第4讲广义线性模型

主讲教师：欧新宇

February 21, 2020

Outlines

- 线性模型的基本概念
- 线性模式的可视化
- 线性回归（Linear Regression）
- 岭回归（Ridge Regression）
- 套索回归（Lasso Regression）

线性模型的基本概念

线性模型是统计学中的一个术语，被广泛应用到基于机器学习的多个领域中，甚至被很多研究人员集成到诸如**神经网络**的复杂系统中。在机器学习中，常见的线性模型包括：线性回归（Linear Regression）、岭回归（Ridge Regression）、套索回归（Lasso Regression)、逻辑回归（Logistic Regression）、线性SVC等。

- **线性模型的数学表达：**

$$\hat{y} = w[0] * x[0] + w[1] * x[1] + \ldots + w[p] * x[p] + b$$

其中，

- $x[0], x[1], \ldots, x[p]$ 是数据集中的特征变量
- 参数 p 表示每个样本都有 p 个特征；
- w 和 b 是模型的参数；
- $\hat{y}$ 是模型对给定数据的预测结果，$\hat{y}$ 读作：$y\ hat$，一般来说 hat 表示估计值。

线性模型的基本概念

- 若数据只有一个特征变量，则线性模型可以被简化为:

$$\hat{y} = w[0] * x[0] + b$$

对于简化模型来说，$\hat{y}$ 就是**一条直线**的方程。

✓ $w[0]$ 是直线的斜率，也称为权重；

✓ b 是 y 轴的偏移量（截距）

模型的预测可以看作输入特征的加权和，参数 w 代表的是每个特征的权值。

线性模型的基本概念

假设有一条直线，其数学表达式为: $y = 0.5x + 3$，将其可视化后可以得到:

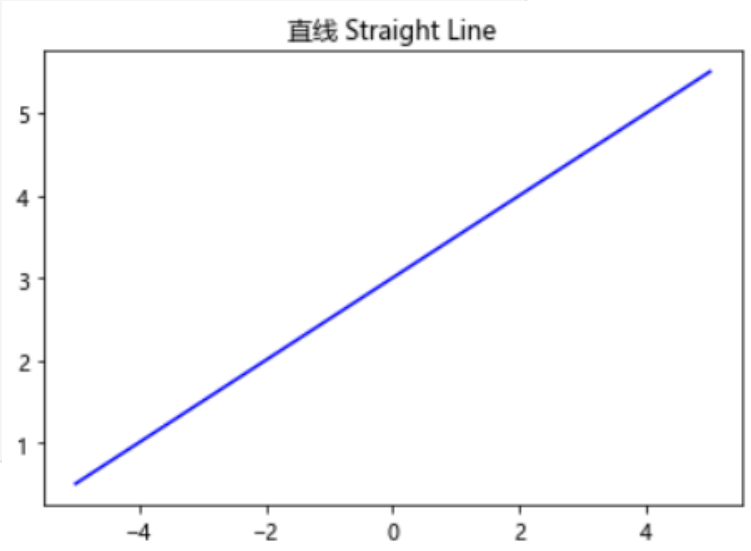
[18]:
```
import numpy as np
import matplotlib.pyplot as plt

# 配置参数使 matplotlib绘图工具可以显示中文
plt.rcParams['font.sans-serif'] = [u'Microsoft YaHei']

# 设置自变量 x：令x为-5到5之间，元素数量为100的等差数列
x = np.linspace(-5, 5, 100)

# 按照方程的数学表达式，定义直线方程
y = 0.5*x + 3

# 设置绘图内容的基本参数
plt.plot(x, y, c = "blue")
# 设置图的题目
plt.title("直线 Straight Line")
# 激活绘图功能，在坐标轴上显示直线
plt.show()
```



线性模型的基本概念

- **线性模型的可视化**
- 生成数据集
- 拆分数据集
- 创建模型并拟合训练集数据
- 输出模型参数（即方程的斜率和截距）
- 可视化方程曲线

线性模型的基本概念

线性模型的优缺点

- **优点**

建模速度快，不需要复杂的计算，特别是在大数据量下依然具有较快的运算速度. 可以根据系数给出每个变量的理解和解释

- **缺点**

不能很好拟合非线性数据，因此需要先判断变量间是否具有线性关系.

线性模型的基本概念

- **为什么线性回归模型依然有效?**
  - 线性回归能够模拟的数据**远不止**线性关系，并且回归中的"线性"指的是系数的线性，通过特征的非线性变换及广义线性模型的推广，输出和特征之间可以是高度非线性的;
  - 线性模型的易解释性让它在物理学、经济学、商学等领域具有不可替代的地位;
  - 逻辑回归（Logistics Regression）目前也是基于深度模型的**目标检测（Detection）**任务中最常用的回归器。

线性回归算法的用法

- **线性回归（Ch0405LinearRegression）**
  - 线性回归模型
  - 线性回归模型（带噪声数据）
  - 线性回归模型（糖尿病数据集）
- **岭回归（Ch0408RidgeRegression）**
  - 糖尿病数据集
  - 波士顿房价数据集
  - 超参数alpha对性能的影响
  - 超参数对训练参数的影响
  - 训练集大小对模型性能的影响
- **套索回归（Ch0413LassoRegression）**
  - 套索回归模型
  - 套索回归和岭回归的对比

欧老师的联系方式

读万卷书 行万里路 只为最好的修炼

QQ：14777591 (宇宙骑士)

Email：ouxinyu@alumni.hust.edu.cn

Tel：18687840023