小圆蓝细胞瘤亚型分类

Classification of small, round blue-cell tumors subtypes by machine learning models

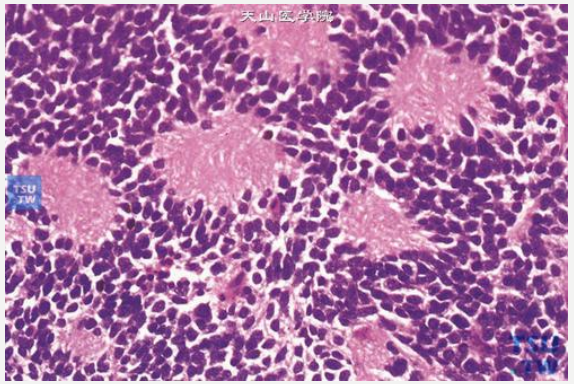
OVERVIEW

- 小圆蓝细胞瘤(SRBCTS)亚型
- “扁平”的数据集

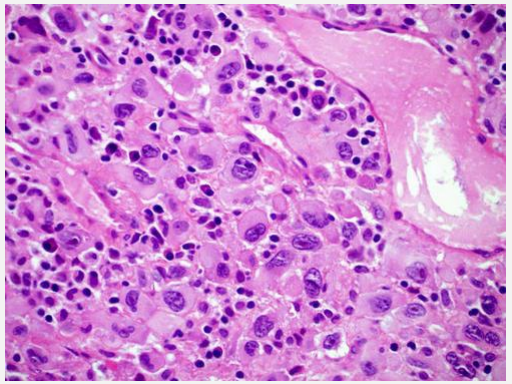
小圆蓝细胞瘤(SRBCTS)亚型

- NB-neuroblastoma 神经母细胞瘤
- RMS-rhabdomyosarcoma 横纹肌肉瘤
- NHL-non-Hodgkin lymphoma 非霍金淋巴瘤
- EWS-the Ewing family of tumors 尤因肿瘤群

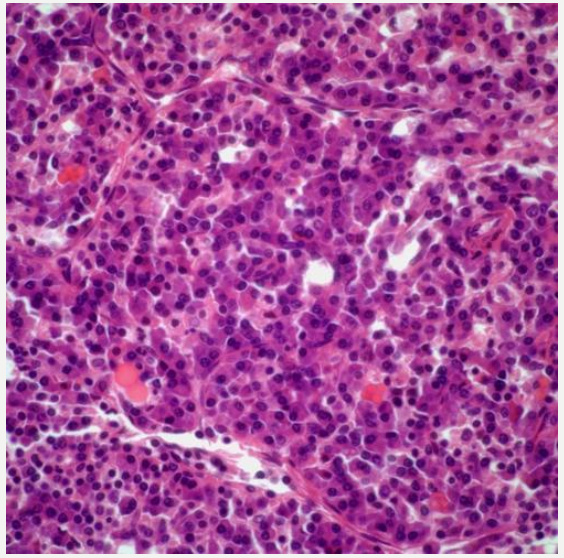
亚型	NB	RMS	NHL	EWS	合计
Train	12	20	8	23	63
Test	6	5	3	6	20
合计	18	25	11	29	83



