CPSC 330 - Applied Machine Learning

Tutorial 4

Course notes

So far, we have worked with various transformers and supervised machine learning models. The goal of this activity is to complete tables that provide an overview of

1. The strengths, weaknesses, and key hyperparameters of different machine learning models
2. The purpose, use cases, and key considerations of various transformers

(This will serve as a handy reference for your upcoming exam and beyond!)

Your task is to engage in group discussions and fill in the designated row in this Google document

- For strengths and weaknesses, some things to consider are:
  * concerns about underfitting
  * concerns about overfitting
  * speed
  * scalability for large data sets
  * interpretability
  * effectiveness on sparse data
  * ease of use for multi-class classification
  * ability to represent uncertainty
  * time/space complexity
  * etc.

Estimators

Model	Strengths	Weaknesses	Key hyperparameters (and their impact)
Decision trees			
KNN			
SVM RBF			
Linear models (logistic regression or linear regression)			

Transformers

Transformation	Purpose	Use cases	Key consideration
Imputation			
Scaling			
One-hot encoding			
Ordinal encoding			
Bag-of-words encoding			

Exercise 2 - Linear models

A professor is trying to create a linear regression model to predict exam scores for their students. The exam is scored out of 100 points. The predictors are hours studied (total), hours of sleep the night before the exam, and class attendance (encoded as low-medium-high).

The model is fitted on unscaled features, resulting in the following coefficients:

Feature	**Coefficient**
Hours studied	+4
Hours slept	+2
Attendance_low	-8
Attendance_medium	+2
Attendance_high	+8
Intercept	20

As the first step, write the linear model as a function below:

Predicted_score = …

Then, answer the following questions:

1. Write a couple of scenarios where the model would predict a score of 100 for the student.
2. What is the meaning of the intercept?
3. If a student was able to increase their attendance from low to medium, how much would their score go up (all other factors being equal)?
4. Can you identify which feature has the greatest impact on the predicted score? Why or why not?
5. A student has high attendance, studies 20 hours, and sleeps the recommended 8 hours before the exam: what is their predicted score? Does this reveal any problem with our model?