

Machine Learning God Knows

Fduzjrqlw

目录

1	机器学习理论	7
1.1	简述机器学习问题.	7
1.2	误差分解公式.	7
1.3	偏差和方差	7
1.4	风险 (risk), 经验风险 (empirical risk) 的含义.	7
1.5	过拟合, 欠拟合和误差的关系.	7
1.6	估计的无偏, 有效, 一致分别在说什么?	7
1.7	为什么样本方差的分母是 $n - 1$?	7
2	数据预处理	8
2.1	归一化 (normalization) 和标准化 (standardization), 以及它们的相同点和不同点.	8
2.2	为什么要使用归一化/标准化?	8
2.3	什么时候用归一化, 什么时候用标准化?	8
2.4	一定要归一化吗? 举出一些不需要归一化的例子.	8
2.5	处理缺失值一般有哪些方法?	8
3	特征工程	8
3.1	特征抽取的目的.	8
3.2	独热编码.	8
3.3	TF-IDF 的公式和含义?	8
4	类别不均衡	8
4.1	什么叫类别不平衡问题?	8
4.2	类别不平衡问题用什么评估?	9

4.3	解决不平衡问题有哪些方法?	9
4.4	欠采样和过采样的缺点?	10
5	性能度量指标 (metrics)	10
6	正则化	10
6.1	做正则化的原因.	10
6.2	模型复杂程度的评价指标.	10
6.3	线性回归需要对偏差项 (bias term) 做正则吗?	10
6.4	限制经验风险最小化 (Constrained ERM) v.s. 惩罚项经验风险最小化 (Penalized ERM) 的等价性.	10
6.5	L_1 正则的作用和原因.	10
6.6	L_2 正则的作用和原因.	11
6.7	为什么要引入 Elastic Net 中的 L_1 和 L_2 的组合正则.	11
6.8	dropout 的原理和过程?	11
6.9	dropout 起到正则化作用和原因?	12
7	优化算法	12
7.1	GD,SGD 与 mini_batch GD 之间的区别和联系.	12
7.2	什么时候用 SGD?mini_batch GD 中 batch_size 的选择.	12
7.3	梯度下降算法的收敛率 (convergence rate).	12
7.4	SGD 的使用技巧.	12
7.5	SGD 的理论依据.	12
7.6	如果测试集数据增加 100 倍,batch 大小 m 需要调整吗?	12
7.7	在线学习的动机和方法.	13
7.8	次梯度和次梯度下降算法.	13
7.9	SMO 算法	13
7.10	SGDM 算法	13
7.11	NAG 算法	13
7.12	AdaGrad 算法	13
7.13	AdaDelta 算法	14
7.14	Adam 算法	14
7.15	Nadam 算法	14
7.16	指数移动平均的偏差修正是什么?	15

7.17 Adam 的缺点有哪些?	15
7.18 如何解决 Adam 的问题?	15
7.19 优化算法的常用 trick 有哪些?	16

8 模型 16

8.1 支持向量机 (Supported Vector Machine)	16
8.1.1 SVM 的损失函数是什么?	16
8.1.2 SVM 的推导过程?	16
8.1.3 为什么要引入对偶问题? 为什么要引入松弛变量 ξ ?	16
8.1.4 Slater 条件是什么, 如何验证 SVM 问题满足 Slater 条件?	16
8.1.5 叙述 SVM 问题的 KKT 条件?	16
8.1.6 模型偏置 bias term b 的作用?	17
8.1.7 什么叫核方法?	17
8.1.8 核化 (kernelized) 的好处有哪些?	17
8.1.9 表示定理以及其作用?	17
8.1.10 如何验证核函数?	17
8.1.11 什么叫径向基函数?	17
8.1.12 RBF 核对应特征空间的维数.	18
8.1.13 正则化 RBF 核.	18
8.1.14 参数 γ 以及 C 的作用	18
8.1.15 叙述 SVM 问题的 KKT 条件?	18
8.1.16 什么是 SVM 的退化?	18
8.1.17 如何构造新的输入数据, 使得一个线形可分的 SVM 问题退化?	18
8.1.18 SVM 的优化算法	19
8.1.19 核函数的选择	19
8.1.20 SVM 的优缺点	19
8.2 感知机 (perceptron)	20
8.2.1 感知机的工作原理?	20
8.2.2 感知机与 SVM 的异同点有哪些?	20
8.2.3 感知机如何优化?	20
8.2.4 感知机解的性质	20
8.3 决策树 (Decision Tree)	20

