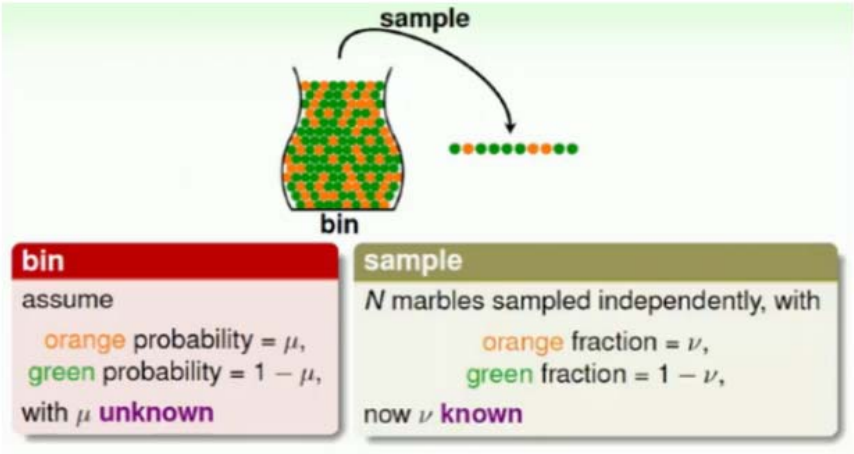
zfmlhw010

1. 阅读作业

阅读博客文章 《 NFL原理与Hoeffding不等式》（http://blog.csdn.net/baimafujinji/article/details/6475824）(http://blog.csdn.net/baimafujinji/article/details/6475824）) 中关于Hoeffding不等式的部分。
注意第2题也涉及到Hoeffding不等式，为了更好地完成后续题目，请务必认真阅读。

二、Hoeffding不等式

为了引出Hoeffding不等式的意义，先来看一个例子。如下图所示，我们有一个罐子，其中装有绿色和橘色两种颜色的小球。整个罐子里，橘色小球所占的比例 u 是未知的，为了推测这个未知的 u，可以从罐子里面随机的抽取一组样本，在被抽到的若干小球里，可以得知橘色小球所占的比例 v。显然，v和u应该是存在某种关系的，这个关系就是Hoeffding不等式。



2. 证明题

假设抛硬币正面朝上的概率为 p, 反面朝上的概率为 $1-p$. 令 $H(n)$ 代表抛 n 次硬币所得正面朝上的次数, 则最多 k 次正面朝上的概率为

$$P(H(n) \leqslant k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} .$$

对 $\delta > 0$, $k = (p-\delta)n$, 有 Hoeffding 不等式

$$P(H(n) \leqslant (p-\delta)n) \leqslant e^{-2\delta^2 n} .$$

据此证明PPT中第8页的公式：

$$P\left(H\left(\boldsymbol{x}\right) \neq f(\boldsymbol{x})\right) = \sum_{k=0}^{\lfloor T/2 \rfloor} \binom{T}{k} (1-\epsilon)^k \epsilon^{T-k}$$

$$\leqslant \exp\left(-\frac{1}{2} T \left(1-2\epsilon\right)^2\right) .$$

由已知得

$$P(H(n) \le (p-\delta)n) = \sum_{i=0}^{(p-\delta)n} \binom{n}{i} p^i (1-p)^{n-i} \le \exp(-2\delta^2 n)$$

令 $p-\delta = \frac{1}{2}$，则 $\delta = p - \frac{1}{2}$

令 $n=T$，$p=1-\varepsilon$

$$\text{上式} = \sum_{k=0}^{\frac{1}{2}T} \binom{T}{k} (1-\varepsilon)^k \varepsilon^{T-k} \le \exp\left(-2T\left(\frac{1}{2}-\varepsilon\right)^2\right)$$

$$= \exp\left(-\frac{1}{2}T(1-2\varepsilon)^2\right)$$

3. 实践题

利用集成学习方法实战一下kaggle中的Titanic项目（https://www.kaggle.com/c/titanic）。要求：(https://www.kaggle.com/c/titanic）。要求：)

1）你必须使用Adaboost模型；
2）使用Python或R；
3）至少引入两种原数据集中未提供的特征（例如Title等，可参考授课内容中相关部分）；
4）用文字回答的方式描述你使用了哪些特征；
5）提交完整的（包括引用必要头文件所需的代码）可以执行的代码，代码部分请不要以截图方式提交，因为无法复制粘贴而不能运行的代码将无法判定正确与否。；
6）截取一张你将预测结果提交到kaggle网站后，系统反馈给你的得分的截图（如果你没有用过kaggle，那么你需要先注册一个账号）。

预测结果提交到kaggle网站后，系统反馈的得分的截图

Overview Data Kernels Discussion Leaderboard Rules Team My Submissions **Submit Predictions**

8067	new	**syedhammad**	0.76555	1	2h
8068	new	**umair2536**	0.76555	1	2h
8069	new	**Malik Ghyaoor Abbas**	0.76555	3	10m
8070	new	**mayi14**	0.76555	1	now

Your Best Entry ↑

Your submission scored 0.76555, which is not an improvement of your best score. Keep trying!

