ECE 157A/272A - Homework 3

ECE 157/272A TAs

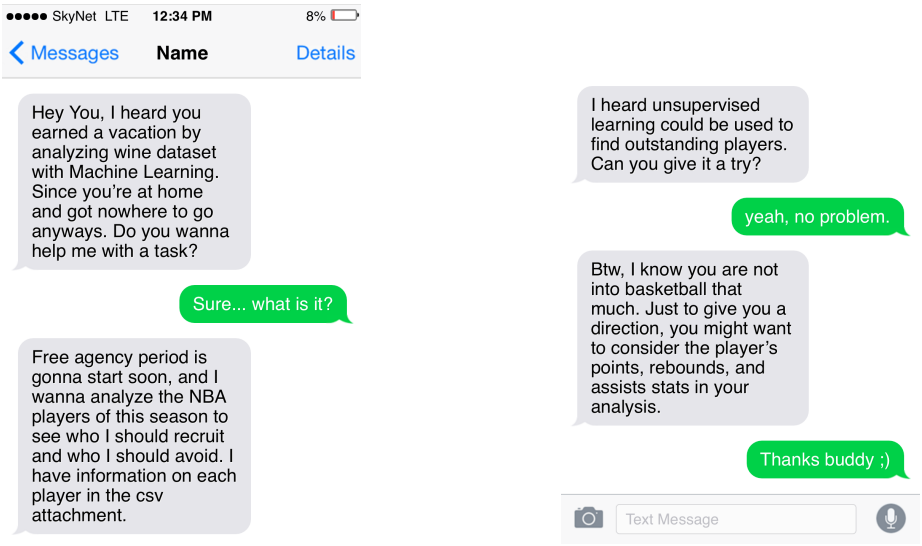
Fall 2021

Due: Monday, 1st of November, 11:59pm
Submit at `https://www.gradescope.com/courses/322684`
GradeScope add code: 74JWNY

1 Introduction

This is the third homework assignment to get you familiar with the machine learning tools. We're introducing another type of machine learning algorithm, which is unsupervised learning. Specifically, we are performing outlier detection with unsupervised learning.

Recall that you went to the employee party last Friday, and you won the wine tasting competition. As a reward, you earned a week of paid vacation. But, due to COVID-19, you have nowhere to go so you are just laying on the couch, watching Netflix. All of sudden you receive a message from your NBA team manager friend and the conversation goes like this:



2 Dataset

Ensure you've download the latest version of the data set from Piazza

We are using data from the *Basketball Reference* website, a data set related to NBA player in the 2020-2021 season. For more details, consult: `https://www.basketball-reference.com/leagues/NBA_2021_per_game.html`.

The file is provided in a format known as Comma Separated Values (CSV). You can open it as raw text to take a look! The data have 30 columns of features describing the players. These fields come in strings, integers, and floating-point numbers. This data set does have missing values.

This dataset doesn't contain any labels because it is meant for unsupervised learning.

You can use any method to load in the CSV file. Refer to the *Getting Started* guide if you are not sure about this step.

3 Recommended Flow

1. Make a function to load the data.
2. Clean the dataset so that all string entries and missing values are converted to floating-point numbers.
3. Make a function to plot the data, to see the distribution of each feature. (Hint: Histogram and Cumulative distribution function (CDF).) Identify three important features. (Hint: If you don't know which features are important, you can try the 3 features mentioned in the introduction.) Approximate the percentage of outlier samples.
4. Make a function to alter/preprocess the data (delete the features that are not important, normalize the data, and so-on...)
5. Train an Outlier Detection model.
6. Score each player based on the model's outlier score. Identify the top three outliers.
7. Verify the top three outliers by plotting the player's feature values onto the graphs from step 3. Check the location of the top 3 outliers with a 3-D scatter plot (the axes of the plot should be the three selected features from step 3.)
8. Save the outlier scores and the plots used to verify the top three outliers.
9. Repeat the previous steps until satisfied.

