机器学习基础

2022.05.19

目录

1. 分类

监督学习的最主要类型

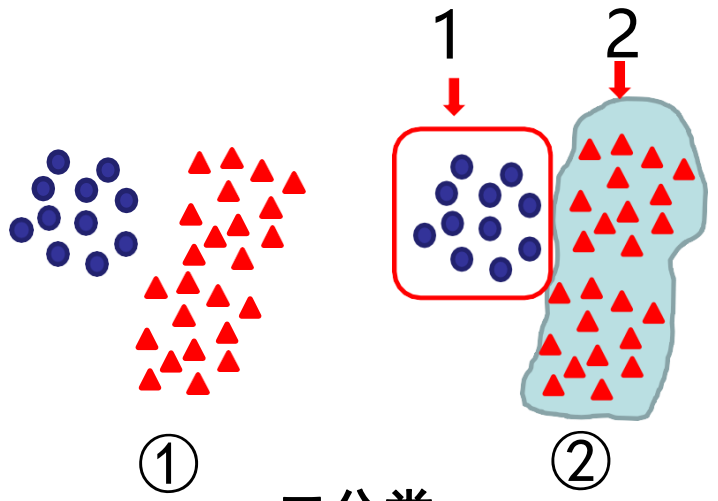
- ✓ 分类（Classification）
  - ✓ 身高1.85m，体重100kg的男人穿什么尺码的T恤？
  - ✓ 根据肿瘤的体积、患者的年龄来判断良性或恶性？
  - ✓ 根据用户的年龄、职业、存款数量来判断信用卡是否会违约？

标签离散

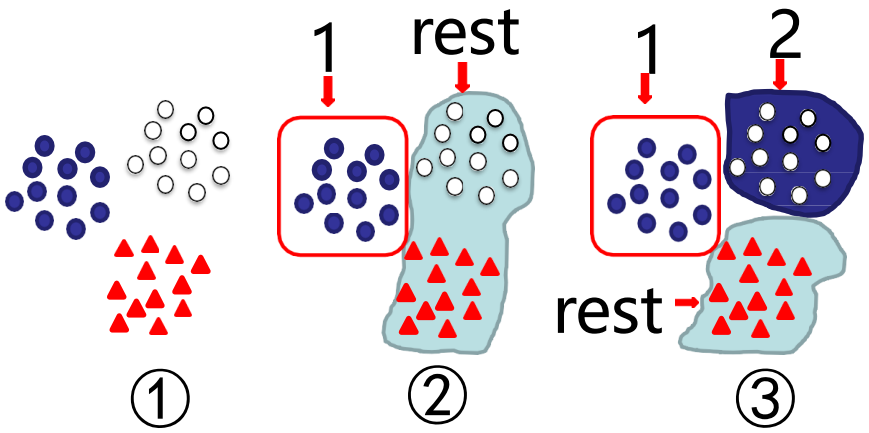
输入变量可以是离散的，也可以是连续的

1. 分类

- ✓ 二分类
  - ✓ 所有数据均可划分为两类中的任意一类
  - ✓ 只需要分类一次，步骤：①->②
- ✓ 多分类
  - ✓ 所有数据均可划分为多种类别
  - ✓ 每次先定义其中一类为类型1（正类），其余数据为负类（rest）；去掉类型1数据，剩余部分再次二分类，分成类型2和负类；如果有n类，那就需要分类n-1次。步骤：①->②->③->……



二分类



One-vs-All (One-vs-Rest)

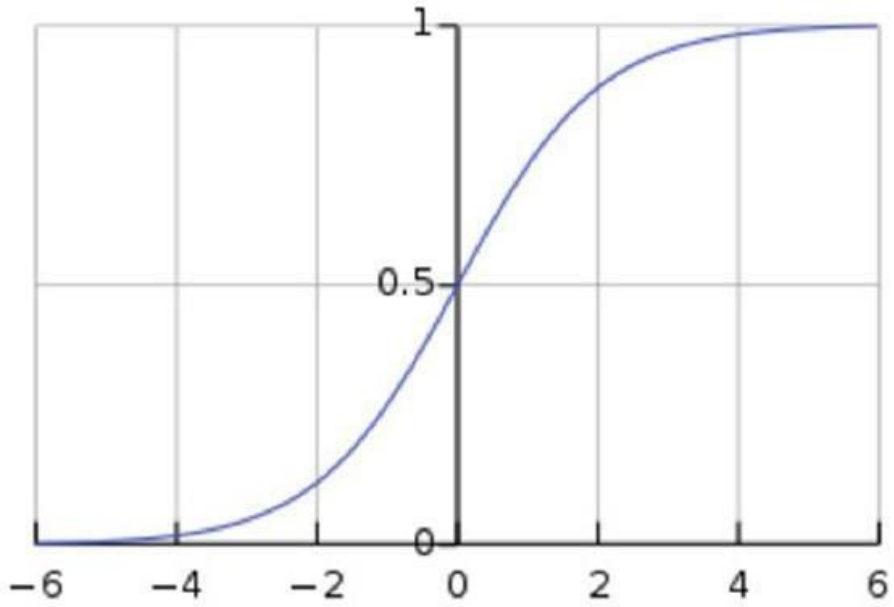
一对多 (一对余)