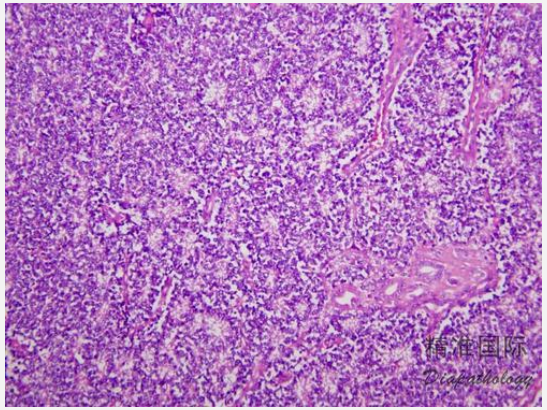
NB



RMS



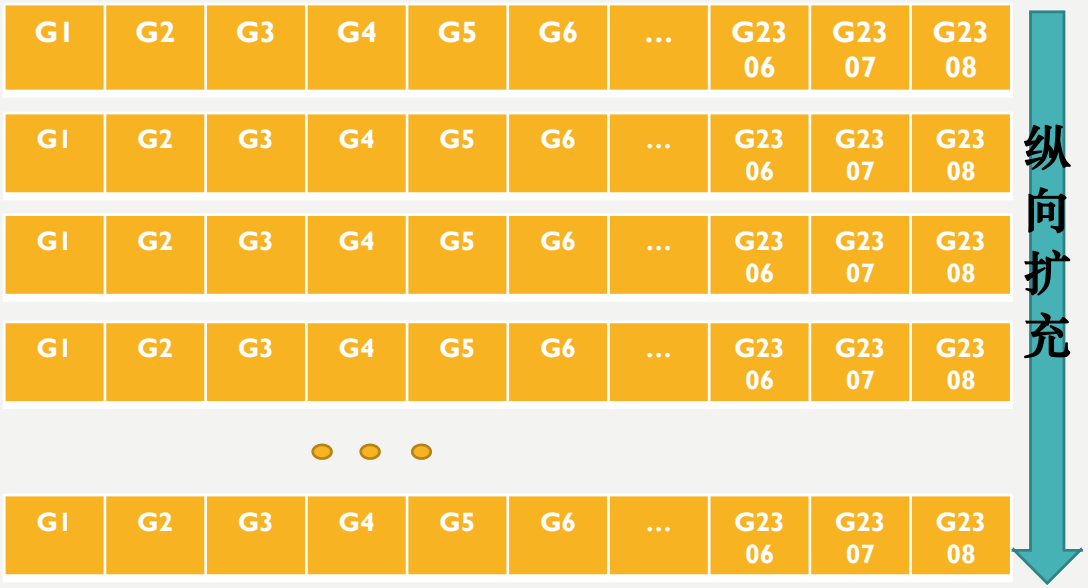
NHL



EWS

“扁平”的数据集


G1 G2 G3 G4 G5 G6 … G2306 G2307 G2308
G1 G2 G3 G4 G5 G6 … G2306 G2307 G2308
G1 G2 G3 G4 G5 G6 … G2306 G2307 G2308
63个样本
2308个基因