4 Problem Formulation

Input	Data Type	Description
Rk	Integer	Rank Based on Last Name
Player	String	Name of the NBA Player
Pos	String	Position of the NBA Player
Age	Integer	Age of the NBA Player
Tm	String	Team of the NBA Player
G	Integer	Number of Games Played
GS	Integer	Number of Games Started with the Player on the Court
MP	Floating-Point	Minutes Played Per Game
FG	Floating-Point	Field Goals Per Game
FGA	Floating-Point	Field Goals Attempts Per Game
FG%	Floating-Point	Field Goal Percentage
3P	Floating-Point	3-Point Field Goals Per Game
3PA	Floating-Point	3-Point Field Goals Attempts Per Game
3P%	Floating-Point	3-Point Field Goal Percentage
2P	Floating-Point	2-Point Field Goals Per Game
2PA	Floating-Point	2-Point Field Goal Attempts Per Game
2P%	Floating-Point	2-Point Field Goal Percentage
eFG%	Floating-Point	Effective Field Goal Percentage
FT	Floating-Point	Free Throws Per Game
FTA	Floating-Point	Free Throws Attempts Per Game
FT%	Floating-Point	Free Throw Percentage
ORB	Floating-Point	Offensive Rebounds Per Game
DRB	Floating-Point	Defensive Rebounds Per Game
TRB	Floating-Point	Total Rebounds Per Game
AST	Floating-Point	Assists Per Game
STL	Floating-Point	Steals Per Game
BLK	Floating-Point	Blocks Per Game
TOV	Floating-Point	Turnovers Per Game
PF	Floating-Point	Personal Fouls Per Game
PTS	Floating-Point	Points Per Game

Your task is to build an outlier model to identify the outlier players in the given CSV. An outlier player can be an extremely good player or an extremely bad player depending on which side of the spectrum the player's stats lie on the distribution graphs.

You are required to use the Python scikit-learn library to construct your models. You are required to use the following three methods:

- **One Class Support Vector Machine (SVM)**
- **Elliptic Envelope**
- **Isolation Forest**

However, you can experiment with any other algorithms you find interesting! Link to the documentation for more methods is available on (**Outlier Detection in SciKit Learn**)

For every algorithm, train it on the data and give an outlier score to each player. Identify the outlier players, based on the outlier model's scores. Then, pick the top three outliers and verify that they are indeed an outlier by using the CDF plot and the 3D scatter plot. Make a new CSV file, *MODEL_NAME_Scores.csv*, with two columns, the player's name and the player's outlier score. The CSV should be **sorted from min to max** based on the player's outlier score. Submit the CSV files with your report.

You also must write a brief report answering the following questions:

- **Describe in your own words what are Outlier Detection and Novelty Detection.** And, how are they different? (Hint: training samples) Does our current problem belong to Outlier Detection or Novelty Detection and why?
- **Explain the data preprocessing/transformation methods you applied.**
- **Show the distribution of the feature values.** Use histogram (Number of Occurrences vs Feature Values) and CDF (Percentage of Player vs Feature Values). Pick three features to train your models. Approximately, how many percent of the players are outliers, and how did you come to that conclusion based on the plots?
- **Describe in your own words how each of the three algorithms creates a model.** How do the model decide the boundary between inliers and outliers? How are the outlier scores calculated? Explain your choice of hyperparameters if any.
- **For each algorithm, pick the top three outliers, verify that they are indeed an outlier, and check whether the player is an outlier due to being bad or being good** Explain your reasoning with the model's outlier scores and the CDF plots.
- **For each algorithm, show a 3-D scatter plot marking the inliers, outliers, and the top three outliers data points.** Describe how the inlier region different from the outlier region.
- **Could this model be used to predict outlier players in the next season?** Justify your answer and state your assumptions.

5 Grading Criteria

Unlike the previous homework, this homework does not have an unknown.csv. Instead, you are scoring the players in the training data and verifying the scores based on your observations. This homework is trying to get you to understand how to apply outlier detection in analyzing a given dataset and extracting useful information.

You will be graded on:

- Your report including **all parts highlighted in bold**, the guiding questions, and the plots in *Problem Formulation*. (See appendix for more details)
- Three CSV files, one for each outlier detection algorithm.

6 What to turn in

Upload the following files to GradeScope at `https://www.gradescope.com/courses/322684`

- `One_Class_SVM_Scores.csv`: A CSV file with two columns containing the player's name and your model's outlier score for the corresponding player.
- `Elliptical_Envelope_Scores.csv`: A CSV file with two columns containing the player's name and your model's outlier score for the corresponding player.
- `Isolation_Forest_Scores.csv`: A CSV file with two columns containing the player's name and your model's outlier score for the corresponding player.
- `report`: A **PDF** document answering the questions, and including all diagrams listed in the *Problem Formulation*.

If you decide to use other outlier algorithms, your CSV file names can be different. The csv files should include a header row with values: Name, Scores. The rows should be sorted by the outlier scores in ascending order.

References

[1] Basketball Reference. 2020-21 nba player stats: Per game. `https://www.basketball-reference.com/leagues/NBA_2021_per_game.html`, 2021.

[2] Scikit-Learn. Novelty and outlier detection. `https://scikit-learn.org/stable/modules/outlier_detection.html`.