8.3.1	简述决策树的工作原理.	20
8.3.2	决策树区域的特性.	21
8.3.3	叙述决策树的建树过程?	21
8.3.4	不纯度是什么?	21
8.3.5	Gini 值的含义?	21
8.3.6	Gini 值和熵的联系?	22
8.3.7	什么是多变量决策树?	22
8.3.8	ID3 决策树的特点有哪些?	22
8.3.9	C4.5 决策树有哪些细节?	22
8.3.10	CART 有哪些细节?	23
8.3.11	决策树的缺失值问题和解决策略?	25
8.3.12	决策树的剪枝策略有哪些?	25
8.3.13	ID3,C4.5,CART 之间的差异有哪些?	25
8.3.14	sklearn 关于决策树的缺失值策略如何实现的?	25
8.3.15	关于决策树的调参, 有哪些技巧?	25
8.4	Bagging	25
8.4.1	统计量 (statistic) 与点估计 (point estimator)	25
8.4.2	估计量的方差有什么含义?	25
8.4.3	如何平衡每组的样本个数与组数	25
8.4.4	自助采样 (bootstrap)	26
8.4.5	Bagging 为什么用平均法? 原理是什么?	26
8.4.6	包外误差估计 (Out-of-Bag Error Estimation) 是什么?	26
8.4.7	Bagging 的适用性?	26
8.4.8	bootstrap 的方差估计	26
8.5	随机森林 (Random Forest)	27
8.6	Boosting	27
8.6.1	简单介绍下 boosting?	27
8.7	Adaboost	27
8.7.1	Adaboost 算法的流程及其推导?	27
8.7.2	Adaboost 的损失函数为什么是指数函数?	27
8.7.3	分类替代函数	27
8.7.4	Adaboost 中权重的含义?	28
8.7.5	如何求加权最优化问题?	28

8.7.6	Adaboost 算法终止的条件?	28
8.7.7	Adaboost 训练集误差的上界估计?	28
8.7.8	为什么 Adaboost 当训练集上 error 变为 0 时, 继续训练可以降低测试集上的 error?	28
8.7.9	Adaboost 什么情况下会过拟合严重?	28
8.7.10	如何缓解 Adaboost 的过拟合现象?	28
8.7.11	前向分布算法和坐标梯度下降法的联系和区别?	29
8.8	主成分分析 (Principal Component Analysis)	29
8.8.1	PCA 的数学推导步骤?	29
8.8.2	PCA 的假设条件是什么?	29
8.8.3	PCA 的核化是什么?	29
8.8.4	PCA 的使用技巧?	29
8.8.5	PCA 的四重境界分别是哪四重?	29
8.9	T-SNE 降维	30
8.9.1	T-SNE 算法步骤是什么?	30
8.9.2	T-SNE 与 PCA 的区别?	30
8.10	线性判别分析 (Linear Discriminate Analysis)	30
8.10.1	LDA 的基本假设	30
8.10.2	LDA 的基本推导	30
8.10.3	LDA 的优缺点有哪些?	31
8.10.4	为什么 LDA 的降维之后的维度最多为 $C - 1$	31
8.11	K 近邻 (Kth Nearest Neighbor)	31
8.11.1	KNN 的算法流程?	31
8.11.2	KNN 的注意事项?	31
8.11.3	KNN 的缺点有哪些?	31
8.11.4	超参数 k 如何选取?	31
8.11.5	KNN 的有哪些改进地方?	31
8.11.6	KD-Tree 的建树和搜索过程?	31
8.12	XGBoost(Extreme Gradient Boosting)	32
8.12.1	XGB 的基本思想是什么?	32
8.12.2	XGB 的损失函数以及分裂函数推导?	32
8.12.3	XGB 处理缺失值的策略是什么?	32
8.12.4	XGB 划分分裂点策略?	32

8.12.5	为什么用 h_i 作为权重?	33
8.12.6	XGB 的加速策略?	33
8.12.7	XGB 的参数有哪些?	33
8.12.8	XGB 的调参过程?	34
8.12.9	XGB 的优缺点?	34
8.12.10	XGB 为什么不适合处理类别特征?	34
8.13	LGB(Light Gradient Boosting machine)	35
8.13.1	LGB 的基本思想?	35
8.13.2	LGB 的 EFB 算法 (Exclusive Feature Bunds)	35
8.13.3	LGB 的 GOSS 算法是什么?	35
8.13.4	直方图算法是什么?	36
8.13.5	直方图算法的具体细节?	36
8.13.6	Leaf-wise 和 Level-wise 的区别?	36
8.13.7	处理类别特征的策略?	36
8.13.8	LGB 是否有用到二阶导数信息?	37