8071	▼782	**Rewant Kedia**	0.76076	2	2mo
8072	▼782	**Sakib Reza**	0.76076	1	2mo
8073	▼782	**Euphor**	0.76076	3	2mo

In [65]:

```
import pandas as pd
from sklearn import ensemble
```

1）读入数据

In [41]:

```
train = pd.read_csv('titanic/train.csv', index_col=0)
train.head()
```

Out[41]:

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Ca
1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	I
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	
3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	I
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C
5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	I

In [42]:

```
test = pd.read_csv('titanic/test.csv', index_col=0)
test.head()
```

Out[42]:

PassengerId	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarl
892	3	Kelly, Mr. James	male	34.5	0	0	330911	7.8292	NaN	
893	3	Wilkes, Mrs. James (Ellen Needs)	female	47.0	1	0	363272	7.0000	NaN	
894	2	Myles, Mr. Thomas Francis	male	62.0	0	0	240276	9.6875	NaN	
895	3	Wirz, Mr. Albert	male	27.0	0	0	315154	8.6625	NaN	
896	3	Hirvonen, Mrs. Alexander (Helga E Lindqvist)	female	22.0	1	1	3101298	12.2875	NaN	

In [43]:

```
train.shape, test.shape
```

Out[43]:

```
((891, 11), (418, 10))
```

2）处理缺失值 & 增加特征

In [44]:

```
#查看那些列存在缺失值
train.isnull().any()
```

Out[44]:

```
Survived    False
Pclass      False
Name        False
Sex         False
Age          True
SibSp       False
Parch       False
Ticket      False
Fare        False
Cabin        True
Embarked     True
dtype: bool
```

In [45]:

```
age = train['Age']

#查看Age字段缺失值的数量
age[age.isnull()].shape
```

Out[45]:

```
(177,)
```

In [46]:

```
# 由于缺失字段较多，增加一个特征来记录缺失情况，缺失为1，未缺失为0
train['age_isnull'] = 0
train['age_isnull'][age.isnull()] = 1

test['age_isnull'] = 0
test['age_isnull'][test['Age'].isnull()] = 1
```

```
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  This is separate from the ipykernel package so we can avoid doing i
mports until
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:6: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
```

In [47]:

```
#用train['Age']的平均值填充缺失
age.mean()
```

Out[47]:

```
29.69911764705882
```

In [48]:

```
train['Age'] = train['Age'].fillna(age.mean())
test['Age'] = test['Age'].fillna(age.mean())
```

In [49]:

```
cabin = train['Cabin']

#查看cabin字段缺失值的数量
cabin[cabin.isnull()].shape
```

Out[49]:

```
(687,)
```

In [50]:

```
# 由于缺失字段较多，增加一个特征来记录缺失情况，缺失为1，未缺失为0
train['cabin_isnull'] = 0
train['cabin_isnull'][cabin.isnull()] = 1
test['cabin_isnull'] = 0
test['cabin_isnull'][test['Cabin'].isnull()] = 1
```

```
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  This is separate from the ipykernel package so we can avoid doing i
mports until
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:5: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  """
```

可以看出cabin的第一个字母代表cabin的等级，所以取出第一个字母作为特征

In [52]:

```
cabin[cabin.isnull()==False].apply(lambda x:x[0][0]).value_counts()
```

Out[52]:

```
C    59
B    47
D    33
E    32
A    15
F    13
G     4
T     1
Name: Cabin, dtype: int64
```

可以看到C最多，但是C,B,D,E的差别并不算很大，所以考虑以一个新值N来填充缺失值

In [55]:

```
train['Cabin'] = train['Cabin'].fillna('N')

train['Cabin'] = train['Cabin'].apply(lambda x:x[0][0])
```

In [58]:

```
test['Cabin'] = test['Cabin'].fillna('N')

test['Cabin'] = test['Cabin'].apply(lambda x:x[0][0])
test['Cabin'].value_counts()
```