Appendix: Grading Details

1. Outlier Detection vs Novelty Detection (4 pts)
   (a) Describe in your own words what is Outlier Detection
   (b) Describe in your own words what is Novelty Detection
   (c) How are they different? Hint: training data
   (d) Does this homework problem belong to Outlier Detection or Novelty Detection and why?
2. Explain the data preprocessing/transformation methods you applied (2 pts)
3. Show the distribution of 3 most important feature values (5 pts)
   (a) Feature 1 - histogram and CDF plots
   (b) Feature 2 - histogram and CDF plots
   (c) Feature 3 - histogram and CDF plots
   (d) Explain why are the three features important based on the histogram and the CDF plots
   (e) Approximate the percent of outlier players, and explain why using the above graphs
4. One Class SVM (Algorithm 1) (6 pts)
   (a) Describe in your own words how the algorithm creates a model
   (b) How does it decide on the outlier score of the samples
   (c) Explain your choice of hyperparameters
   (d) Identify the top three outliers. Mark the feature values of the top 3 outlier samples on the CDF graphs
   (e) Plot a scatter plot that shows the inliers, outliers and top 3 players with respect to the 3 selected features
   (f) With the outlier score, the CDF graphs and scatter plot, explain whether they are outliers for being outstandingly bad or outstandingly good
5. Elliptic Envelope (Algorithm 2) (6 pts)
   (a) Describe in your own words how the algorithm creates a model
   (b) How does it decide on the outlier score of the samples
   (c) Explain your choice of hyperparameters
   (d) Identify the top three outliers. Mark the feature values of the top 3 outlier samples on the CDF graphs

(e) Plot a scatter plot that shows the inliers, outliers and top 3 players with respect to the 3 selected features

(f) With the outlier score, the CDF graphs and scatter plot, explain whether they are outliers for being outstandingly bad or outstandingly good

6. Isolation Forest (Algorithm 3) (6 pts)

   (a) Describe in your own words how the algorithm creates a model

   (b) How does it decide on the outlier score of the samples

   (c) Explain your choice of hyperparameters

   (d) Identify the top three outliers. Mark the feature values of the top 3 outlier samples on the CDF graphs

   (e) Plot a scatter plot that shows the inliers, outliers and top 3 players with respect to the 3 selected features

   (f) With the outlier score, the CDF graphs and scatter plot, explain whether they are outliers for being outstandingly bad or outstandingly good

7. Overall, how does the inlier region different from the outlier region? (1 pts)

8. Could these models be used to predict outlier players in the next season? (2 pts)

   - Justify your answer and state your assumptions

9. Three score CSVs (1 pt each)

Appendix: Sample Graphs

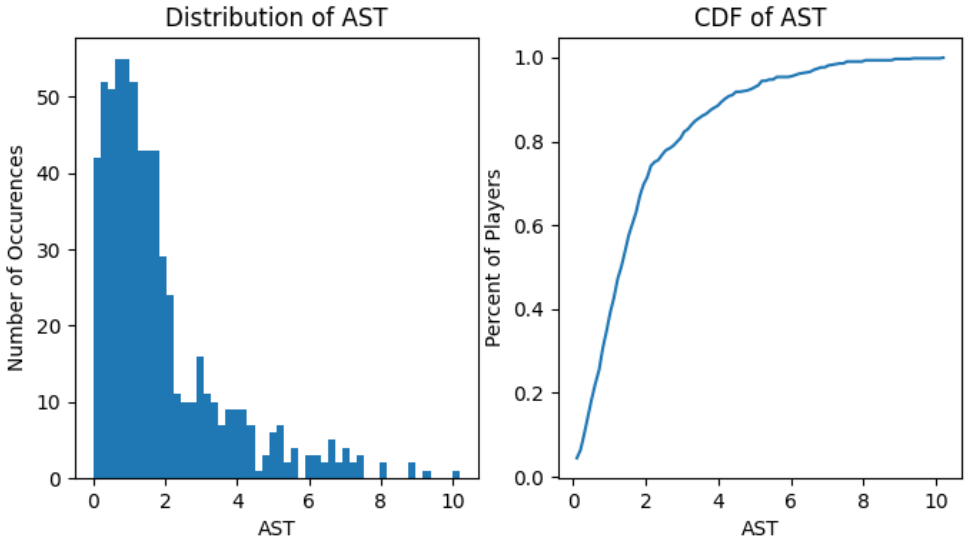


Figure 1: Example of a histogram and CDF plots for assist feature from 2018-19 NBA dataset