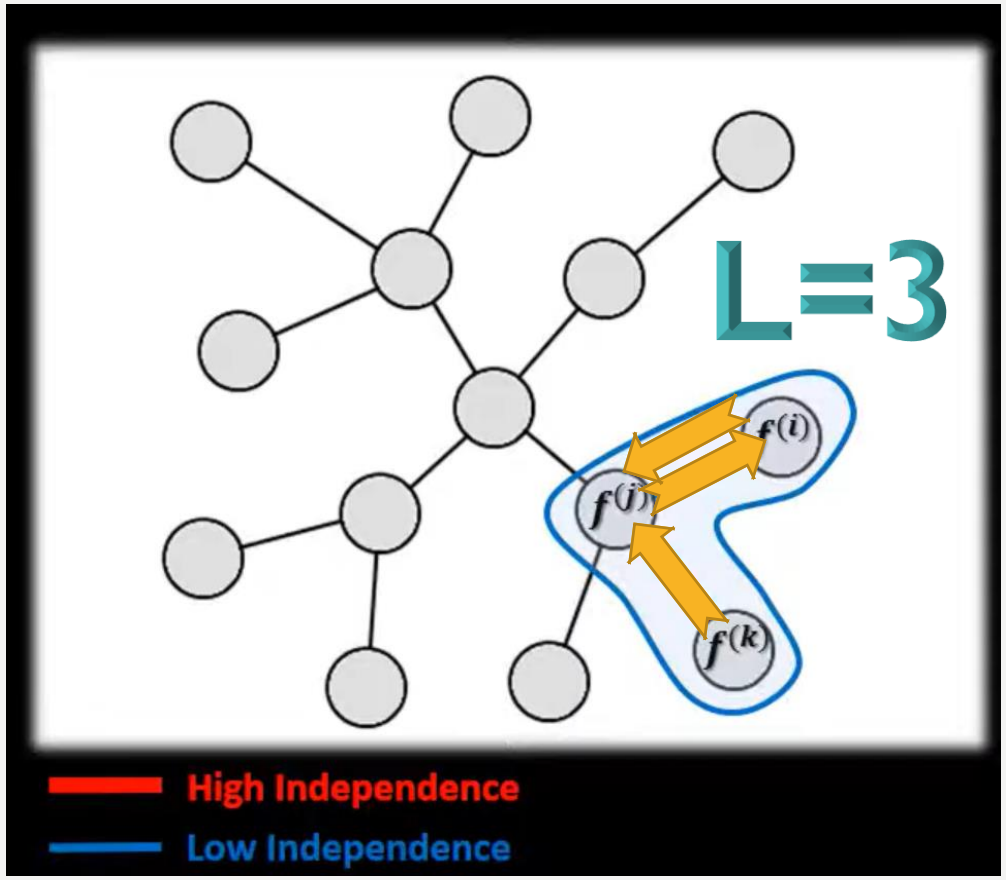

G1 G2 G3 G4 G5 G6 … G2306 G2307 G2308
纵向扩充


横向降维
G1 G2 G3 … G40
G1 G2 G3 … G40
G1 G2 G3 … G40

VISUALIZE & DIMENSIONALITY REDUCTION

- INFINITE FEATURE SELECTION降维算法
- 降维后效果

降维算法

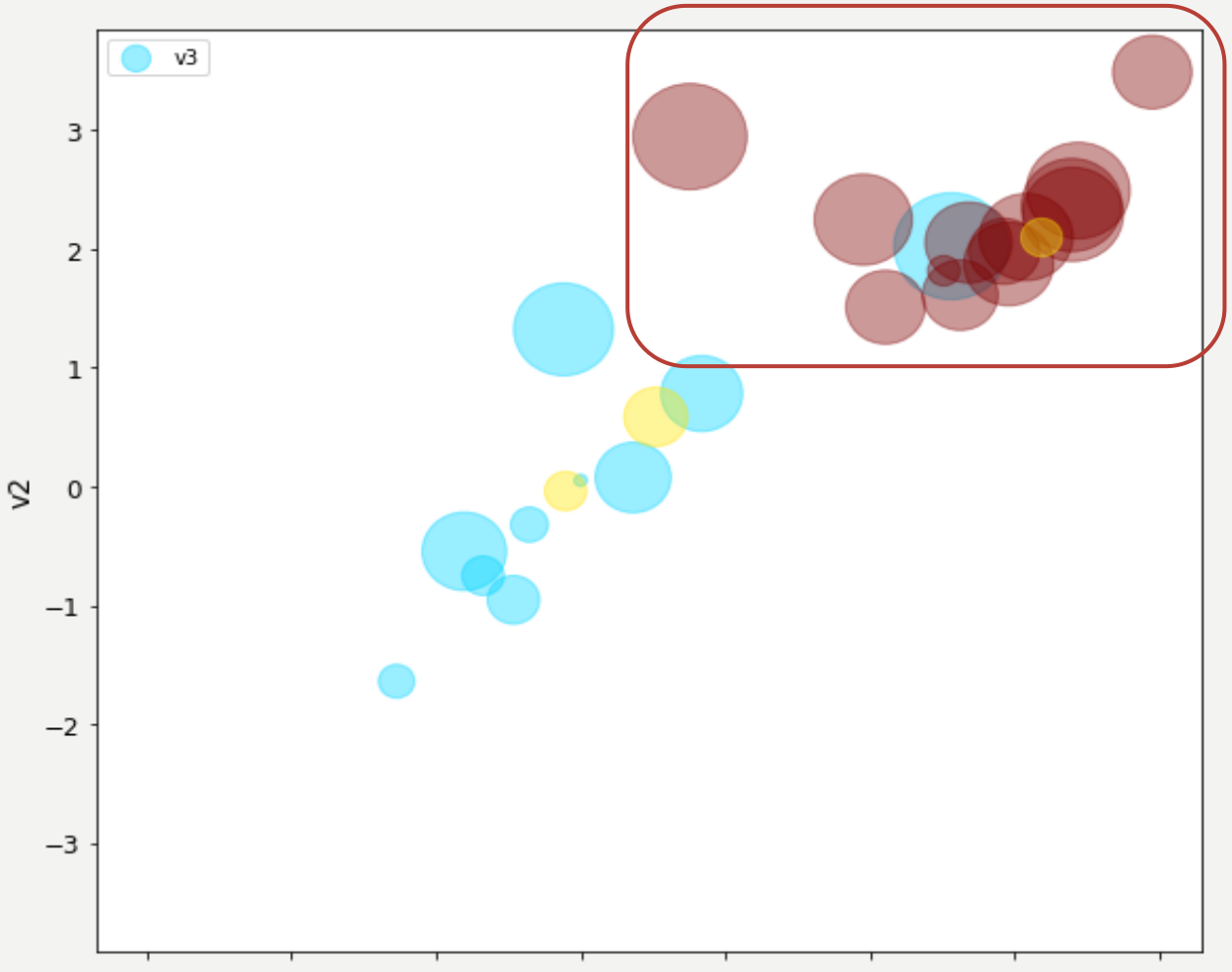


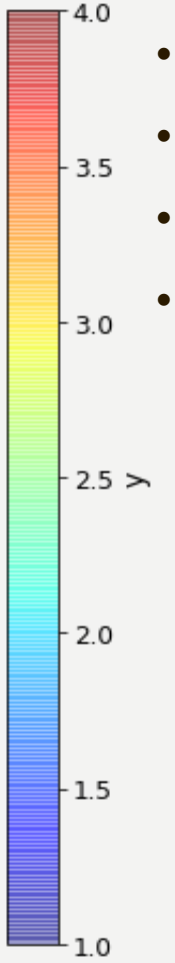
- **图中元素所对应的数据特性:**
- 特征子集= 图中的路径
- 点 = 特征
- 边表示两个特性的相互独立性
- 路径的能量 = 其边的分值乘积。

- **算法目标:**
- 高能量路径 = 获得高分值边 = 高独立性特征
- Inf-FS 考虑给定长度L时的所有路径:

