CIE架构设计

1. 数据读入DataReader
2. csv、excel
3. mongodb
4. mysql
5. json
6. 数据结构标准化。包括稀疏存储和非稀疏存储
7. 分类/回归问题数据
8. 聚类问题数据
9. 标注问题/语料形式
10. EDA
11. 目标统计
12. 特征统计：特征直方图、目标特征图
13. 变量-变量相关性热图

5、特征工程

1. 数据预处理
2. 缺省值：统计；舍弃，NAN，拟合
3. 哑变量onehot
4. 数据变换：log、指数、box-cox
5. 归一化/标准化
6. 离散化
7. 特征选择
8. Filter: Variance, Correlation, Chi2, Entropy
9. Wrapper: RFE
10. Embedded: Regularization, Tree
11. 降维
12. PCA
13. LDA

6、模型选择

1. Classification
2. LinearSVC/SVC
3. NB
4. KNeighborsClassifier
5. SGDClassifier
6. LogisticRegression
7. Kernel approximation：RBF, AdditiveChi2, SkewedChi2
8. EnsembleClassifiers(GradientBoosting, Adaboost, bagging, extraTrees, randomforest, voting, xgboost, lightgbm)
9. Regression
10. LinearRegression
11. SVR
12. EnsembleRegressors(GradientBoosting, Adaboost, bagging, extraTrees, randomforest, xgboost, lightgbm)
13. SGDRegressor
14. ALS-WR
15. Cluster
16. KMeans
17. Spectral Clustering
18. GMM

7、模型训练

1. 上采样/下采样
2. 学习曲线
3. grid search/cross validation

8、模型融合

1. voting
2. bagging
3. stacking
4. blending

9、模型评价

1. 准确率
2. 精准、召回、F1
3. AUC
4. 敏感度、特异度
5. 混淆矩阵