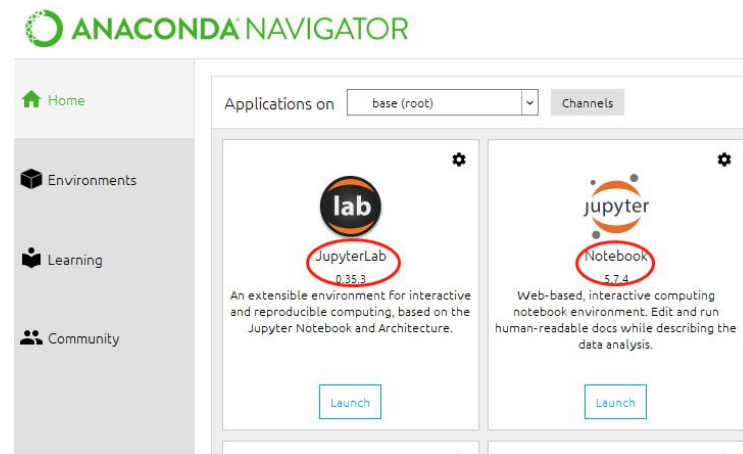


花加技术-AI机器学习实践

FLŌWERPLUS 花加

学习资料

- 莫愁Python: <https://morvanzhou.github.io/learning-steps/>
- 阿里云大数据人工智能应用: <https://ai.aliyun.com/>
- 在线练习比赛: <https://www.kaggle.com/>
- 泰坦尼克号数据分析例子: <https://www.kaggle.com/vgadevik/titanic-analysis/>
- 代码:
https://github.com/herderwu/DIY_ML_Systems_with_Python_1st_Edition/blob/master/Chapter_4/kaggle_titanic.ipynb
- 工具: Anaconda (数据科学和机器学习工具库, 学python都可以用, 写好代码运行一下直接看效果):
<https://www.anaconda.com/distribution/>, 使用JupyterLab和Jupyter Notebook。
- 内网代码: 架构\机器学习\机器学习代码\kaggle_titanic\kaggle_titanic.ipynb



目录

- AI人工智能概要
- 题目介绍
- 代码实战
- 应用场景

目录

- AI人工智能概要
- 题目介绍
- 代码实战
- 应用场景

AI人工智能概要

- 在线体验:

<https://data.aliyun.com/ai#/image-tag>



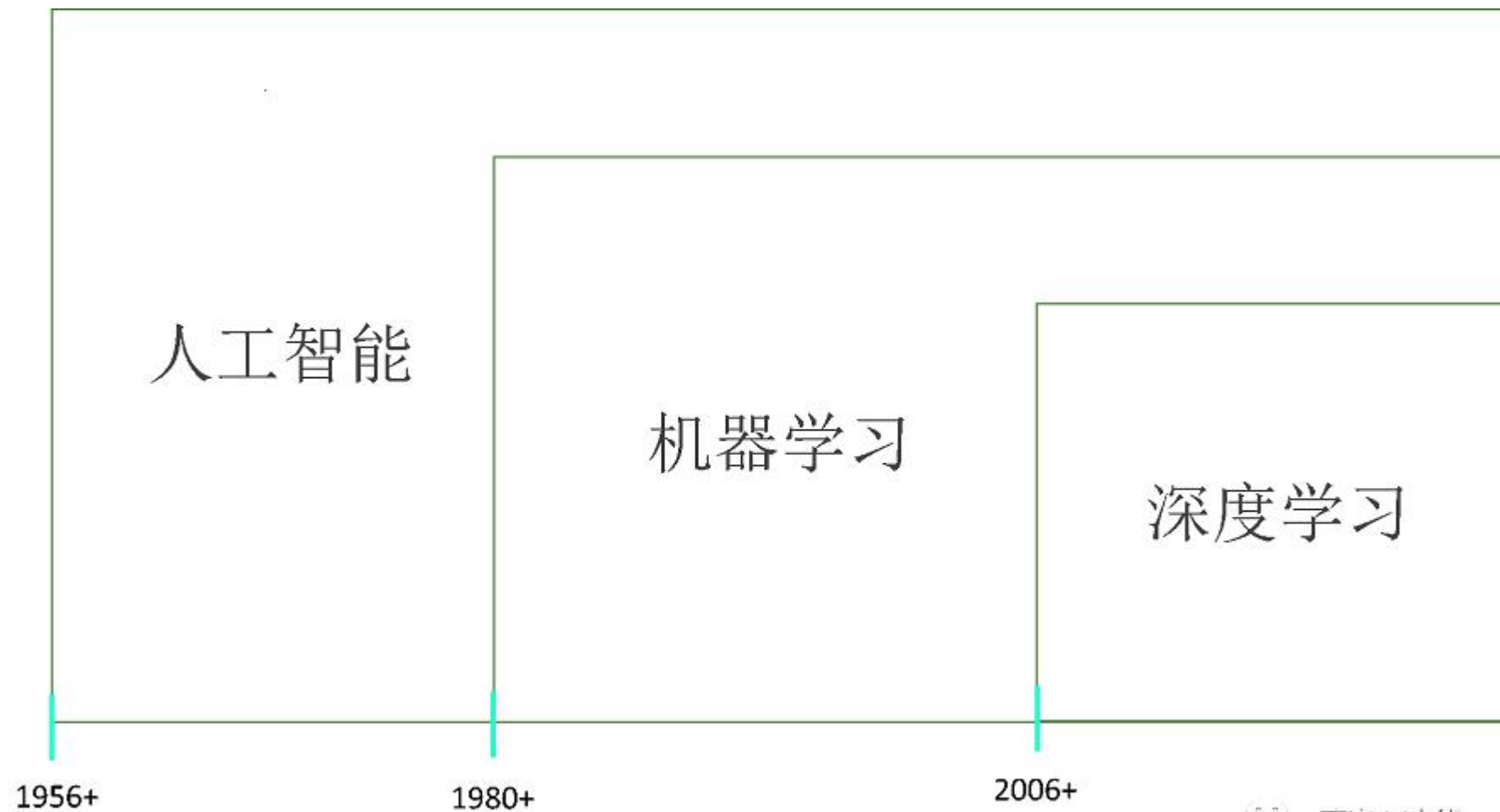
图像标签:

美食	85%
草莓	14%
菠萝	13%
英式甜点	13%
盘子	12%

本地上传 | 请输入图片URL | 识别

AI人工智能概要

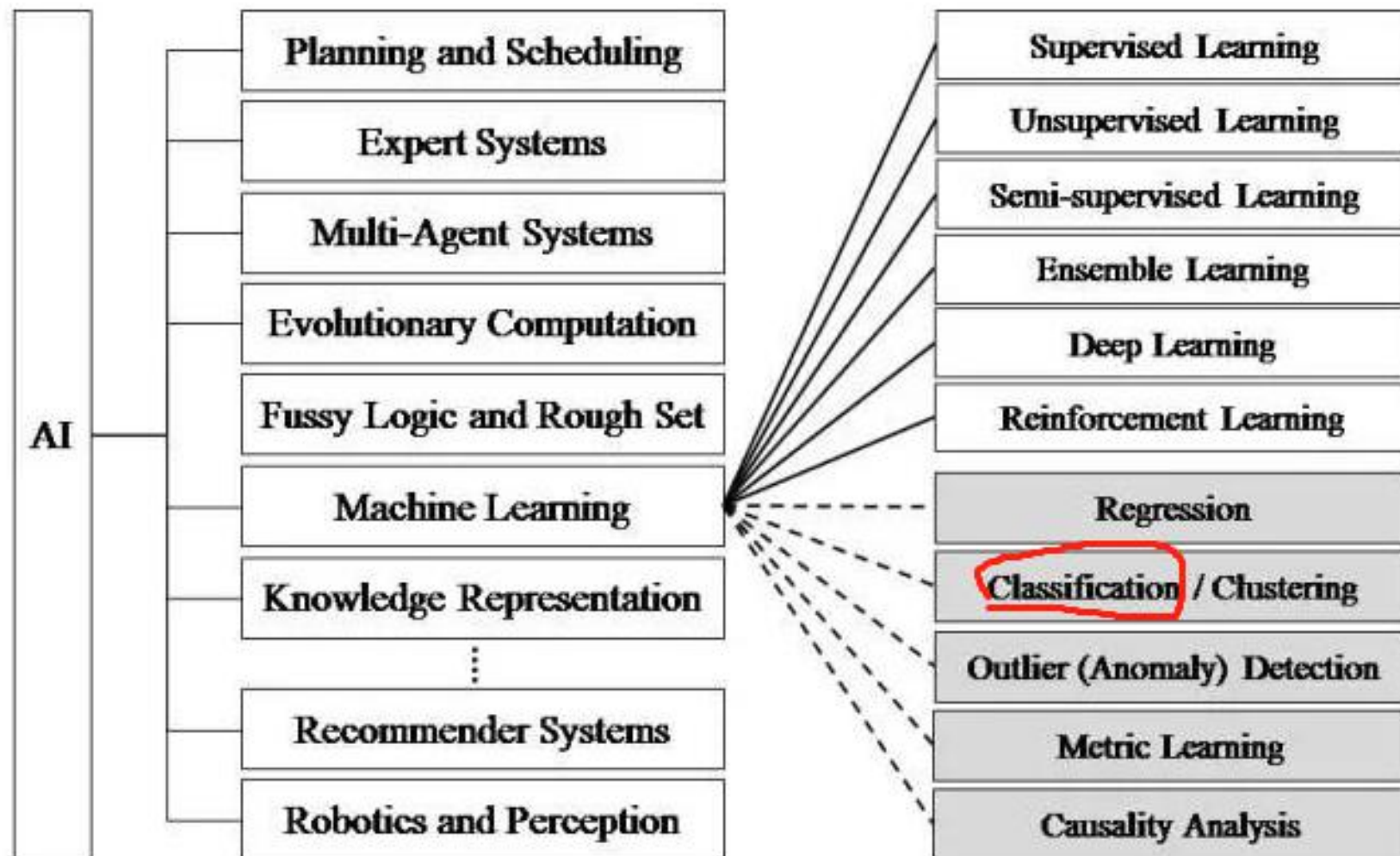
人工智能，机器学习，深度学习的关系：



数据来源：产业创新创投数据平台饮鹿网 (innov100)

AI人工智能概要

人工智能，机器学习，深度学习的关系：



目录

- AI人工智能概要
- 题目介绍
- 代码实战
- 应用场景

题目介绍

- 泰坦尼克号的幸存者的因数分析

- 题目地址：


<https://www.kaggle.com/c/titanic>

- 数据：

<https://www.kaggle.com/c/titanic/data>

- 作业提交

<https://www.kaggle.com/c/titanic/submit>

2 submissions for Joseph Herder		Sort by	Most recent
All Successful Selected			
Submission and Description		Public Score	Use for Final Score
xgbc_submission.csv 12 days ago by Joseph Herder XGBClassifier		0.77511	<input type="checkbox"/>
rfc_submission.csv 17 days ago by Joseph Herder add submission details 		0.73684	<input type="checkbox"/>
No more submissions to show			

题目介绍

- 泰坦尼克号的幸存者的因数分析

- 数据：

<https://www.kaggle.com/c/titanic/data>

- train.csv: 训练数据, 包括参数和幸存数据
- test.csv: 预测输入
- gender_submission.csv: 预测输入的格式

目录

- AI人工智能概要
- 题目介绍
- **代码实战**
- 应用场景

代码实战

- 加载库
- 读取数据
- 查看数据

```
In [1]: 1+1
```

```
Out[1]: 2
```

```
In [2]: # 参考python2的代码: https://github.com/herderwu/DIY\_ML\_Systems\_with\_Python\_1st\_Edition/blob/master/Chapter\_4/Chapter\_4.2.ipynb

# 这是python 3的代码: https://github.com/herderwu/DIY\_ML\_Systems\_with\_Python\_1st\_Edition/blob/master/Chapter\_4/kaggle\_titanic.ipynb

# 机器学习教程: https://morvanzhou.github.io/learning-steps/
```