$$S_l(\mathrm{i}) = \Sigma_j \in vA^l(\mathrm{i}, j)$$

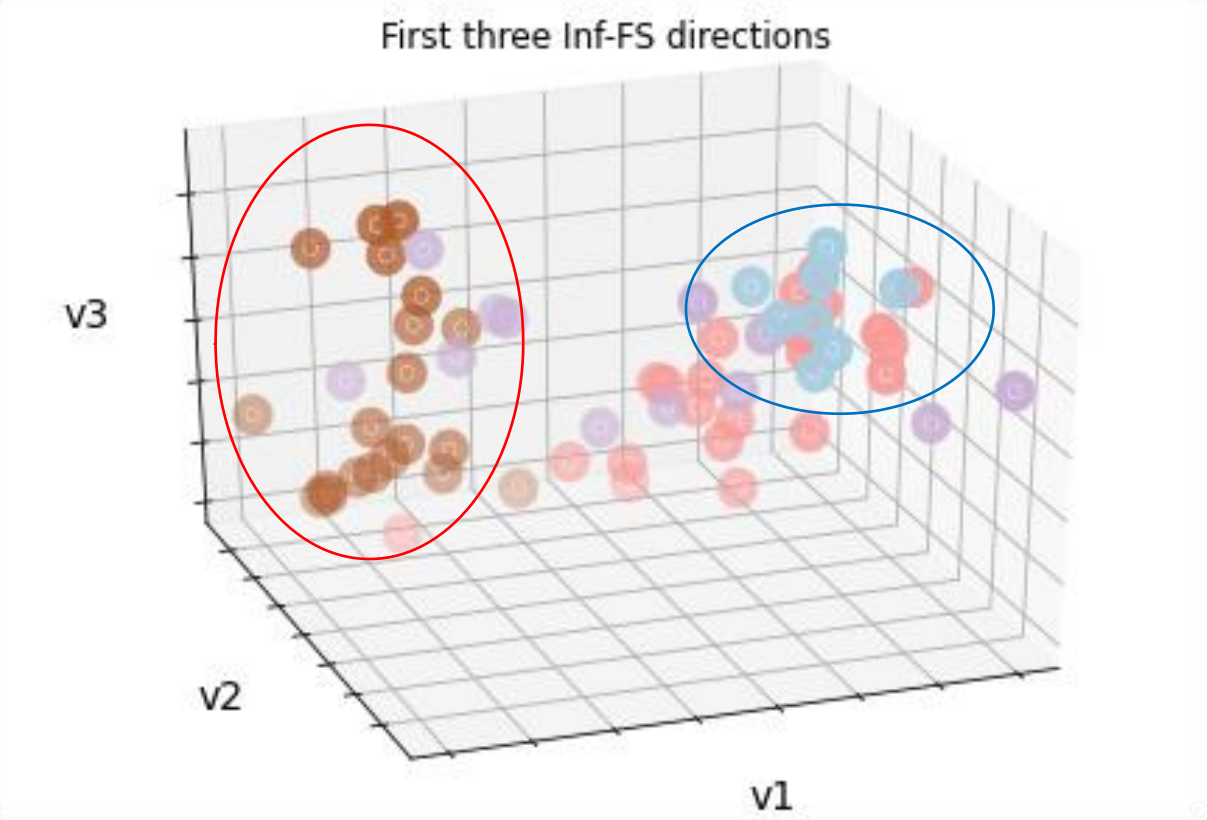
降维后可视化-2D





- X 轴 —— V1
- Y轴 —— V2
- 圆点半径 —— V3
- 圆点颜色 —— 所属类别

降维后可视化-3D



- X 轴 —— V1
- Y轴 —— V2
- Z轴 —— V3
- 圆点颜色 —— 所属类别

MODELS

- 模型选择
- 预测试
- 模型调参/模型训练
- 测试

模型选择

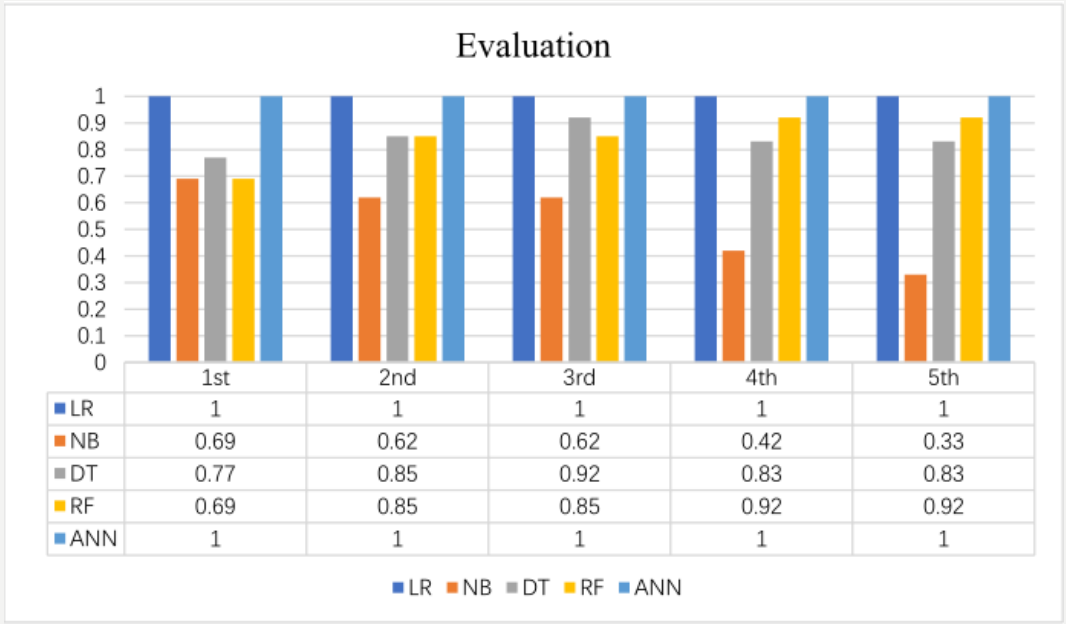
- Logistic regression
- Naïve Bayes
- Decision Trees / Random Forest
- Artificial Neural Network

