19.11.2018

CMP5130 – Machine Learning and Pattern Recognition

Project Proposal

The Dataset

Selected dataset for this ML project will be "Mushroom Classification" by UCI Machine Learning that can be seen here. The dataset consists of 22 features and 1 class, with 8123 instances. Features are almost exclusively physical, such as shape, color, and odor; and they are all categorical. Class of the mushrooms is binary, with "edible" and "poisonous". These two classes are distributed almost 50/50, so class imbalance will not be an issue.

The Aim of the Project & Possible Methodology

In this binary classification problem, the ultimate goal will be classifying mushrooms as correctly as possible. To be more specific, our performance measures will focus on not only accuracy, but also recall and precision of the model. If we had to value one of the criteria more than the others, it could be said the more important metric is recall. This is because missing a poisonous mushroom is much more critical (and possibly deadly) than not being able to eat some harmless mushrooms.

Methodology of the project will probably include a shorter data acquisition and preprocessing period than usual. As data will be acquired through Kaggle, first one will be easy; and data cleaning will be most likely minimal. After exploring the data, depending on our findings, feature reduction techniques may be required. Finally, different machine learning algorithms will be considered. Some of these might be Naïve Bayes, Decision Trees, Logistic Regression, and Support Vector Machines. Although some of our features might be correlated and violate our assumption, Naïve Bayes will be a good base to start. As our features are completely categorical, Decision Trees and Random Forests must be considered. Other algorithms might be included or excluded after a thorough examination of the dataset.

Motivation for This Dataset

On a personal level, I am not a huge fan of mushrooms. But one can still accept it is fascinating that two things that are extremely similar in appearance can kill you or feed you, depending on which one of the two you choose. It is compelling to learn more about this phenomenon.

On a scholarly level, although I have been using Linear Regression for years, classification algorithms are an area I am less experienced. Moreover, I have never studied a completely categorical dataset, and this is an opportunity in that regard - which was another motivation for selecting this dataset.