

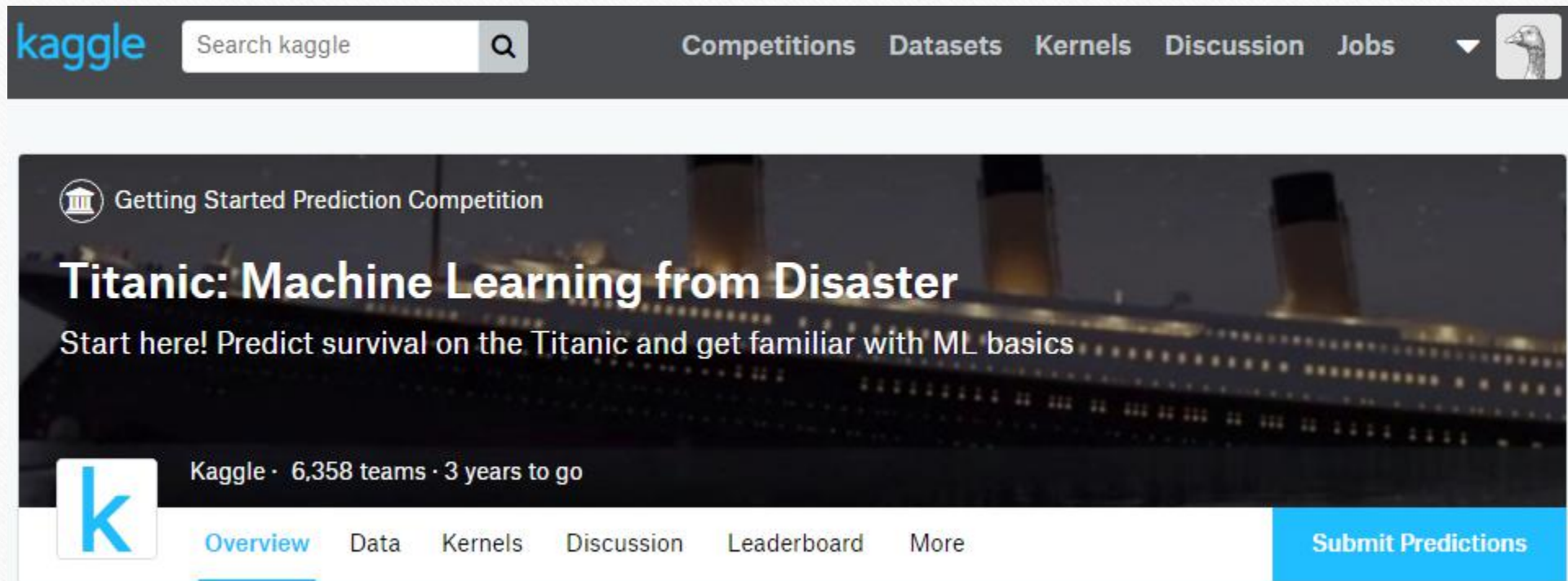
本周工作

学生：屠晓涵

本周工作

- 机器人通信模块数据压缩编码、 解码
- 机器人数据产生模拟器制作
- 机器人数据存储模拟器制作
- 智能检测机器人——工业大数据
- 大论文修改

kaggle机器学习竞赛



The image shows a screenshot of the Kaggle website. At the top, there is a navigation bar with the Kaggle logo, a search bar, and links for Competitions, Datasets, Kernels, Discussion, and Jobs. Below the navigation bar, there is a large banner for the 'Titanic: Machine Learning from Disaster' competition. The banner features a background image of the Titanic ship at night. The text on the banner includes 'Getting Started Prediction Competition', 'Titanic: Machine Learning from Disaster', and 'Start here! Predict survival on the Titanic and get familiar with ML basics'. Below the banner, there is a section with the Kaggle logo, the text 'Kaggle · 6,358 teams · 3 years to go', and a navigation menu with links for Overview, Data, Kernels, Discussion, Leaderboard, and More. A blue button labeled 'Submit Predictions' is also visible.