\+ **Softmax Function**

Natively support

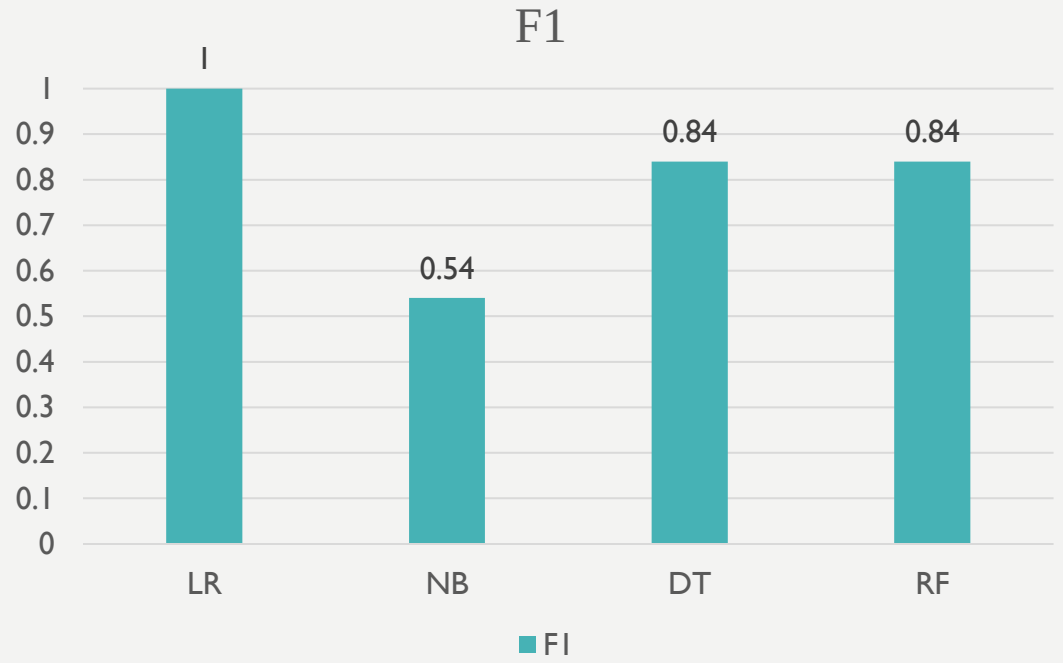
预测试

- Default hyperparameters / settings
- 5-fold cross-validation (scoring=“accuracy”)



	1st	2nd	3rd	4th	5th
LR	1	1	1	1	1
NB	0.69	0.62	0.62	0.42	0.33
DT	0.77	0.85	0.92	0.83	0.83
RF	0.69	0.85	0.85	0.92	0.92
ANN	1	1	1	1	1