Out[58]:

```
N    327
C     35
B     18
D     13
E      9
F      8
A      7
G      1
Name: Cabin, dtype: int64
```

In [59]:

```
train['Embarked'].value_counts()
```

Out[59]:

```
S    644
C    168
Q     77
Name: Embarked, dtype: int64
```

In [60]:

```
train['Embarked'][train['Embarked'].isnull()].shape
```

Out[60]:

```
(2,)
```

由于Embark的缺失值很少，所以用S填充缺失值

In [61]:

```
train['Embarked'] = train['Embarked'].fillna('S')
```

In [62]:

```
test['Embarked'] = test['Embarked'].fillna('S')
```

增加名称的称谓特征

In [77]:

```
train['title'] = train['Name'].apply(lambda x:x[(x.index(',')+2):x.index('.')].stri
train['title'].value_counts()
```

Out[77]:

```
Mr              517
Miss            182
Mrs             125
Master           40
Dr                7
Rev               6
Col               2
Mlle              2
Major             2
Ms                1
Capt              1
Lady              1
the Countess      1
Sir               1
Jonkheer          1
Don               1
Mme               1
Name: title, dtype: int64
```

做一些同义替换

In [80]:

```
train['title'][train['title']=='Ms'] = 'Mrs'
train['title'][train['title']=='Lady'] = 'Mrs'
train['title'][train['title']=='Sir'] = 'Mr'
train['title'].value_counts()
```

```
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  """Entry point for launching an IPython kernel.
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)

/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  This is separate from the ipykernel package so we can avoid doing i
mports until
```

Out[80]:

```
Mr              518
Miss            182
Mrs             127
Master           40
Dr                7
Rev               6
Major             2
Col               2
Mlle              2
Jonkheer          1
Mme               1
Capt              1
the Countess      1
Don               1
Name: title, dtype: int64
```

In [81]:

```
test['title'] = test['Name'].apply(lambda x:x[(x.index(',')+2):x.index('.')].strip(
test['title'][test['title']=='Ms'] = 'Mrs'
test['title'][test['title']=='Lady'] = 'Mrs'
test['title'][test['title']=='Sir'] = 'Mr'
test['title'].value_counts()
```

```
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)

/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:3: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  This is separate from the ipykernel package so we can avoid doing i
mports until
/home/ian/installed/anaconda3/lib/python3.6/site-packages/ipykernel_l
auncher.py:4: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation: http://pandas.pydata.org/pandas
-docs/stable/indexing.html#indexing-view-versus-copy (http://pandas.p
ydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy)
  after removing the cwd from sys.path.
```

Out[81]:

```
Mr        240
Miss       78
Mrs        73
Master     21
Rev         2
Col         2
Dona        1
Dr          1
Name: title, dtype: int64
```

到此为止，引入了4个特征，age_isnull用来标记age是否缺失，cabin_isnull用来标记cabin是否缺失，同时选cabin的首字母作为一个新的特征，用title来标记称谓

删除 Name 和 Ticket

In [82]:

```
del train['Name']
del test['Name']
del train['Ticket']
del test['Ticket']
```

In [83]:

```
train.head()
```

Out[83]:

PassengerId	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked	age_i
1	0	3	male	22.0	1	0	7.2500	N	S	
2	1	1	female	38.0	1	0	71.2833	C	C	
3	1	3	female	26.0	0	0	7.9250	N	S	
4	1	1	female	35.0	1	0	53.1000	C	S	
5	0	3	male	35.0	0	0	8.0500	N	S	

In [84]:

```
test.head()
```

Out[84]:

PassengerId	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked	age_isnull	cabi
892	3	male	34.5	0	0	7.8292	N	Q	0	
893	3	female	47.0	1	0	7.0000	N	S	0	
894	2	male	62.0	0	0	9.6875	N	Q	0	
895	3	male	27.0	0	0	8.6625	N	S	0	
896	3	female	22.0	1	1	12.2875	N	S	0	

2）把分类变量用数字替代

In [88]:

```
Sex_mapping = dict(zip(train['Sex'].unique(),range(len(train['Sex'].unique()))))
Cabin_mapping = dict(zip(train['Cabin'].unique(),range(len(train['Cabin'].unique()))
Embarked_mapping = dict(zip(train['Embarked'].unique(),range(len(train['Embarked'].
title_mapping = dict(zip(train['title'].unique(),range(len(train['title'].unique()))
```