网址: <https://www.kaggle.com/c/titanic#evaluation>

Titanic: Machine Learning from Disaster数据集

乘客编号 是否幸存 船舱等级 姓名 性别 年龄 兄弟姊妹人数 父母人数 船票信息 票价 船舱型号 上船地址

	A	B	C	D	E	F	G	H	I	J	K	L
1	PassengerId	Survived	Class	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
2	1	0	3	Braund, M	male	22	1	0	A/5 21171	7.25		S
3	2	1	1	Cumings, female	female	38	1	0	PC 17599	71.2833	C85	C
4	3	1	3	Heikkinen, female	female	26	0	0	STON/O2.	7.925		S
5	4	1	1	Futrelle, female	female	35	1	0	113803	53.1	C123	S
6	5	0	3	Allen, M	male	35	0	0	373450	8.05		S
7	6	0	3	Moran, M	male		0	0	330877	8.4583		Q
8	7	0	1	McCarthy, male	male	54	0	0	17463	51.8625	E46	S
9	8	0	3	Palsson, male	male	2	3	1	349909	21.075		S
10	9	1	3	Johnson, female	female	27	0	2	347742	11.1333		S
11	10	1	2	Nasser, M	female	14	1	0	237736	30.0708		C
12	11	1	3	Sandstrom, female	female	4	1	1	PP 9549	16.7	G6	S
13	12	1	1	Bonnell, female	female	58	0	0	113783	26.55	C103	S
14	13	0	3	Saunders, male	male	20	0	0	A/5. 2151	8.05		S
15	14	0	3	Andersson, male	male	39	1	5	347082	31.275		S
16	15	0	3	Vestrom, female	female	14	0	0	350406	7.8542		S
17	16	1	2	Hewlett, female	female	55	0	0	248706	16		S
18	17	0	3	Rice, M	male	2	4	1	382652	29.125		Q
19	18	1	2	Williams, male	male		0	0	244873	13		S

任务和思路

- 任务：根据船员信息预测是否被获救
- 思路：
 - 理解问题和数据信息，针对问题的特性来选择算法
 - 对比算法，选择更好的机器学习算法
 - 生成更好的特征，用更好的特征来预测
 - 根据模型或者算法本身存在的不合理的地方，然后提出新的假设，从而去优化模型或算法。

数据预处理

- 将年龄空缺值用整体年龄的平均值补全
- 将性别空缺值用统计出的人数最多的性别填充
- 性别字符串转换为整型0、1,
- 将上船地址空缺值用登陆最多的地点'S' (统计得出S) 填充
- 将上船地址字符串转换为整型0、1、2……

线性回归算法做预测

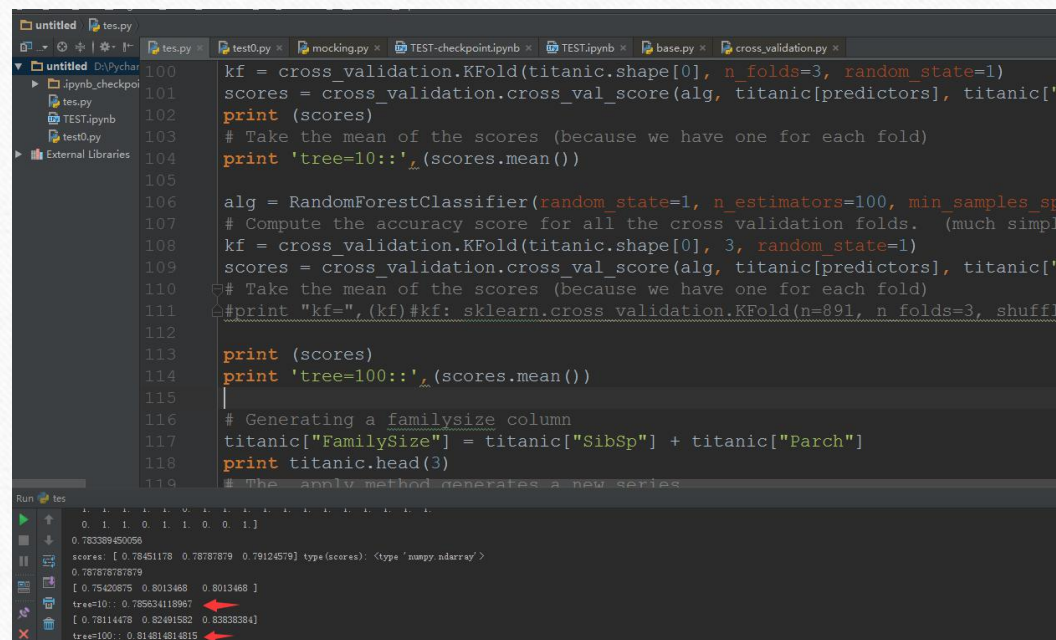
- 选取特征["Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked"]
- 采用k-折交叉验证, k=3
- 实验指标: 准确率
- 结果: 0.783389450056
- 注释1: k-折交叉检验就是把原始的数据随机分成K个部分。在这K个部分中, 选择一个作为测试数据, 剩下的K-1个作为训练数据。
- 注释2: 实际上是把实验重复做K次, 每次实验都从K个部分中选择一个不同的部分作为测试数据(保证K个部分的数据都分别做过测试数据), 剩下的K-1个当作训练数据进行实验, 最后把得到的K个实验结果平均。

逻辑回归算法做预测

- 选取特征["Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Embarked"]
- 采用k-折交叉验证, k=3
- 实验指标: 准确率
- 结果: 0.787878787879 大于 线性回归结果: 0.783389450056

随机森林算法做预测

- 在我们的数据上构建一个随机森林并且生成交叉验证预测
- 实验指标：准确率
- 10颗树时结果：0.785634118967
- 100颗树时结果：0.814814814815