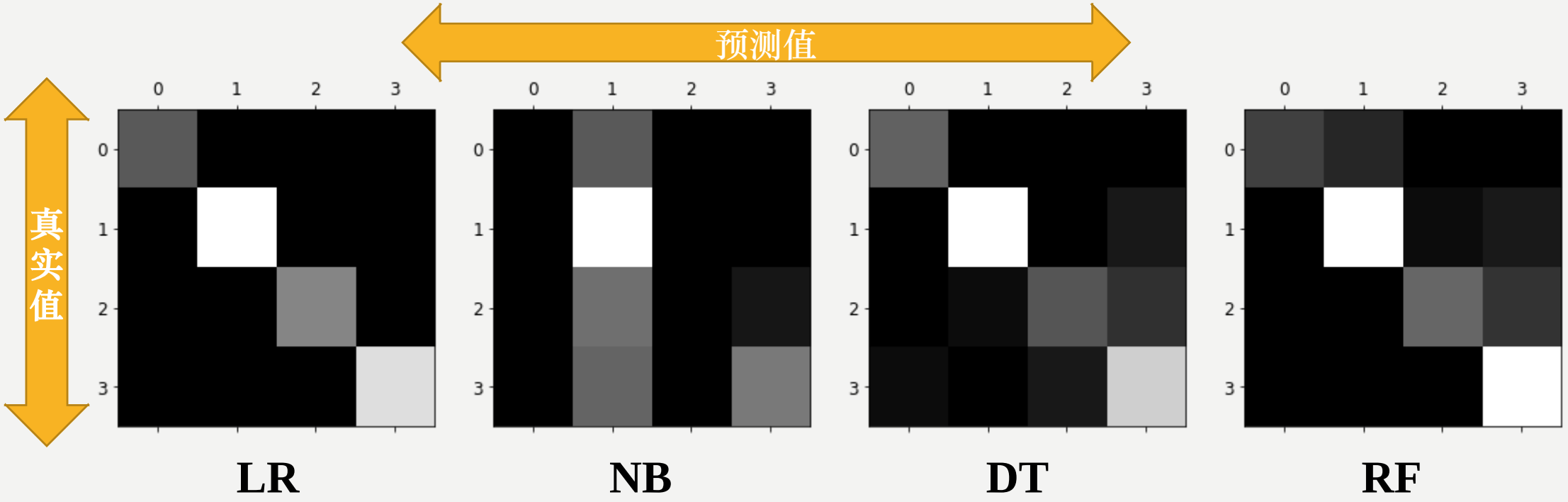
Accuracy



F1 value

预测试

- Confusion matrix



模型调参/模型训练

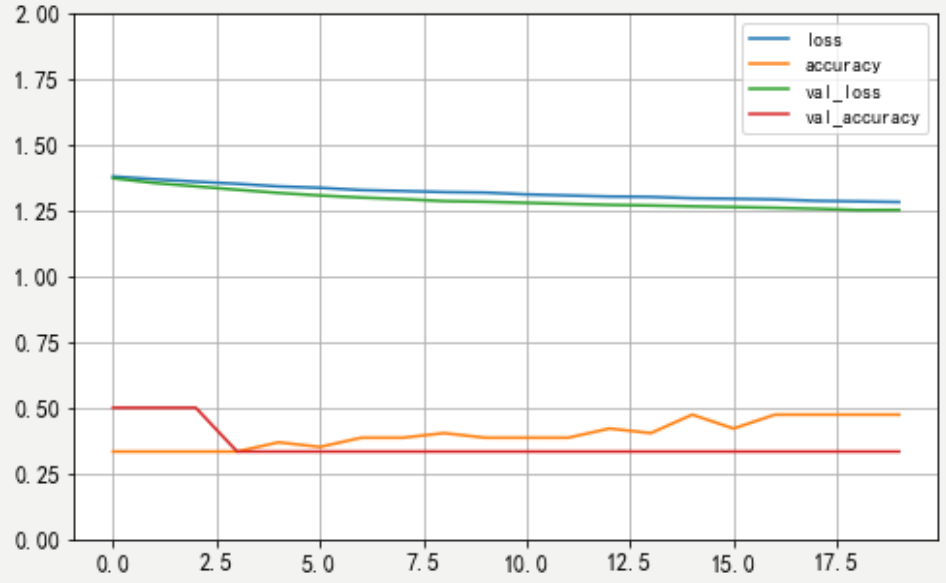
Logistic Regression

Hyperparameters	Values
solver	“lbfgs”,“sag”,“newton-cg”
multi-class	“ovr”,“multinomial”
C	4,6,8,10

