Discriminative vs Generative Classification:

A case study on the performance of Logistic Regression and Naive Bayes on different types of datasets

Overview

Classification and prediction is a ubiquitous problem tackled in statistics, machine learning, pattern recognition and data mining. Generative and discriminative learning are two of the major paradigms for solving prediction problems in machine learning, each offering important distinct advantages. These algorithm utilize a vastly different technique from each other in solving the classification problem and have their own advantages and drawbacks with respect to the other approach.

The goal of this project is to apply these algorithms on the several datasets, differentiating in size, features and the number of classes to identify the avenues of comparison with respect to performance as well as the empirical and theoretical advantages of each approach.

Background

Generative Classifiers

Generative classifiers, such as **Normal-based Discriminant Analysis** and the **Naive Bayes** classifier, model the joint distribution $P(x, y)$ of the measured features x and the class labels y factorized in the form $P(x|y)P(y)$, and learn the model parameters through maximization of the likelihood given by $P(x|y)P(y)$.

In other words, a generative classifier tries to learn the model that generates the data behind the scenes by estimating the assumptions and distributions of the model. It then uses this to predict unseen data, because it assumes the model that was learned captures the real model.

Discriminative Classifiers

Discriminative classifiers, such as logistic regression, model the conditional distribution $P(y|x)$ of the class labels given the features, and learn the model parameters through maximizing the conditional likelihood based on $P(y|x)$.

A discriminative classifier tries to model by just depending on the observed data. It makes fewer assumptions on the distributions but depends heavily on the quality of the data (For e.g. Is it representative? Is there a lot of data?).

Generative models allow you to make explicit claims about the process that underlies a dataset. For example, generative graphical models allow you to describe conditional dependencies between model parameters. If your model has a good fit to your data set, it strengthens your claim that your model accurately reflects the generative process that actually created the data that you are modeling.

However, since Generative classifiers learn about the conditional probability indirectly, they can get the wrong assumptions of the data distribution. Quoting Vapnik from Statistical Learning Theory –

"One should solve the [classification] problem directly and never solve a more general problem as an intermediate step [such as modeling P(X|Y)]."

In such a case discriminative model should be preferred over generative model.

Objective

We compare the generative and discriminative classifiers by applying Logistic Regression and Naïve Bayes algorithm on different datasets and gathering inferences based on their performance with respect to time, accuracy, error rate and various other factors on the data.