机器学习笔记

目录

[1. 机器学习算法的验证与评价 2](#_Toc530005910)

[2. Linear regression 线性回归 2](#_Toc530005911)

[3. Logistic regression 2](#_Toc530005912)

[4. KNN 4](#_Toc530005913)

[5. PCA（Principle Component Analysis）主成分分析 4](#_Toc530005914)

[6. LDA 5](#_Toc530005915)

[7. MDS（Multiple Dimensional Scaling） 5](#_Toc530005916)

[8. ISOMAP 6](#_Toc530005917)

[9. LLE(Locally Linear Embedding) 6](#_Toc530005918)

[10. Laplacian Eigenmap 7](#_Toc530005919)

## 1. 机器学习算法的验证与评价

留出法

交叉验证法

自助法

查准率、查全率 P-R曲线

F1度量

ROC&AUC

## 2. Linear regression 线性回归

多变量线性回归模型如下



作如下变换，则可得到变换后的模型：







下面求最优的参数









则得到多变量线性回归模型如下：



## 3. Logistic regression

sigmoid函数如下：



如多变量线性回归，代入



作变换







接下来需要确定，采用极大似然法



取对数，得：



利用







求得：



于是，得到：  


等价于



利用牛顿迭代法求



即





## 4. KNN

## 5. PCA（Principle Component Analysis）主成分分析

对于给定数据集，希望使用一个超平面来表达，主要由2个准则（最近重构性、最大可分性），使用这2者进行表达可以得到等价的问题表达。

输入：样本集，原维度，目标维度

输出：投影矩阵，进而降维样本集

（1）最近重构性，即通过降维后的新坐标和投影矩阵重构出的坐标要和原维度坐标的距离最小



（2）最大可分性，即投影后样本方差最大



使用拉格朗日乘子法进行求解：







即为矩阵特征值求解问题，对进行特征值分解后选取最大的个特征值对应的特征向量组成投影矩阵

这里这个协方差矩阵一定是实对称矩阵，所以一定可以进行特征值分解，

其中为对角线元素为特征值的对角矩阵，是正交矩阵（这是因为n阶实对称矩阵必然有n个线性无关的特征向量，于是必然能通过正交化方法构造出正交矩阵）

## 6. LDA

## 7. MDS（Multiple Dimensional Scaling）

算法过程

输入：

样本集，由此计算出距离矩阵，其中

原本的维度为，要求降维到维度

输出：

降维后的样本集

设矩阵

则可以通过如下公式计算矩阵：



其中



计算好后进行特征值分解，，是特征值对角阵

考虑到



取其中最大的个特征值和对应的特征向量，得到和



得到

## 8. ISOMAP

## 9. LLE(Locally Linear Embedding)

首先求系数矩阵



得到最优化问题



设

使用拉格朗日乘子法，得到



于是



又有





所以



接下来求降维后的样本集



得到最优化问题：



设

由拉格朗日乘子法得：



因此求出最大的指定维度的特征向量即组成降维后的样本集

## 10. Laplacian Eigenmap

有个邻接矩阵W，降维后数据集Z，最小化如下式子



可得到最优化问题



其中

为拉普拉斯矩阵

使用拉格朗日乘子法，得到



求解此广义特征值问题，得到最大的那些特征向量即组成降维后的数据集Z

可通过转化为一般特征值求解问题