point is labeled by 1 (blue) or -1 (red). Construct a Support Vector Machine (SVM) model with a linear kernel.

x_1	x_2	Class
3	2	1
2	3	1
2	2	1
4	1	-1
4	2	-1
5	1	-1



1. **(20 points)** Draw the decision boundary and identify the support vectors.

2. **(5 points)** How would the classification boundary change if we used a non-linear kernel, such as the Radial Basis Function (RBF) kernel?

3. **(5 points)** Explain how the C parameter in SVM affects the decision boundary and generalization of the model.

Problem 4: Comparative Analysis and Extension (10 points)

1. **(10 points)** Compare the Bayesian Networks, SVM, and Decision Tree models in terms of their assumptions, strengths, and weaknesses. Provide at least two real-world application scenarios where one model might be preferred over the others.

2. **Optional Bonus (10 points)** Read the paper *Random Forests* by Leo Breiman (2001). Propose a variant of the Random Forest algorithm that you think could potentially outperform the original version. Justify your proposal. The number of points given will be based on the quality and thoughtfulness provided in your answer.