In [89]:

```
train['Sex'] = train['Sex'].map(Sex_mapping)
train['Cabin'] = train['Cabin'].map(Cabin_mapping)
train['Embarked'] = train['Embarked'].map(Embarked_mapping)
train['title'] = train['title'].map(title_mapping)
```

In [91]:

```
test['Sex'] = test['Sex'].map(Sex_mapping)
test['Cabin'] = test['Cabin'].map(Cabin_mapping)
test['Embarked'] = test['Embarked'].map(Embarked_mapping)
test['title'] = test['title'].map(title_mapping)
```

In [93]:

```
train.head()
```

Out[93]:

	Survived	Pclass	Sex	Age	SibSp	Parch	Fare	Cabin	Embarked	age_isn
PassengerId										
1	0	3	0	22.0	1	0	7.2500	0	0	
2	1	1	1	38.0	1	0	71.2833	1	1	
3	1	3	1	26.0	0	0	7.9250	0	0	
4	1	1	1	35.0	1	0	53.1000	1	0	
5	0	3	0	35.0	0	0	8.0500	0	0	

3）训练模型

In [95]:

```
clf = ensemble.AdaBoostClassifier()
```

In [96]:

```
clf.fit(train.values[:,1:],train.values[:,0])
```

Out[96]:

```
AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None,
          learning_rate=1.0, n_estimators=50, random_state=None)
```

In [97]:

```
clf.predict(test.values)
```

```
---------------------------------------------------------------------------
ValueError                                Traceback (most recent call last)
<ipython-input-97-b3e107bc1ead> in <module>()
----> 1 clf.predict(test.values)

~/installed/anaconda3/lib/python3.6/site-packages/sklearn/ensemble/weight_boosting.py in predict(self, X)
    600             The predicted classes.
    601         """
--> 602         pred = self.decision_function(X)
    603
    604         if self.n_classes_ == 2:

~/installed/anaconda3/lib/python3.6/site-packages/sklearn/ensemble/weight_boosting.py in decision_function(self, X)
    659         """
    660         check_is_fitted(self, "n_classes_")
--> 661         X = self._validate_X_predict(X)
    662
    663         n_classes = self.n_classes_

~/installed/anaconda3/lib/python3.6/site-packages/sklearn/ensemble/weight_boosting.py in _validate_X_predict(self, X)
    267                 isinstance(self.base_estimator,
    268                            (BaseDecisionTree, BaseForest))):
--> 269             X = check_array(X, accept_sparse='csr', dtype=DTYPE)
    270
    271         else:

~/installed/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py in check_array(array, accept_sparse, dtype, order, copy, force_all_finite, ensure_2d, allow_nd, ensure_min_samples, ensure_min_features, warn_on_dtype, estimator)
    451                               % (array.ndim, estimator_name))
    452         if force_all_finite:
--> 453             _assert_all_finite(array)
    454
    455     shape_repr = _shape_repr(array.shape)

~/installed/anaconda3/lib/python3.6/site-packages/sklearn/utils/validation.py in _assert_all_finite(X)
     42             and not np.isfinite(X).all()):
     43         raise ValueError("Input contains NaN, infinity"
---> 44                          " or a value too large for %r." % X.dtype)
     45
     46

ValueError: Input contains NaN, infinity or a value too large for dtype('float32').
```

预测是报错，提示test中有空值。处理test中的空值

In [98]:

```
test.isnull().any()
```

Out[98]:

```
Pclass           False
Sex              False
Age              False
SibSp            False
Parch            False
Fare              True
Cabin            False
Embarked         False
age_isnull       False
cabin_isnull     False
title             True
dtype: bool
```

In [99]:

```
title_mapping
```

Out[99]:

```
{'Mr': 0,
 'Mrs': 1,
 'Miss': 2,
 'Master': 3,
 'Don': 4,
 'Rev': 5,
 'Dr': 6,
 'Mme': 7,
 'Major': 8,
 'Mlle': 9,
 'Col': 10,
 'Capt': 11,
 'the Countess': 12,
 'Jonkheer': 13}
```

In [100]:

```
test['Fare'] = test['Fare'].fillna(train['Fare'].mean())
test['title'] = test['title'].fillna(0)
```

In [103]:

```
result = pd.DataFrame()
```

In [104]:

```
result['PassengerId'] = test.index
```

In [113]:

```
result['Survived'] = clf.predict(test.values)
```

In [106]:

```
result.head()
```

Out[106]:

	PassengerId	Survived
0	892	0.0
1	893	0.0
2	894	0.0
3	895	0.0
4	896	1.0

In [114]:

```
result.to_csv('result.csv')
```

end

end

end

end

end

end

end

end

end

end