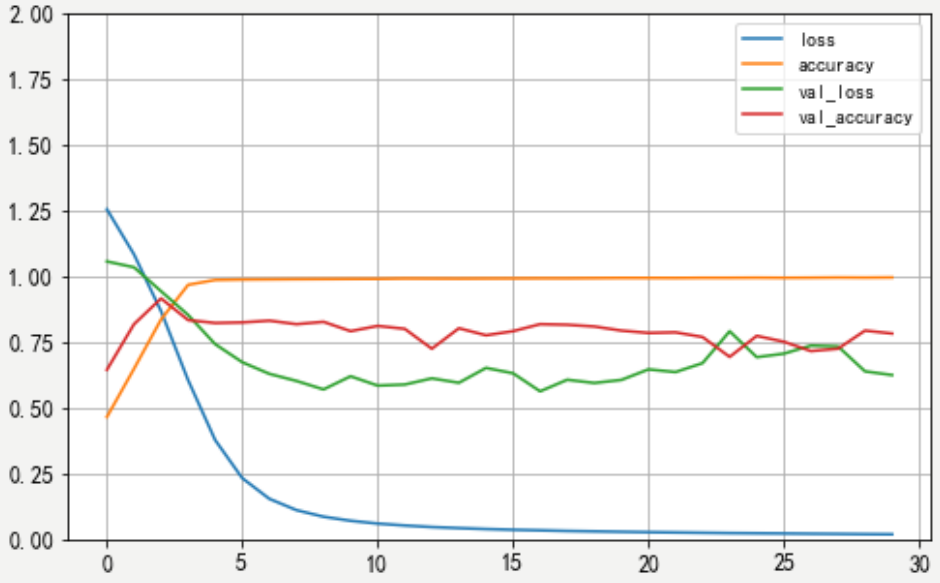
Decision Trees / Random Forest

Hyperparameters	Values
N_estimators	100,150,200,300
max_depth	4,5,6

模型调参/模型训练



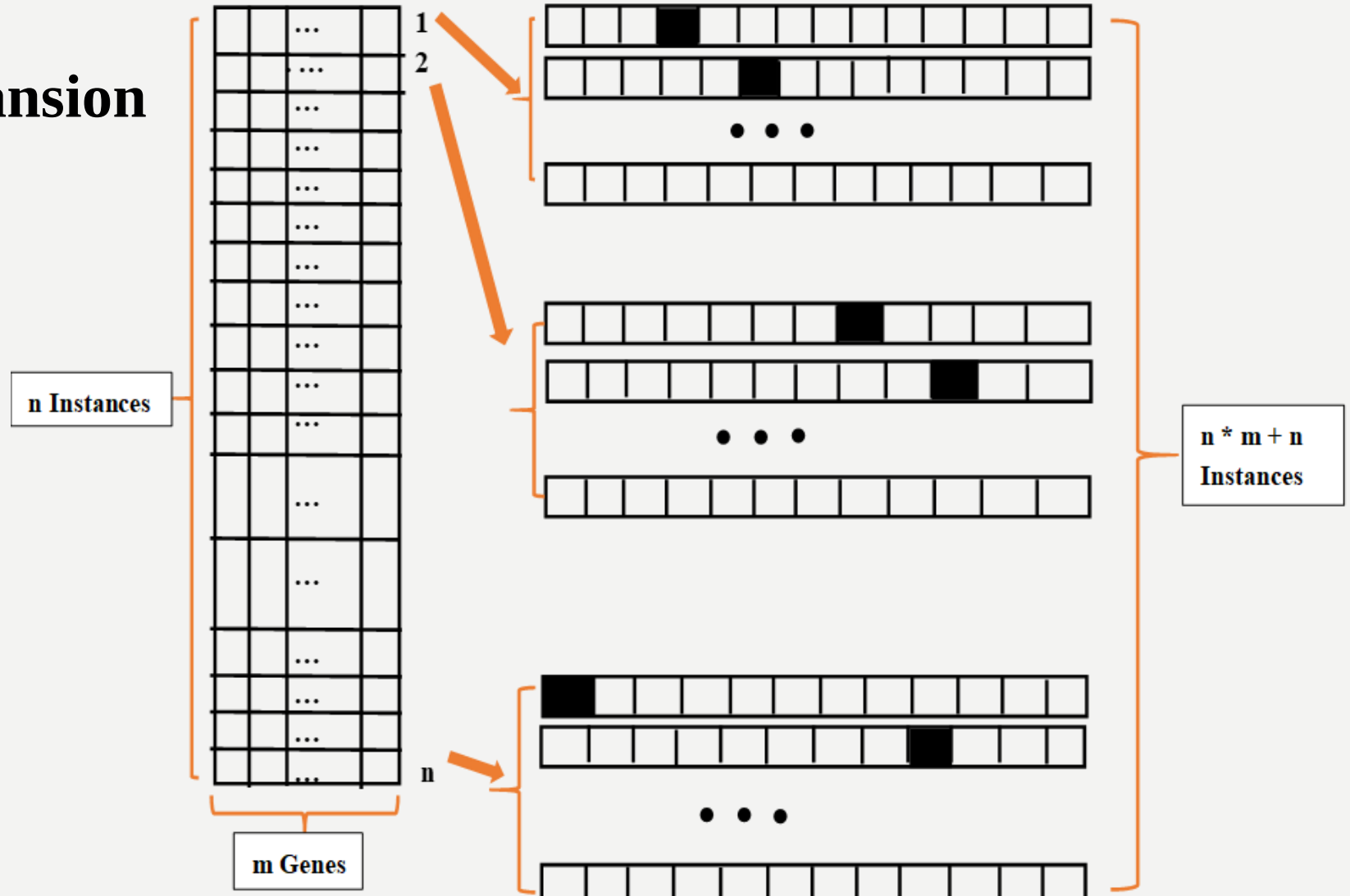
Before samples expansion



After samples expansion

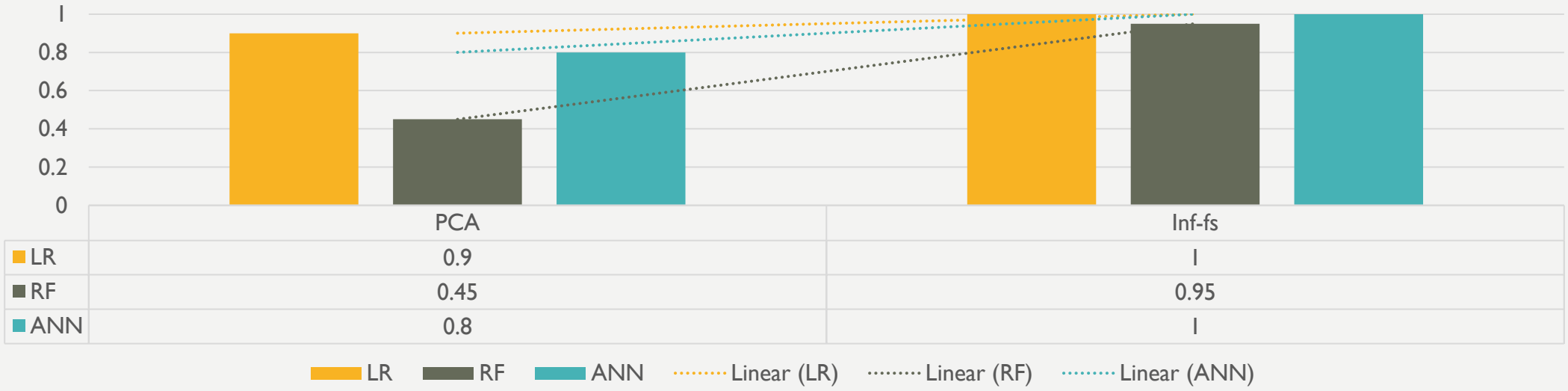
模型调参/模型训练

Samples expansion

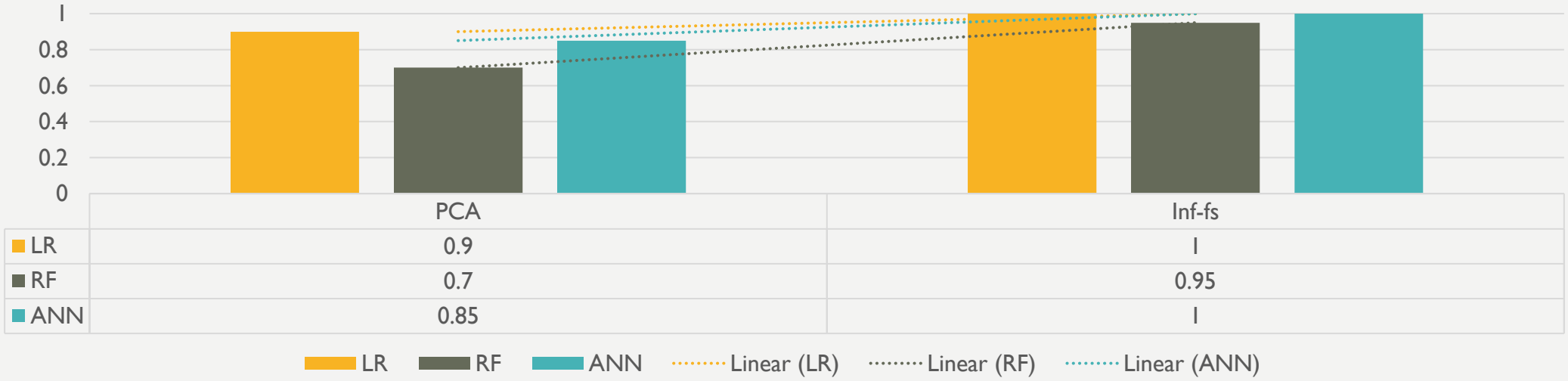


测试

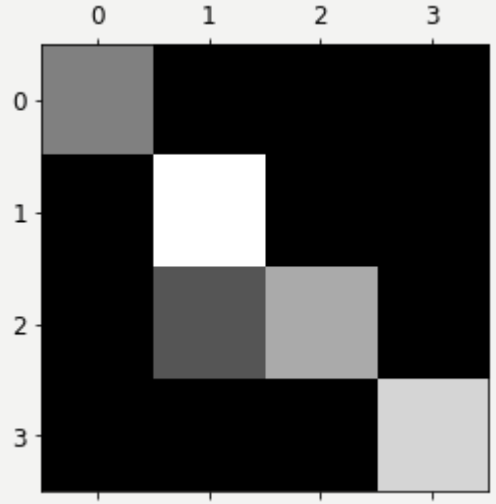
Accuracy



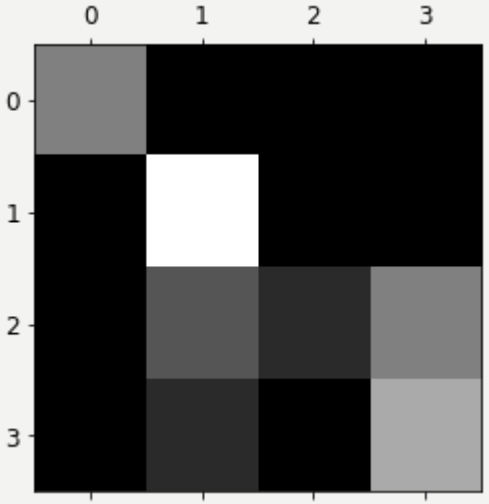