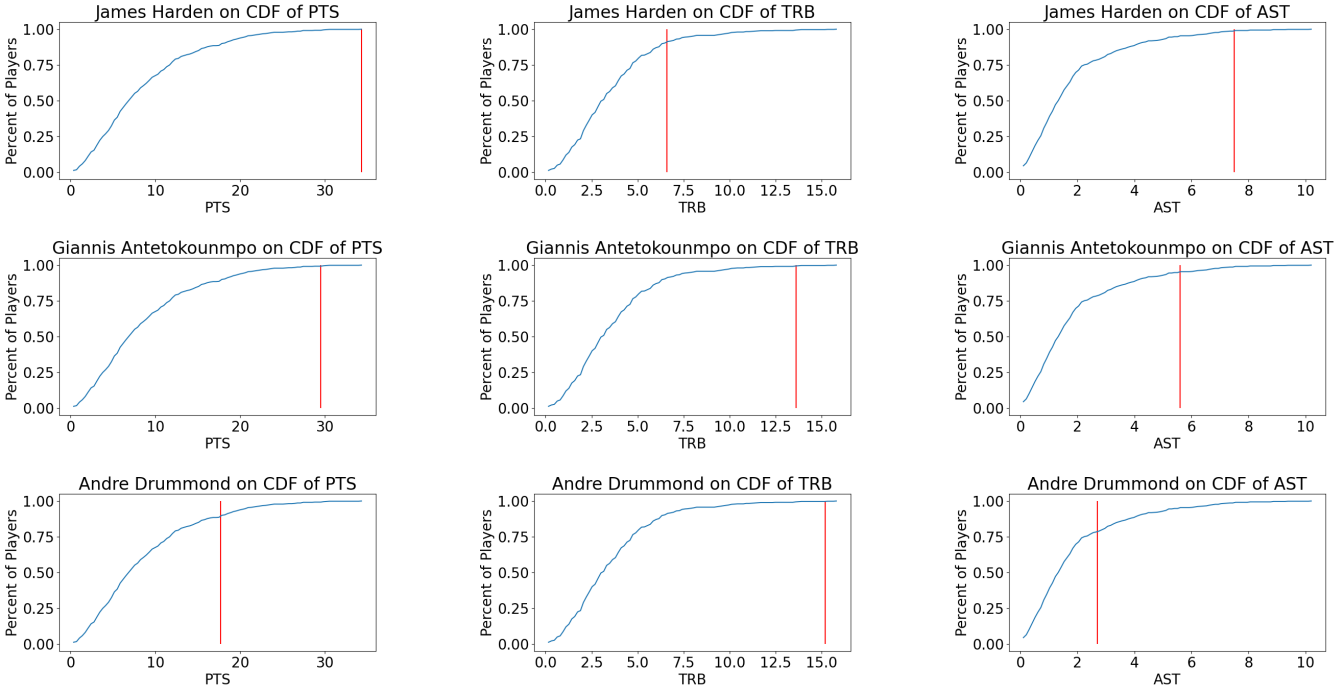


Figure 2: Example of outlier players' feature values marked on CDF plots for the 2018-19 NBA dataset

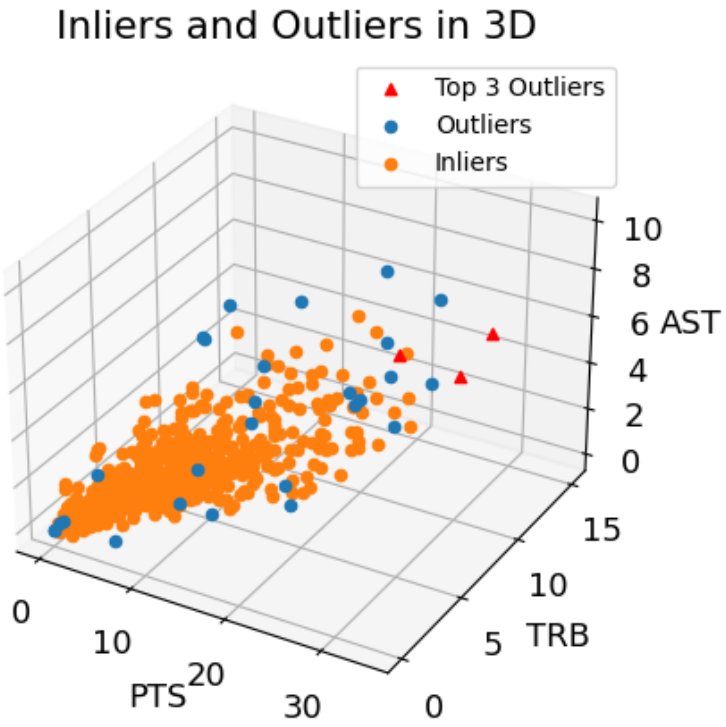


Figure 3: Example of scatter plot marking inliers (orange), outliers (blue), and top 3 players (red) for the 2018-19 NBA dataset.