```
In [3]: import pandas as pd
train=pd.read_csv('../kaggle_titanic/titanic/train.csv')
test=pd.read_csv('../kaggle_titanic/titanic/test.csv')
print(train.info())
print(test.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 12 columns):
PassengerId    891 non-null int64
Survived        891 non-null int64
Pclass         891 non-null int64
Name           891 non-null object
Sex            891 non-null object
Age           714 non-null float64
SibSp          891 non-null int64
Parch          891 non-null int64
Ticket         891 non-null object
Fare           891 non-null float64
Cabin          204 non-null object
Embarked       889 non-null object
dtypes: float64(2), int64(5), object(5)
```

代码实战

- 准备测试数据
- 看数据的完整性

```
In [4]: selected_feature=['Pclass', 'Sex', 'Age', 'Embarked', 'SibSp', 'Parch', 'Fare']
```

```
In [5]: X_train=train[selected_feature]  
X_test=test[selected_feature]
```

```
In [6]: y_train=train['Survived']
```

```
In [7]: print(X_train['Embarked'].value_counts())  
print(X_test['Embarked'].value_counts())
```

```
S    644  
C    168  
Q     77  
Name: Embarked, dtype: int64  
S    270  
C    102  
Q     46  
Name: Embarked, dtype: int64
```