F1 Value



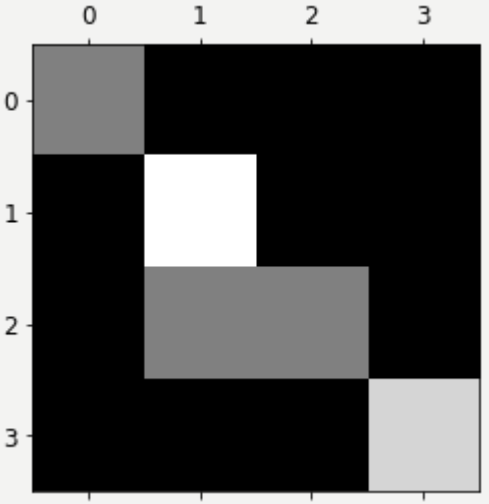
测试



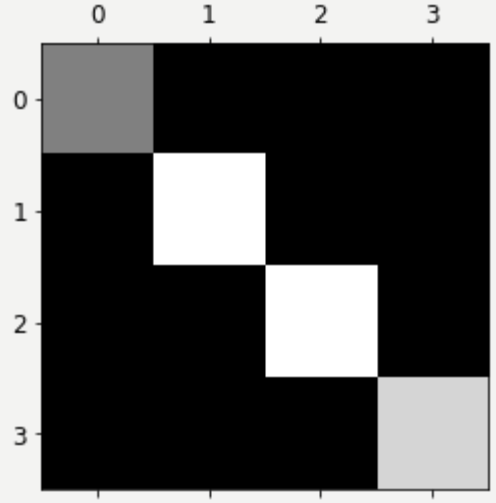
LR_PCA



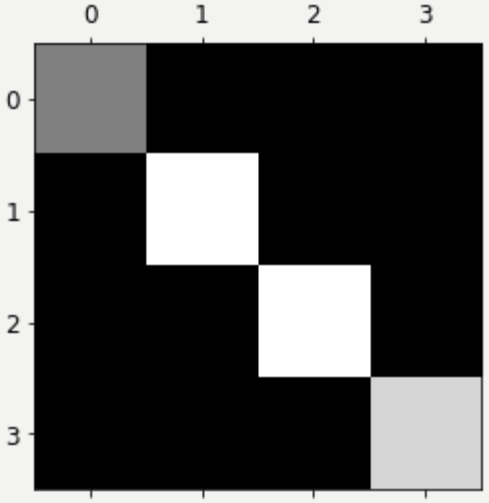
RF_PCA



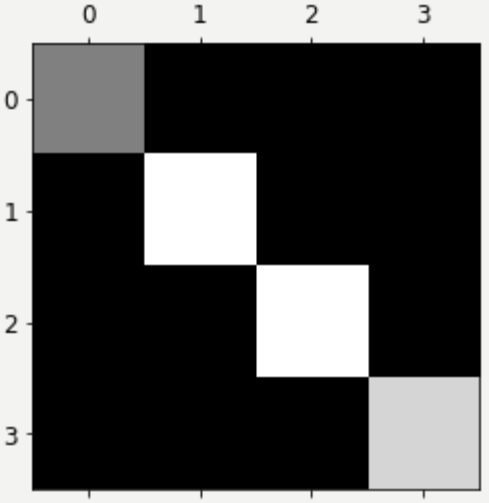
ANN_PCA



LR_Inf



RF_Inf



ANN_Inf

CONCLUSION

- 发现
- 不足

发现

- 高维性
- 样本数目少

Logistic Regression (1.0) = ANN (1.0) > Random Forest (0.95)

- 利用Inf-FS算法进行降维
- 利用SE对降维数据进行样本扩充

不足

- SE1DCNN and SESAE
- 对Inf-FS算法的理解
- 调参
- 高维数据可视化

谢谢观看，有不足，请多多包涵。