ECE 157A - Homework 2

ECE 157 TAs

Fall 2021

Due: Sunday, 24th of October, 11:59pm
Submit at `https://www.gradescope.com/courses/322684`
GradeScope add code: 74JWNY

1 Introduction

This is the second homework assignment to get you familiar with the machine learning tools. We're going to deal with another important type of supervised learning: *Regression* models.

Recall that you did an excellent job analyzing the diabetes data last week. Your boss was happy and he announced an employee party over the weekend. You've received the following invitation:

> `From: bossman.bossy@icinc.med`
> `To: you@icinc.med`
> `Subject: Wine tasting competition`
>
> To celebrate our current progress and also introduce our diverse company culture, we're going to hold a wine tasting competition next week! You may not know that drinking is one of my favorite hobbies (surprise!). In fact, I've been a certified sommelier since I joined the company in 1990, and I've collected a lot of information about different wines over the years.
>
> For the competition, I would like you to recommend the most tasting wine out of a dozens of options to me. If you pick the one with high quality, you can have an extra week of paid vacation!
>
> I know most of you are more professional being a data analyst than a sommelier, so I would like to provide enough wine data for you to leverage your expertise!
>
> Different from the diabetes outcome, the wine quality is scored from 0 to 10, instead of being labeled as good or bad. **You can either use *Regression* techniques to predict a real number, or still use *classification* methods to predict multiple categories**. A lot of classifiers you've used also have variants for regression tasks. This should be an interesting off-work activity for you!

2 Dataset

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

We are using data from the *White Wine Quality Data Set*, a data set related to white variants of the Portuguese "Vinho Verde" wine. For more details, consult: `https://www.vinhoverde.pt/en/about-vinho-verde` or [1].

Due to privacy and logistic issues, the features are physicochemical stats and qualities are sensory variables from three wine experts (e.g. there is no data about grape types, wine brand, wine selling price, etc.). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).

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 columns have 11 variables on quantifying the chemical properties of each wine: `fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, `pH`, `sulphates`, and `alcohol`. All of these fields come in floating point numbers. This data set doesn't have any missing values.

`Quality` is the one you want to predict! It is a categorical variable ranged from 0 to 10, where 0 indicates poor quality and 10 indicates high quality.

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

3 Recommended Flow

The flow is similar to homework 1, but with different objectives. You can reuse the data loading and preprocessing functions from the last assignment.

4 Problem Formulation

Input	Range	Description
fixed acidity	[3.8,14.2]	Various types of acids found in grapes
volatile acidity	[0.08,1.1]	Acids that are distilled out
citric acid	[0.0,1.66]	One of the fixed acids which gives freshness
residual sugar	[0.6,65.8]	Natural sugar from grapes after fermentation
chlorides	[0.009,0.346]	Major contributor to saltiness
free sulfur dioxide	[2.0,289.0]	Amount of the SO_2 that is not bound
total sulfur dioxide	[9.0,440.0)	Sum total of the bound and the free SO_2
density	[0.99,1.04]	Mass of wine per unit volume
pH	[2.72,3.82]	A value to specify the acidity/basicity
sulphates	[0.22,1.08]	Mineral salts connected to fermentation
alcohol	[8.0,14.2]	% vol or alcohol by volume (ABV)

Output	Range	Description
quality	[3,9]	Wine quality between 0 (very bad) and 10 (very excellent)

Your task is to create a classifier that converts physicochemical data of the white wines into a prediction for how good is the quality of the wine.

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

- **Logistic Regression**
- **Linear Regression**
- **Multi Layer Perceptrons Classifier (MLPC)**
- **Multi Layer Perceptrons Regressor (MLPR)**
- **Support Vector Classification (SVC)**
- **Support Vector Regression (SVR)**
- **Gaussian Process Classifier (GPC)**
- **Gaussian Process Regressor (GPR)**

However, you can experiment any other algorithms you find interesting! Links to the documentation for more methods is available on **Supervised Learning in SciKit Learn**)

After you have trained your algorithms and selected the one you think is best, train it on the whole training set, then predict on the data in *unknowns.csv*. Make a new CSV file, *scores.csv*, with only one column, the predicted wine quality, and submit it with your report.

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

1. **Show the distribution of the quality variable** (Hint: use histogram). How many samples are in each category? Is this a balanced data set?
2. **Compute the correlation matrix between any two features**. What correlations do you observe?
3. **How regression methods are different from the classification methods in terms of their prediction results?**
4. The following questions are intended to test your understanding of the models listed above: (Hint: Answers can be found in lecture slides and/or SciKit Learn website.)

   (a) Mathematically Linear Regression solves a problem of the form:

   $$\min_{w} \|Xw - y\|_2^2$$

   Explain the above formula, what does X, w, and y represent respectively?

(b) Logistic regression, despite its name, is a linear model for classification. Compared to the Linear Regression, it applies a so called logistic function (or sigmoid function) to the hypothesis. Explain what is a logistic function?

(c) What is the difference between a Multi-layer Perceptron Classifier with logistic activation function and a Logistic Regressor?

(d) What is the activation function used in Multi-layer Perceptron Regressor? Can we use sigmoid as its activation function, if not, why?

(e) The output of SVC's decision function for a given sample x can be written as:

$$\sum_{i \in SV} y_i \alpha_i K(x_i, x) + b$$

Explain the above decision function, what does α and K represent respectively?

(f) Give two linearly separable classes, the objective of SVM is to find a hyperplane that separates these two classes with minimum error while also maximizes the the perpendicular distance between the two closes points from either of these two classes. Explain how SVM can be used for regression? (Hint: SVR introduces a slack variable ϵ)

(g) Why Gaussian Process has slow performance for larger datasets (Hint: look into the kernel function for GP)?

(h) How is Gaussian Process fundamentally different from the above models?

5. Evaluate your choices of model with cross-validation (CV). What is the CV accuracy on training data and validation data for each model? **Use a table or scatter plot to show the accuracy results.** Which one has the best performance? Do you observe overfitting/underfitting?

5 Grading Criteria

You must choose one model as your "best" for this data. You will run your model on the "unknown" data for which you will not be given the quality labels. The top five students by accuracy of their predictions on this unknown data will recieve a small amount of bonus points.

You will be graded on:

- Achieving the **accuracy requirement (over 60% on the unknown data).**
- Your report including **all parts highlighted in bold** in *Problem Formulation*

6 What to turn in

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

- `scores.csv`: A CSV file with one column containing your model's predictions for the wine quality in *unknowns.csv*.
- `report`: A PDF document answering the questions, and including all diagrams listed in the *Problem Formulation*.

References

[1] Paulo Cortez. Wine quality data set, 2009. A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal.