deepshare.net
深度之眼

Kaggle-生物分子预测

导师：Karl

1. 验证集划分
2. 模型结构修改
3. 优化器修改
4. SVD 特征降维
5. 其余模型尝试
6. 提交融合

验证集划分

提前划分好
写入硬盘中

- 稀疏矩阵划分
- 稠密矩阵划分

- 随机划分
- 按受试者划分
- 按时间划分

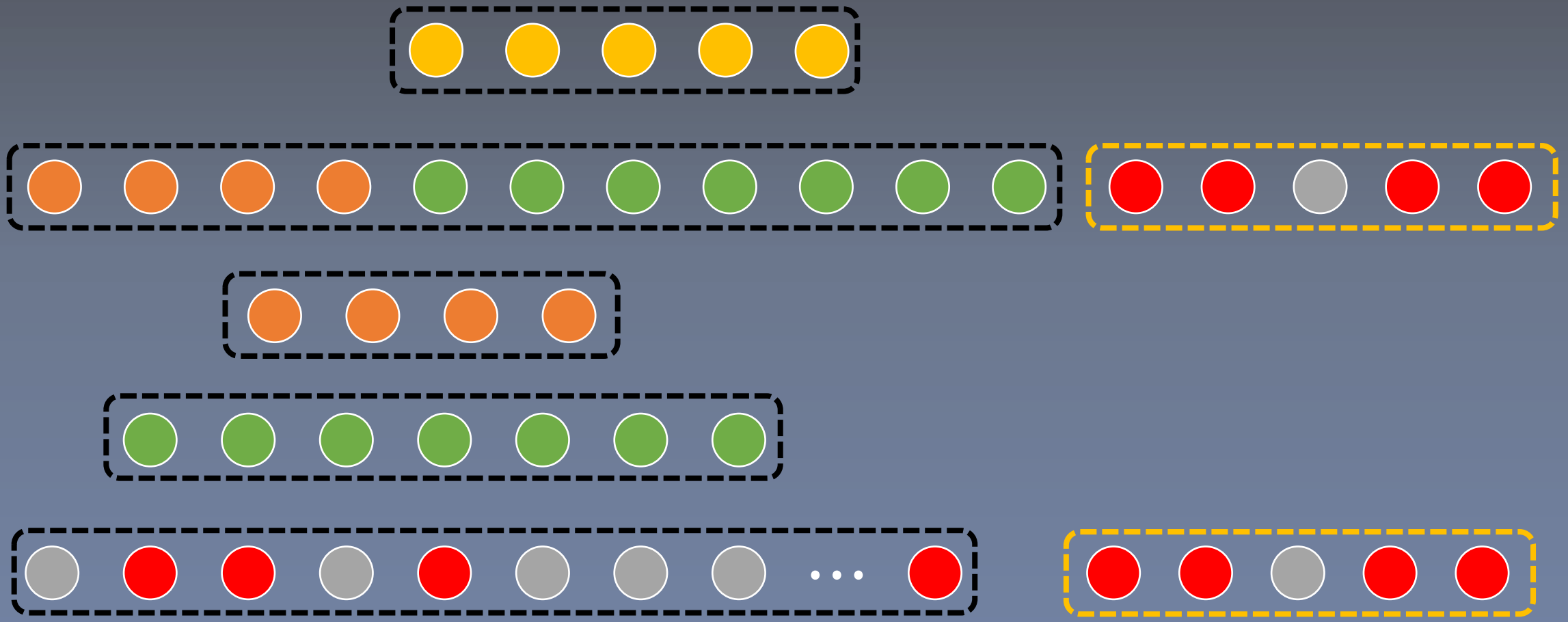
验证集划分目的

1. 方便快速迭代
2. 为后面 K-fold 融合准备

验证集划分原则

1. 和线上得分一致
2. 和线上得分趋势一致

模型结构修改



优化器修改

Sgd→ Adam

MULT : 0.66xx
CITE : 0.89xx

多做实验

SVD特征降维

Mathematically, truncated SVD applied to training samples X produces a low-rank approximation X:

$$X \approx X_k = U_k \Sigma_k V_k^\top$$

After this operation, $U_k \Sigma_k$ is the transformed training set with k features (called `n_components` in the API).

To also transform a test set X, we multiply it with V_k:

$$X' = X V_k$$

线性降维　　近似于PCA

问题：为什么不用PCA？　答：SVD能有效处理稀疏矩阵

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html

其余模型尝试

Lgb 模型 (损失函数 MSE)
(https://www.kaggle.com/code/jsmithperera/citeseq-ambrosm-s-x-keras-lgbm-tabnet)

Cab 模型 (损失函数 MSE)
(https://www.kaggle.com/code/xiafire/lb-t15-msci-multiome-catboostregressor)

Ridge/Lasso 模型 (损失函数 MSE)
(https://www.kaggle.com/code/stautxie/svd-based-decomposition-for-cite-prediction)

TabNet 模型 (损失函数 MSE)
(https://www.kaggle.com/code/tamaryo/lb-0-807-citeseq-tabnet-baseline)(pytorch-tabnet)

PS：除 cab/TabNet 外，需要结合 MultiOutputRegressor

提交融合

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X * \sigma_Y} = \frac{\mathbb{E}[(X - \mathbb{E}(X)) * (Y - \mathbb{E}(Y))]}{\sigma_X * \sigma_Y}$$

Now take another solution $X' = C_1 + C_2 * X$

$$corr(X',Y) = corr(C_1 + C_2 * X, Y) = \frac{\mathbb{E}[(C_1 + C_2 * X - \mathbb{E}[C_1 + C_2 * X]) * (Y - \mathbb{E}(Y))]}{\sigma_{[C_1 + C_2 * X]} * \sigma_Y}$$

$$= \frac{\mathbb{E}[(C_1 + C_2 * X - C_1 + C_2 * \mathbb{E}(X)) * (Y - \mathbb{E}(Y))]}{C_2 * \sigma_X * \sigma_Y}$$

$$= \frac{C_2 * \mathbb{E}[(X - \mathbb{E}(X)) * (Y - \mathbb{E}(Y))]}{C_2 * \sigma_X * \sigma_Y}$$

$$= \frac{\mathbb{E}[(X - \mathbb{E}(X)) * (Y - \mathbb{E}(Y))]}{\sigma_X * \sigma_Y}$$

$$= corr(X,Y)$$

https://www.kaggle.com/code/vslaykovsky/lb-0-811-normalized-ensembles-for-pearson-s-r

提交融合

线性变换不影响最终的评分
融合前（不同模型量纲可能不一致）做标准化
然后赋予权重融合在一起
权重既可以是等权也可以适当调整
例如根据提交的评分调整权重

多做实验

PS：实际上是在拟合PB

Q&A