1. 分类

- ✓ Sigmoid函数
  - ✓ $\sigma(z)$代表一个常用的逻辑函数（logistic function）为S形函数（Sigmoid function)

$$\sigma(z) = g(z) = \frac{1}{1+e^{-z}}, z = w^T x + b$$

$$\sigma'(z) = \sigma(z)(1-\sigma(z))$$



  - ✓ 回归模型的假设函数

$$\mathrm{L}(\hat{y}, y) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}, z = w^T x + b$$

其中，y是真实的标签，$\hat{y}$是预测值。

- ✓ $\sigma(z)$大于等于0.5时，预测y=1
- ✓ $\sigma(z)$小于0.5时，预测y=0

1. 分类

- ✓ 逻辑回归求解

$$p(y=1 \mid x ; w)=h(x)$$

$$p(y=0 \mid x ; w)=1-h(x)$$

  - ✓ 假设一个二分类模型：

$$p(y \mid x; w) = (h(x))^{y}(1-h(x))^{1-y}$$

  - ✓ 逻辑回归模型的假设是：$h(x) = g(w^{T}x) = g(z)$
  - ✓ 逻辑函数（logistic function）公式为

$$g(z) = \frac{1}{1+e^{-z}}, g'(z) = g(z)(1-g(z))$$

  - ✓ 损失函数

$$\mathrm{L}(\hat{y}, y) = -y\log(\hat{y}) - (1-y)\log(1-\hat{y})$$

1. 分类

✓ 逻辑回归求解

✓ 损失函数

$$J(w) = -\frac{1}{m} l(w) = -\frac{1}{m} \sum_{i=1}^{m} (y^{(i)} \log(\hat{y}^{(i)}) + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}))$$

$$\hat{y}^{(i)} = \sigma(z^{(i)}) = \frac{1}{1 + e^{-z^{(i)}}}, z^{(i)} = w^T x^{(i)} + b$$

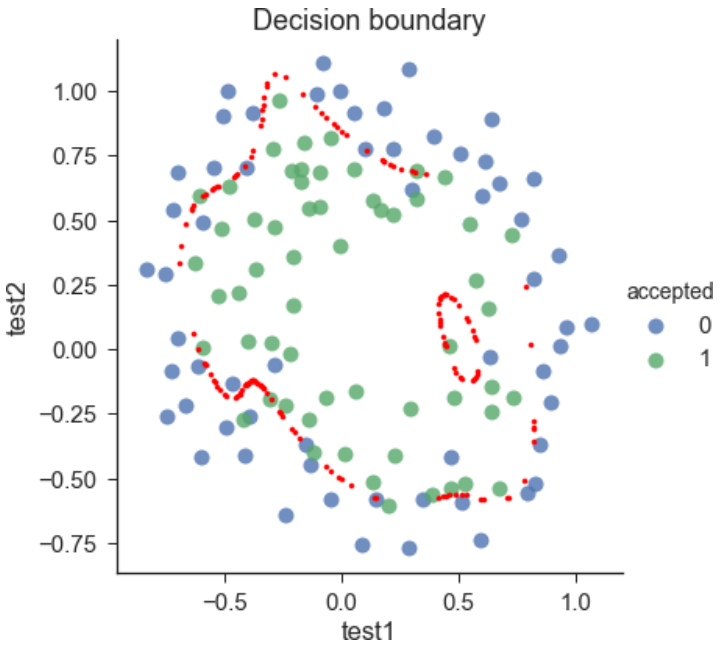
✓ 梯度下降

$$w_j := w_j - \alpha \frac{\partial J(w)}{\partial w} \quad \frac{\partial}{\partial w_j} J(w) = \frac{1}{m} \sum_{i=1}^{m} (\hat{y}^{(i)} - y^{(i)}) x_j^{(i)}$$

