Machine Learning

Machine Learning is the subfield of computer science that gives computers the ability to perform tasks without being explicitly programmed. There are several different tasks that fall under the domain of machine learning and several different algorithms for "learning". In this chapter, we are going to focus on Classification and two classic Classification algorithms: Naive Bayes and Logistic Regression.

Classification

In classification tasks, your job is to use training data with feature/label pairs $(\mathbf{x}, \text{y})$ in order to estimate a function $\hat{y} = g(\mathbf{x})$. This function can then be used to make a prediction. In classification the value of y takes on one of a \textbf{discrete} number of values. As such we often chose $g(\mathbf{x}) = \underset{y}{\operatorname{argmax}} \ \hat{P}(Y = y | \mathbf{X})$.

In the classification task you are given N training pairs: $\left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), \ldots, \left(\mathbf{x}^{(N)}, y^{(N)}\right)$ Where $\mathbf{x}^{(i)}$ is a vector of m discrete features for the ith training example and $y^{(i)}$ is the discrete label for the ith training example.

In our introduction to machine learning, we are going to assume that all values in our training data-set are binary. While this is not a necessary assumption (both naive Bayes and logistic regression can work for non-binary data), it makes it much easier to learn the core concepts. Specifically we assume that all labels are binary $y^{(i)} \in \{0, 1\} \ \forall i$ and all features are binary $x_j^{(i)} \in \{0, 1\} \ \forall i, j$.