Generative models for imbalances data

Assume that we are interested is classifying cases into two classes based on some covariate information. We then have data that consist of a binary outcome y and a p-dimensional covariate vector $\mathbf{x}$. The classic way of doing the classification is then to specify a discriminative model, i.e. a model for $Y|\mathbf{X} = \mathbf{x}$, i.e. conditioning on the covariate values in order to avoid specifying a distribution for those. The logistic regression model is an example of such a model. The model parameters are then typically estimated using the conditional likelihood of $Y|\mathbf{X} = \mathbf{x}$.

An alternative is to use generative models for discrimination. The model for $Y|\mathbf{X} = \mathbf{x}$ is then specified indirectly via the conditional distributions of $\mathbf{X}|Y = y$ for $y = 0, 1$ and the marginal probability $\pi_Y = P(Y = 1)$, as

$$P(Y = 1|\mathbf{x}) = (1 - \pi_Y)p(\mathbf{x}|Y = 0) + \pi_Y p(\mathbf{x}|Y = 1).$$

In this setting, the parameters are typically estimated via the likelihood functions of $\mathbf{X}|Y = y$ separately for the two classes.

This project consists in investigating this type of approach in the setting of class imbalanced data. Data are considered class imbalanced if one of the classes (typically $y = 1$) constitutes a much smaller fraction of the data (20% or less) than the other class, which occurs in many applications such as fraud detection, medical applications with rare diseases, etc. The aim is then to construct a model for predicting the class of a new incoming case, based on a collection of previous cases, with an associated class indicator y.

As there is much more data from the class $y = 0$, it could be best to have a complex model for this class, and a (much) simpler model for the class $y = 1$, for which data are scarce. The aim of this project is to investigate how well such an approach works, compared to using the same model complexity for both classes, depending on how severe the class imbalance is, as well as other characteristics of the data, on simulated data. The assumed models for $\mathbf{X}|Y = y$ for $y = 0, 1$ will be multivariate normal, at least to begin with, with a possibility to expand to the more general case with a Gaussian copula and possibly non-normal marginal distributions.