

# Spark MLlib中算法总结

## 1. 线性模型

---

### 1.1 二分类 (Binary classification)

二分类算法是将目标分为两个类别，正例和负例。MLlib中包含两种线性二分类算法：线性支持向量机 (linear support vector machines) 和逻辑回归 (logistic regression)。对于这两种方法，MLlib支持L1和L2正则变体

#### 1.1.1 线性支持向量机 (SVMs)

线性支持向量机 ([SVMs](#)) 是用于大规模分类任务的标准方法，他的损失函数如下：

$$L(\mathbf{w}; \mathbf{x}, y) := \max\{0, 1 - y\mathbf{w}^T \mathbf{x}\}$$

线性SVMs在默认情况下使用L2正则化，同时也可选L1正则，在这种情况下问题就变成线性问题。

线性支持向量机算法输出SVM模型，输入一个未知的数据点 $\mathbf{x}$ ，模型根据 $\mathbf{w}^T \mathbf{x}$ 预测结果，默认情况下如果 $\mathbf{w}^T \mathbf{x} \geq 0$ 则输出为正，否则为负。

#### 1.1.2 逻辑回归

逻辑回归在二分类中广泛应用，损失函数表示如下：

$$L(\mathbf{w}; \mathbf{x}, y) := \log(1 + e^{-y\mathbf{w}^T \mathbf{x}})$$

逻辑回归算法输出为逻辑回归模型，给定一个数据点 $\mathbf{x}$ ，模型运用逻辑方程进行预测

$$f(z) = \frac{1}{1 + e^{-z}}$$

其中 $z = \mathbf{w}^T \mathbf{x}$ ，默认情况下，如果 $f(\mathbf{w}^T \mathbf{x}) > 0.5$ ，则输出为正，否则为负，与线性支持向量机不同，逻辑回归模型的输出 $f(z)$ 可以预测输出为正的的概率。

### 1.1.3 评价矩阵