

机器学习实验报告

学院: 软件学院

专业: 软件工程

|  |  |
| --- | --- |
| *作者:* | *指导教师:*  谭明奎 |
|  |  |
| *学号:*  202010100001 | *班级:* |

2020-10-24

# 逻辑回归和支持向量机

**摘要**

逻辑回归是这样的一个过程：面对一个分类问题，建立代价函数，然后通过优化方法迭代求解出最优的模型参数，然后测试验证我们这个求解的模型的好坏。

逻辑回归的思想是在之前的一般线性回归模型基础上添加logistic函数进行包裹，是一个被logistic方程归一化后的线性回归，只用于处理二分类问题。

而支持向量机(support vector machine，常简称为SVM，又名支持向量网络)是在分类与回归分析中分析数据的监督式学习模型与相关的学习算法。给定一组训练实例，每个训练实例被标记为属于两个类别中的一个或另一个，SVM训练算法创建一个将新的实例分配给两个类别之一的模型，使其成为非概率二元线性分类器。SVM模型是将实例表示为空间中的点，这样映射就使得单独类别的实例被尽可能宽的明显的间隔分开。然后，将新的实例映射到同一空间，并基于它们落在间隔的哪一侧来预测所属类别。

除了进行线性分类之外，SVM还可以使用所谓的核技巧有效地进行非线性分类，将其输入隐式映射到高维特征空间中。

当数据未被标记时，不能进行监督式学习，需要用非监督式学习，它会尝试找出数据到簇的自然聚类，并将新数据映射到这些已形成的簇。将支持向量机改进的聚类算法被称为支持向量聚类，当数据未被标记或者仅一些数据被标记时，支持向量聚类经常在工业应用中用作分类步骤的预处理。

## **介绍**

1. 对比理解梯度下降和批量随机梯度下降的区别与联系。
2. 对比理解逻辑回归和线性分类的区别与联系。
3. 进一步理解SVM的原理并在较大数据上实践。

## **方法和理论**

1实验介绍

在本次实验汇总，主要使用了两种分类模型对a9a进行分类，它们分别为逻辑回归与支持向量机。

1.1逻辑回归

1. 读取实验训练集和验证集。
2. 逻辑回归模型参数初始化（可以考虑全零初始化，随机初始化或者正态分布初始化）。
3. 选择Loss函数及对其求导，过程详见课件ppt。
4. 自行确定batch\_size大小，随机取出部分样本，求得部分样本对Loss函数的梯度 。
5. 使用SGD优化方法更新参数模型，鼓励额外尝试Adam优化方法。
6. 选择合适的阈值，将验证集中计算结果大于阈值的标记为正类，反之为负类。在验证集上测试并得到Loss函数值Lvalidation。
7. 重复步骤4-6若干次，画出Lvalidation随迭代次数的变化图。

1.2支持向量机

1. 读取实验训练集和验证集。
2. 逻辑回归模型参数初始化（可以考虑全零初始化，随机初始化或者正态分布初始化）。
3. 选择Loss函数及对其求导，过程详见课件ppt。
4. 自行确定batch\_size大小，随机取出部分样本，求得部分样本对Loss函数的梯度 。
5. 使用SGD优化方法更新参数模型，鼓励额外尝试Adam优化方法。
6. 选择合适的阈值，将验证集中计算结果大于阈值的标记为正类，反之为负类。在验证集上测试并得到Loss函数值Lvalidation。
7. 重复步骤4-6若干次，画出Lvalidation随迭代次数的变化图。

2逻辑回归

2.1实验模型

首先定义线性模型：



其中为权重矩阵，为自变量，Y为数据标签，b为偏置量，在本实验中，在X中增加一个特征，该特征数据初始化为1，作为偏置量的训练数据。因此模型可以简化为

然后定义逻辑函数：



由逻辑函数可得：



其中y为分类结果。

2.2损失函数

其中

即：

其中

2.3梯度求解



其中

2.4参数优化

输入：



输出：

随机挑选一个样本为依据的损失函数对的导数。



3支持向量机

2.1实验模型

给定一个两类分类器数据集，其中 𝑦(𝑖)∈{+1, 1}，其中 x(𝑖) 表示第 i 个样本的特征值组成的向量，𝑦(𝑖) 表示第i个样本的分类结果，因此如果两类样本是线性可分的，即存在一个超平面:



我们定义整个数据集 D 中所有样本中到分割超平面的距离最短的那个点的距离为间隔（Margin）



如果间隔越大，其分割超平面对两个数据集的划分越稳定，不容易受噪声点等 因素影响。支持向量机的目标是寻找一个超平面使得最大，即:



显然各个样本中落在平行于分割超平面的边界的点就是我们要找的，我们假设这些点刚好落在上,即公式为：



为了能够容忍部分不满足线性约束的样本，我们可以引入松弛变量 𝜉，将优化问题变为:



𝜉𝑖 可以理解为第 𝑖 个样本点离其正确类别点的边界的距离,

1. 如果𝜉𝑖 = 0则其刚好落在边界点，即其是支持向量，则分对了。

2. 如果0 < 𝜉𝑖 < 1则其在分割面的正确的一边，只是没在边界面内，也能正

确分类。

3. 如果𝜉𝑖 > 1则其分在了分割面的另一个类别，明显是错误分类了。

2.2损失函数

在本实验操作中，由于将偏置量b合并到了参数中了，因此可以直接获取损失函数



其中成为hinge损失函数

2.3梯度求解



2.4参数优化

输入：



输出：

随机挑选一个样本为依据的损失函数对的导数。



## **实验**

### 数据集

此部分展示数据集的相关信息，例如内容、数据数量、训练集、验证集等。

### 实现

实验中的所有详细实现：初始化，处理过程，结果，各种参数。一言以蔽之，清楚地描述做了什么工作和工作方式。图和表应进行标记和编号，如表I和图1所示。

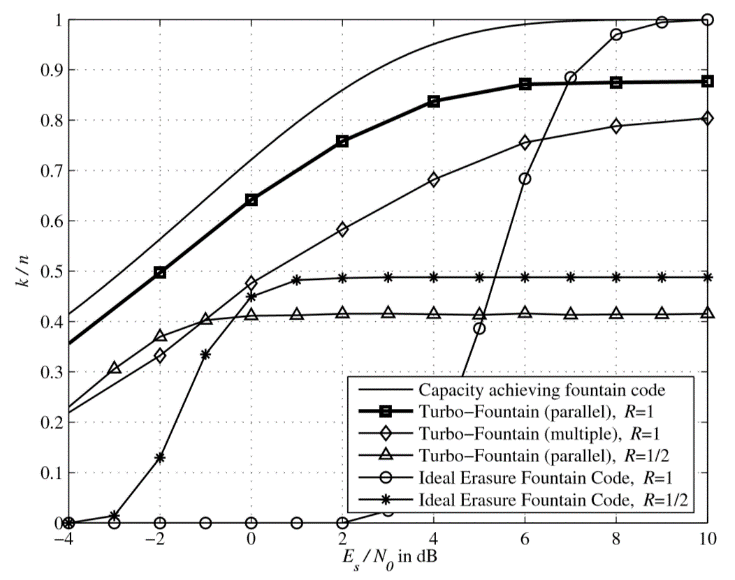
TABLE I

SIMULATION PARAMETERS

|  |  |
| --- | --- |
| Information message length | *k* = 16000 bit |
| Radio segment size | *b* = 160 bit |
| Rate of component codes | R = 1/3 |

## **结论**

本节总结了实验报告。也可以在此处写下在实验中的收获和灵感。

Figure. 1. Simulation results on the AWGN channel.