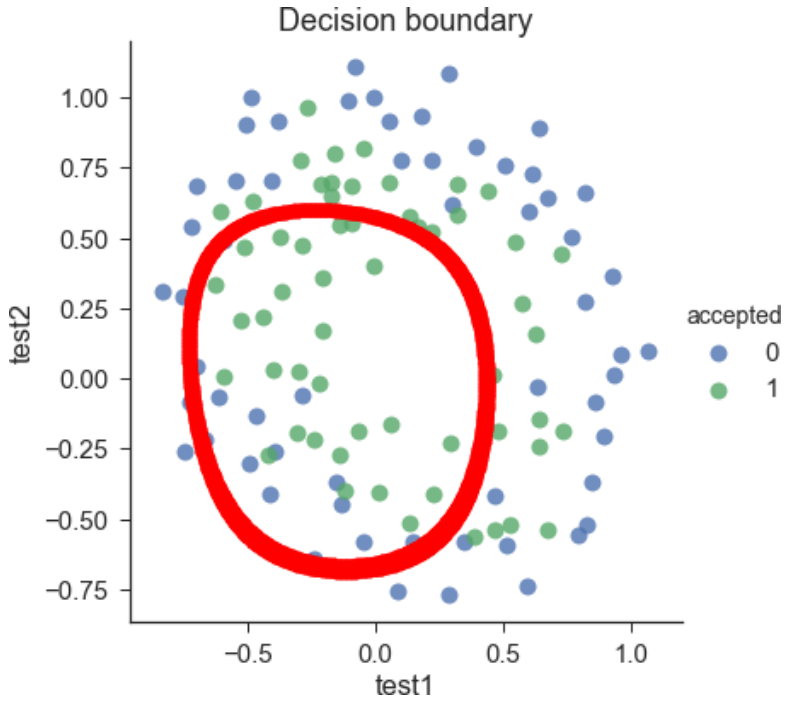
1. 分类

✓逻辑回归求解—正则化，防止过拟合

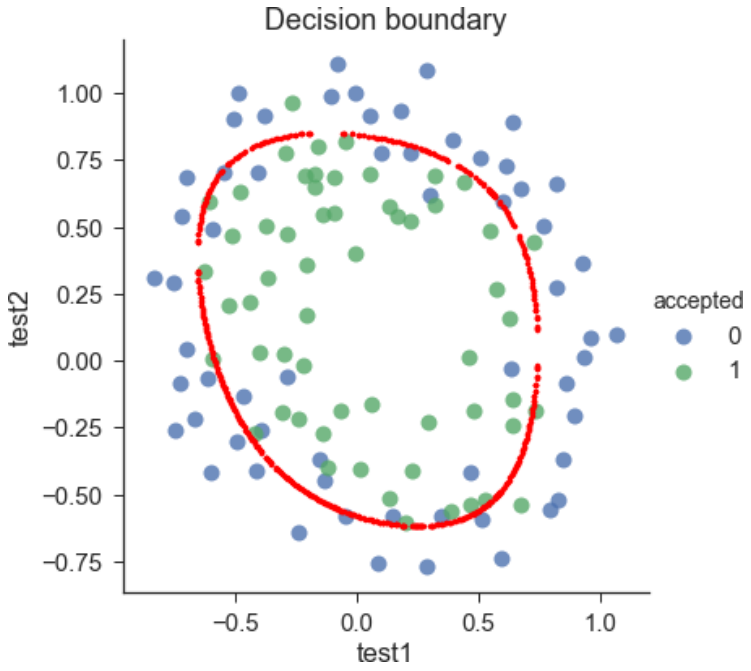
$$J(w)=-\frac{1}{m}l(w)=-\frac{1}{m}\sum_{i=1}^{m}(y^{(i)}\log(\hat{y}^{(i)})+(1-y^{(i)})\log(1-\hat{y}^{(i)}))+\frac{\lambda}{2m}\sum_{j=1}^{n}w_j^2$$



没有正则化，过拟合



正则化过度，欠拟合



适当的正则化

2. 实验任务

- 1. 在给定文本数据集完成大学生录取预测分类训练，画出训练loss曲线图、数据可视化图。（数据集课上发布）
- 2. （选做）在给定文本数据集完情感预测预测分类训练，此题不要求画曲线图。（数据集使用和朴素贝叶斯相同的数据集）

- 要求
  - 设计合适的网络结构，选择合适的损失函数，利用训练集完成网络训练，画出数据可视化图、loss曲线图，计算模型收敛后的分类准确率。
  - 需要提交简要报告+代码
  - 压缩包：学号_姓名_作业编号.zip，如 20331234_张三_实验7.zip。
  - 截止日期：2022.5.25　23:59

附录

矩阵求导： https://zhuanlan.zhihu.com/p/137702347

矩阵运算库Numpy教程：

https://www.runoob.com/numpy/numpy-tutorial.html

Matplotlib可视化教程:

https://www.runoob.com/matplotlib/matplotlib-tutorial.html

Thanks!