1 Classification

1.1 Introduction

Classification is a supervised machine learning approach to categorizing data into **distinct number of classes** where we can assign label to each class. Given a set of data $\{x^{(i)}, y^{(i)}\}$ where x is the feature space in $m \times (n+1)$ dimension, y is the classfication output such $y \in \{0, 1\}$ for binary output or $\{1, 2, ...n\}$ for multiclass output. Classification algorithms are most used for Spam detection, Voice and image recognition, sentiment analysis, fraud detection and many more.

Machine learning classification algorithms come in different types and they include:

- Logistic regression
- Naive Bayes
- K-Nearest Neighbor
- Support vector machines
- Decision trees
- Boosted trees (Adaboost)
- Random Forest

1.2 Logistic Regession

Logistic regression is a discriminative model since it focuses only on the posterior probability of each class $Pr(Y|x;\beta)$. It is also a generalized linear model, mapping output of linear multiple regression to posterior probability of each class $Pr(Y|x;\beta) \in \{0, 1\}$. probabilty a data-sample belongs to class 1 is given by

$$Pr(Y = 1|X = x; \beta) = \sigma(z), where\ z = \beta^T x \tag{1}$$

$$P(Y = 1|X = x; \beta) = \sigma(\beta^T x) \tag{2}$$

where

$$Pr(Y = 1|X = x; \beta) + Pr(Y = 0|X = x; \beta) = 1 \tag{3}$$

$$Pr(Y = 0|X = x; \beta) = 1 - Pr(Y = 1|X = x; \beta) \tag{4}$$

Hence, probability that a data-sample belongs to class 0 is give by:

$$Pr(Y = 0|X = x; \beta) = 1 - \sigma(z) \tag{5}$$

$\sigma(z)$ is called the **logistic sigmoid function** and is give by

$$\sigma(z) = \frac{1}{1 + \exp^{-z}} \tag{6}$$

The uniqueness of this function is that it maps all real numbers $\mathbb{R}$ to range { 0, 1}.

Again, we know

$$log(odds(Pr(Y = 1|X = x; \beta))) = \frac{Pr(Y = 1|X = x; \beta)}{Pr(Y = 0|X = x; \beta)} = \frac{Pr(Y = 1|X = x; \beta)}{1 - Pr(Y = 1|X = x; \beta)} \tag{7}$$

Assuming $P(Y = 1|X = x; \beta) = p(x)$, the next most obvious idea is to let $logp(x)$ be a linear function of x, so that changing an input variable multiplies the probability by a fixed amount. This is done by taking a *log* transformation of $p(x)$.

Formally, $logit(p(x)) = \beta_0 + \beta^T x$ making

$$logit(p(x)) = log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta^T x \tag{8}$$

Simplifying for $p(x)$ and $1 - p(x)$ we have

$$\frac{p(x)}{1 - p(x)} = \exp(\beta_0 + \beta^T x) \tag{9}$$

$$p(x) = (1 - p(x)) \exp(\beta_0 + \beta^T x) \tag{10}$$

$$p(x) = \exp(\beta_0 + \beta^T x) - p(x) \cdot \exp(\beta_0 + \beta^T x) \tag{11}$$

$$p(x) + p(x) \cdot \exp(\beta_0 + \beta^T x) = \exp(\beta_0 + \beta^T x) \tag{12}$$

$$p(x)(1 + \exp(\beta_0 + \beta^T x)) = \exp(\beta_0 + \beta^T x) \tag{13}$$

$$p(x) = \frac{\exp(\beta_0 + \beta^T x)}{1 + \exp(\beta_0 + \beta^T x)} = \frac{\frac{1}{exp(\beta_0+\beta^T x)} \cdot \exp(\beta_0 + \beta^T x)}{\frac{1}{exp(\beta_0+\beta^T x)} + \frac{exp(\beta_0+\beta^T x)}{exp(\beta_0+\beta^T x)}} = \frac{1}{1 + \exp -(\beta_0 + \beta^T x)} \tag{14}$$

$$1 - p(x) = \frac{1}{1 + \exp(\beta_0 + \beta^T x)} \tag{15}$$

1.2.1 Log likelihood

Assume that $P(Y = 1|X = x; \beta) = P(x; \beta)$, for some function p parameterized by β. Further assume that observations are independent of each other. The conditional likelihood function is given by Bernoulli sequence:

$$\Pi_{i=1}^{n} Pr(Y = y_i|X = x_i; \beta) = \Pi_{i=1}^{n} p(x_i; \beta)^y (1 - p(x_i; \beta)^{(1-y_i)}) \tag{16}$$

The probability of a class is p, if $y_i = 1$, or $1 - p$, if $y_i = 0$. The likelihood is then

$$L(\beta_0, \beta) = \Pi_{i=1}^{n} p(x_i)^y (1 - p(x_i)^{(1-y_i)}) \tag{17}$$

taking the log of this likelhood we have

$$l(\beta_0, \beta) = \sum_{i=1}^{n} y_i logp(x_i) + (1 - y_i) log(1 - p(x_i)) \tag{18}$$

$$= \sum_{i=1}^{n} y_i logp(x_i) + y_i log(1 - p(x_i)) + log(1 - p(x_i)) \tag{19}$$

$$= \sum_{i=1}^{n} log(1 - p(x_i)) + y_i log \left(\frac{p(x_i)}{1 - p(x_i)} \right) \tag{20}$$

replace $log \left(\frac{p(x)}{1-p(x)} \right)$ with $\beta_0 + x \cdot \beta$ as seen in equation (8) and $(1 - p(x))$ with $\frac{1}{1+\exp(\beta_0+x\cdot\beta)}$. Hence,

$$l(\beta_0, \beta) = \sum_{i=i}^{n} log \left(\frac{1}{\exp(\beta_0 + x \cdot \beta)} \right) + y(\beta_0 + x \cdot \beta) \tag{21}$$

$$= \sum_{i=i}^{n} -log(1 + \exp(\beta_0 + x \cdot \beta)) + y(\beta_0 + x \cdot \beta) \tag{22}$$

$$\bigtriangledown l(\beta_0, \beta) = -\sum_{i=1}^{n} \frac{1}{1 + \exp(\beta_0 + x \cdot \beta)} x_{ij} \exp(\beta_0 + x \cdot \beta) + \sum_{i=1}^{n} y_i x_i \tag{23}$$

$$= \sum_{i=1}^{n} (y_i - p(x_i; \beta_0, \beta)) x_{ij} \tag{24}$$

Since this is a transcendental equation and there is no closed-form solution, we would apply the gradient ascent optimization algorithm.

1.2.2 Gradient ascent

Algorithm:
Using the cost function equal to the negative average of log-likelihood. eq(18)

$$\max_{\beta} J(\beta) = -\frac{1}{m} \left[\sum_{i=1}^{m} y_i logp(x_i) + (1 - y_i) log(1 - p(x_i)) \right] \tag{25}$$

Repeat {

$$\beta_{j+1} = \beta_j + \alpha \cdot \bigtriangledown J(\beta) \tag{26}$$

}simultaneously updating β_j

Where $\bigtriangledown J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (y_i - p(x_i; \beta_0, \beta)) x_{ij}$ and $p(x)$ remains logistic function given by $\frac{1}{1+\exp -(x \cdot \beta)}$. Newtons numerical optimization method is another approach to solving this problem- a case where the second order differential in required.