Partial dependence functions

Partial dependence plots are a graphical way to analyze and quantify the marginal effect of one feature on the target response (output variable). The partial dependence functions are obtained by considering the interaction between the target and a feature while marginalizing out the other features in the input dataset. In other words, we can interpret partial dependence plots as the expected target response (in our case the predicted probability function) as a function of the considered feature.

More formally, calling X_i the feature with respect to which we want to compute dependence and X_{-i} the set of complementary features we define the partial dependence of the response f as:

$$pd_{X_i}(x_i) = E_{X_{-i}}[f(x_i, X_i)] = \int f(x_i, x_{-i}) p(x_{-i}) dx_{-i}$$

The integral above cannot be computed analytically and it is therefore approximated as:

$$pd_{X_i}(x_i) \approx \frac{1}{n_{samples}} \sum_{j=1}^{n} f(x_i, x^{(j)}{}_{-i})$$

where $x^{(j)}{}_{-i}$ is the value of the jth sample for features X_{-i}. The plot is produced when this integral is computed over several values of x_i.