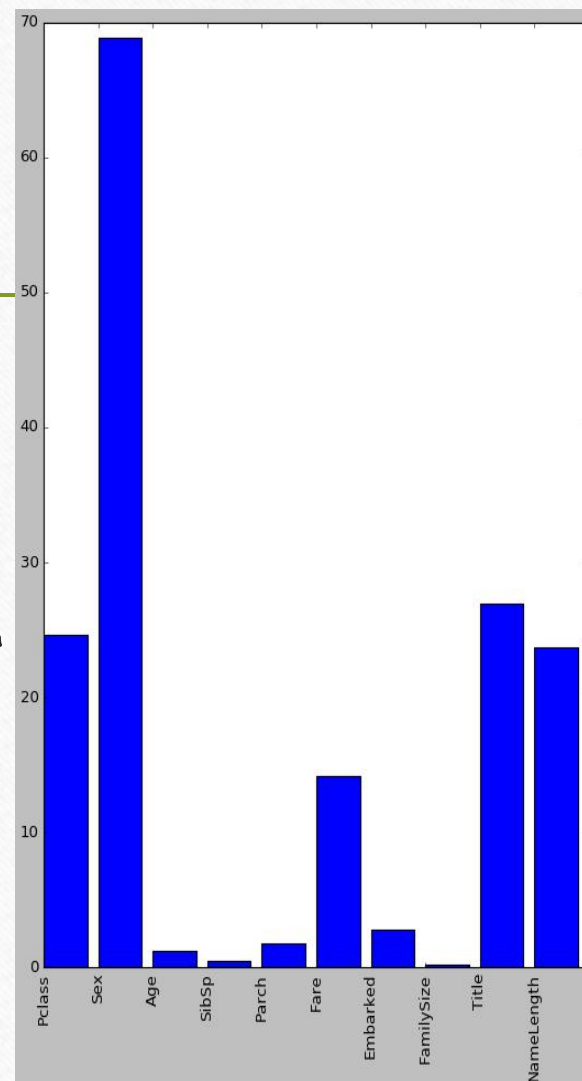
```
100 kf = cross_validation.KFold(titanic.shape[0], n_folds=3, random_state=1)
101 scores = cross_validation.cross_val_score(alg, titanic[predictors], titanic['
102 print (scores)
103 # Take the mean of the scores (because we have one for each fold)
104 print 'tree=10::', (scores.mean())
105
106 alg = RandomForestClassifier(random_state=1, n_estimators=100, min_samples_sp
107 # Compute the accuracy score for all the cross validation folds. (much simpl
108 kf = cross_validation.KFold(titanic.shape[0], 3, random_state=1)
109 scores = cross_validation.cross_val_score(alg, titanic[predictors], titanic['
110 # Take the mean of the scores (because we have one for each fold)
111 print "kf=", (kf) #kf: sklearn.cross_validation.KFold(n=891, n_folds=3, shuffl
112
113 print (scores)
114 print 'tree=100::', (scores.mean())
115
116 # Generating a familysize column
117 titanic["FamilySize"] = titanic["SibSp"] + titanic["Parch"]
118 print titanic.head(3)
119 # The apply method generates a new series
```

Run tes

```
0.783389450056
scores: [ 0.78451178  0.78787879  0.79124579] type(scores): <type 'numpy.ndarray'>
0.787878787879
[ 0.78420675  0.8013468  0.8013468 ]
tree=10:: 0.785634118967
[ 0.78114478  0.82491582  0.83838384 ]
tree=100:: 0.814814814815
```

随机森林算法改进策略

- 1, 生成新特征, 使用新特征提高预测准确率, 新特征如下:
 - 名字长度 (名字长度和富穷有关系, 可能决定获救的几率)
 - 一个家庭总人数 (一个人在船上的家庭人数, 可能决定获救的几率)
 - 头衔特征 (头衔的格式是Master., Mr., Mrs.)
 - 通过家庭总人数连接某些人的姓来得到一个家庭编号。然后基于他们的家庭编号给每个人赋值一个特征 (可能决定获救的几率)
- 2, 找出最好的特征: Pclass, Sex, Fare, Title (如图)



随机森林算法改进策略

- 3, 重新选择树的数量, 树的深度, 观察是否会提升预测准确
- 4, 集成一个线性回归和一个随机森林, 观察是否会提升预测准确
- 5, 集成一个逻辑回归和一个随机森林, 观察是否会提升预测准确
- 6, 集成线性回归和随机森林结果: 0.81593714927
- 7, 集成逻辑回归和随机森林结果: 0.817059483726,

下步工作

- 特征工程
 - 观察家庭大小特征是否会有帮助——一个家庭中女性的数量多使全家更可能幸存？
 - 尝试用和船舱相关的特征。
- 算法方面
 - 支持向量机是否会很有效
 - 神经网络是否会更有效
- 集成方法
 - 多数表决是否是比概率平均更好的集成方法？

谢谢观看指导！