Multiclass Prediction

SoftMax Regression, One-vs-All & One-vs-One for Multi-class Classification

In Multi-class classification, we classify data into multiple class labels. Unlike classification trees and nearest neighbors, the concept of Multi-class classification for linear classifiers is not as straightforward. We can convert logistic regression to Multi-class classification using multinomial logistic regression or SoftMax regression; this is a generalization of logistic regression. SoftMax regression will not work for Support Vector Machines (SVM); One vs. All (One-vs-Rest) and One vs One are two other multi-class classification techniques that can convert most two-class classifiers to a multi-class classifier.

SoftMax Regression

SoftMax regression is similar to logistic regression, the SoftMax function converts the actual distances i.e. dot products of x with each of the parameters θ_i for K classes in the range from 0 to K-1. This is converted to probabilities using the following formula.

$$\text{softmax}(x, i) = \frac{e^{-\theta_i^T x}}{\sum_{j=1}^{K} e^{-\theta_j^T x}} \quad (1)$$

The training procedure is almost identical to logistic regression using cross-entropy, but the prediction is different . Consider the three-class example where $y \in \{0, 1, 2\}$ i.e y can equal 0,1,2. We would like to classify x. We can use the SoftMax function to generate a probability of how likely the sample belongs to each class. We then make a prediction using the $argmax$ function:

$$\hat{y} = argmax_i(softmax(x, i)) \quad (2)$$

Let's do an example, consider sample x_1, we will start by creating a table where each column will be the $i-th$ values of the SoftMax function. The index of each column is the same as the class.

probability of $\hat{y} = 0$	probability of $\hat{y} = 1$	probability of $\hat{y} = 2$
$softmax(x_1, 0)$	$softmax(x_1, 1)$	$softmax(x_1, 2)$
$i = 0$	$i = 1$	$i = 2$

Table 1. Each column will be the i-th values of the SoftMax function. The index of each column is the same as the class.

Let's add some real probabilities , this is the models estimate of how likely a sample belongs to each class.

0.97	0.02	0.01