代码实战

• 数据填充

```
In [ ]: X_train['Embarked'].fillna('S', inplace=True)
        X_test['Embarked'].fillna('S', inplace=True)
```

```
In [9]: X_train['Age'].fillna(X_train['Age'].mean(), inplace=True)
        X_test['Age'].fillna(X_test['Age'].mean(), inplace=True)
        X_test['Fare'].fillna(X_test['Fare'].mean(), inplace=True)
```

```
In [11]: print(X_train.info())
         print(X_test.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 891 entries, 0 to 890
Data columns (total 7 columns):
Pclass      891 non-null int64
Sex          891 non-null object
Age          891 non-null float64
Embarked     891 non-null object
SibSp        891 non-null int64
Parch        891 non-null int64
Fare         891 non-null float64
dtypes: float64(2), int64(3), object(2)
memory usage: 48.8+ KB
None
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 7 columns):
Pclass      418 non-null int64
Sex          418 non-null object
Age          418 non-null float64
Embarked     418 non-null object
SibSp        418 non-null int64
Parch        418 non-null int64
Fare         418 non-null float64
dtypes: float64(2), int64(3), object(2)
memory usage: 22.9+ KB
None
```

代码实战

- 加载机器学习库：分类

```
In [16]: from sklearn.ensemble import RandomForestClassifier  
rfc = RandomForestClassifier()
```

```
In [17]: from xgboost import XGBClassifier  
xgbc = XGBClassifier(booster='gblinear')
```

代码实战

- 随机森林：训练，预测和生成csv结果

```
In [23]: rfc.fit(X_train, y_train)
```

```
In [24]: rfc_y_predict = rfc.predict(X_test)

rfc_submission = pd.DataFrame({'PassengerId': test['PassengerId'], 'Survived': rfc_y_predict})
rfc_submission.to_csv('../kaggle_titanic/titanic/rfc_submission.csv', index=False)
```


代码实战

- 梯度提升决策树分类：训练，预测和生成csv结果

```
In [28]: from xgboost import XGBClassifier
```

```
xgbc = XGBClassifier()
```

```
xgbc.fit(X_train, y_train)
```

```
# 打印模型的属性:
```

```
# print(xgbc.coef_)
```

```
# print(xgbc.intercept_)
```


```
Out[28]: XGBClassifier(base_score=0.5, booster='gbtree', colsample_bylevel=1,  
                      colsample_bytree=1, gamma=0, learning_rate=0.1, max_delta_step=0,  
                      max_depth=3, min_child_weight=1, missing=None, n_estimators=100,  
                      n_jobs=1, nthread=None, objective='binary:logistic', random_state=0,  
                      reg_alpha=0, reg_lambda=1, scale_pos_weight=1, seed=None,  
                      silent=True, subsample=1)
```

```
In [29]: xgbc_y_predict = xgbc.predict(X_test)  
xgbc_submission = pd.DataFrame({'PassengerId': test['PassengerId'], 'Survived': xgbc_y_predict})  
xgbc_submission.to_csv('../kaggle_titanic/titanic/xgbc_submission.csv', index=False)
```

代码实战

- 作业提交

<https://www.kaggle.com/c/titanic/submit>

2 submissions for Joseph Herder		Sort by Most recent
All Successful Selected		
Submission and Description	Public Score	Use for Final Score
xgbc_submission.csv 12 days ago by Joseph Herder XGBClassifier	0.77511	<input type="checkbox"/>
rfc_submission.csv 17 days ago by Joseph Herder add submission details 	0.73684	<input type="checkbox"/>
No more submissions to show		

目录

- AI人工智能概要
- 题目介绍
- 代码实战
- 应用场景

应用场景

• 阿里云大数据人工智能应用: <https://ai.aliyun.com/>

数据更智能 业务才更智能



智能语音交互



人脸识别



图像识别



图像搜索



内容安全

录音文件识别

提供的是将语音转写成文字的服务。

实时语音转写

对音频流做实时转写，达到“边说边出文字”的效果。

一句话识别

对时长较短（一分钟以内）的语音进行转写。

语音合成

语音合成服务（TTS），就是将文本转成语音的服务。

语音合成声音定制

为企业提供深度定制TTS声音的能力。

语言模型自学习工具

通过文本数据自学习训练语音模型，以达到定制效果。



印刷文字识别



自然语言处理



机器翻译



机器学习PAI



大数据计算



大数据搜索与分析



数据开发



数据可视化



大数据应用

交流

FLŌWERPLUS 花加