

实验操作： 中文医学短文本分类

宗辉，李作峰

- 了解词向量和TF-IDF的概念。
- 基于Sklearn+JupyterNotebook, 学会机器学习模型的基本使用。
- 能够训练模型解决医学临床中的自然语言处理问题(如文本分类等)。
- 基于Pyramid+Docker, 将模型搭建为网页应用。

实验目标

- 成功搭建自己的AI应用

Welcome!



DOCS

GITHUB

DOWNLOAD

Chinese Medical Short Text Classification

Main Steps

1. Input a short medical text.
2. Select a classification method.
☒ MyModel(new) ☐ LR(existed) ☐ SVM(existed) ☐ KNN(existed) ☐ NB(existed) ☐ CNN ☐ Bert



实验介绍

- 在本次实验中，我们给定事先定义好的15种类别和一系列中文临床短文本的描述语句，要求返回每一条输入短文本的具体类别。如：

No.	Category	No.	Category
1	Addictive Behavior	9	Laboratory Examinations
2	Age	10	Life Expectancy
3	Allergy Intolerance	11	Organ or Tissue Status
4	Compliance with Protocol	12	Pharmaceutical Substance or Drug
5	Consent	13	Risk Assessment
6	Diagnostic	14	Smoking Status
7	Disease	15	Therapy or Surgery
8	Enrollment in other studies		

ID	输入（筛选标准）	输出（类别）
s1	年龄>80 岁	Age
s2	近期颅内或椎管内手术史	Therapy or Surgery
s3	血糖<2.7mmol/L	Laboratory Examinations

实验数据

- **数据特点：**自由文本格式，短文本，非结构化，中文，医学相关。
- **训练数据8000条：**

	sentence	category
0	(2)若伴便秘者符合罗马IV功能性便秘诊断标准，若伴夜尿症者符合夜尿症的诊断标准；	Diagnostic
1	(1) 患者拟行急症手术；	Therapy or Surgery
2	c) 在过去6个月内参加过I、II期临床试验或者3个月内参加过III、IV期临床试验；	Enrollment in other studies
3	4. 身高体重指数(BMI)>=25得病人	Risk Assessment
4	5.愿意参与该研究并配合调查者；	Consent

- **测试数据2000条：**

	sentence
0	1.符合脓毒症诊断Sepsis 3.0版标准；
1	1) 符合WHO对不孕症的诊断标准；
2	3. 肿瘤直径 ≥8mm且≤30mm；
3	8.合并其他运动可能加重的神经、肌肉、骨骼肌、风湿性疾病；
4	2. 符合国际疾病分类（ICD-10）编码 J21 的毛细支气管炎诊断标准。

训练数据类别&数目：

	category	count
0	Addictive Behavior	196
1	Age	638
2	Allergy Intolerance	430
3	Compliance with Protocol	294
4	Consent	874
5	Diagnostic	794
6	Disease	1177
7	Enrollment in other studies	358
8	Laboratory Examinations	755
9	Life Expectancy	101
10	Organ or Tissue Status	258
11	Pharmaceutical Substance or Drug	578
12	Risk Assessment	442
13	Smoking Status	33
14	Therapy or Surgery	1072


- **实验平台:** Anaconda3, JupyterNotebook, Python3, Docker

已经预装好

- **所需python包:** pandas, numpy, codecs, scikit-learn==0.21,
jieba, wordcloud, cornice, xlrd, pickle5,
joblib, scipy

```
pip install -r requirements.txt
```

词向量-TFIDF

- **TF-IDF**全称Term Frequency - Inverse Document Frequency，是一种文本特征提取算法，由两部分组成。
 - **词频(TF)**: 文本中各个词的出现频率统计, 是词语出现的次数除以该文件的总词语数。
 - **逆文档频率(IDF)**:是文档频率(Df)的倒数，Df是出现某词语的文档数除以总文档数。
-
- TF-IDF可通过scikit-learn计算。
 - 按词频排序，选取前1000个词，显示如右所示。
- 



机器学习-分类模型

- 逻辑回归, Logistic Regression
- 支持向量机, Support Vector Machine
- K近邻算法, K Nearest Neighbors
- 朴素贝叶斯, Naive Bayes
- 神经网络, Neural network

以上模型实现的代码，均在文件夹notebooks中

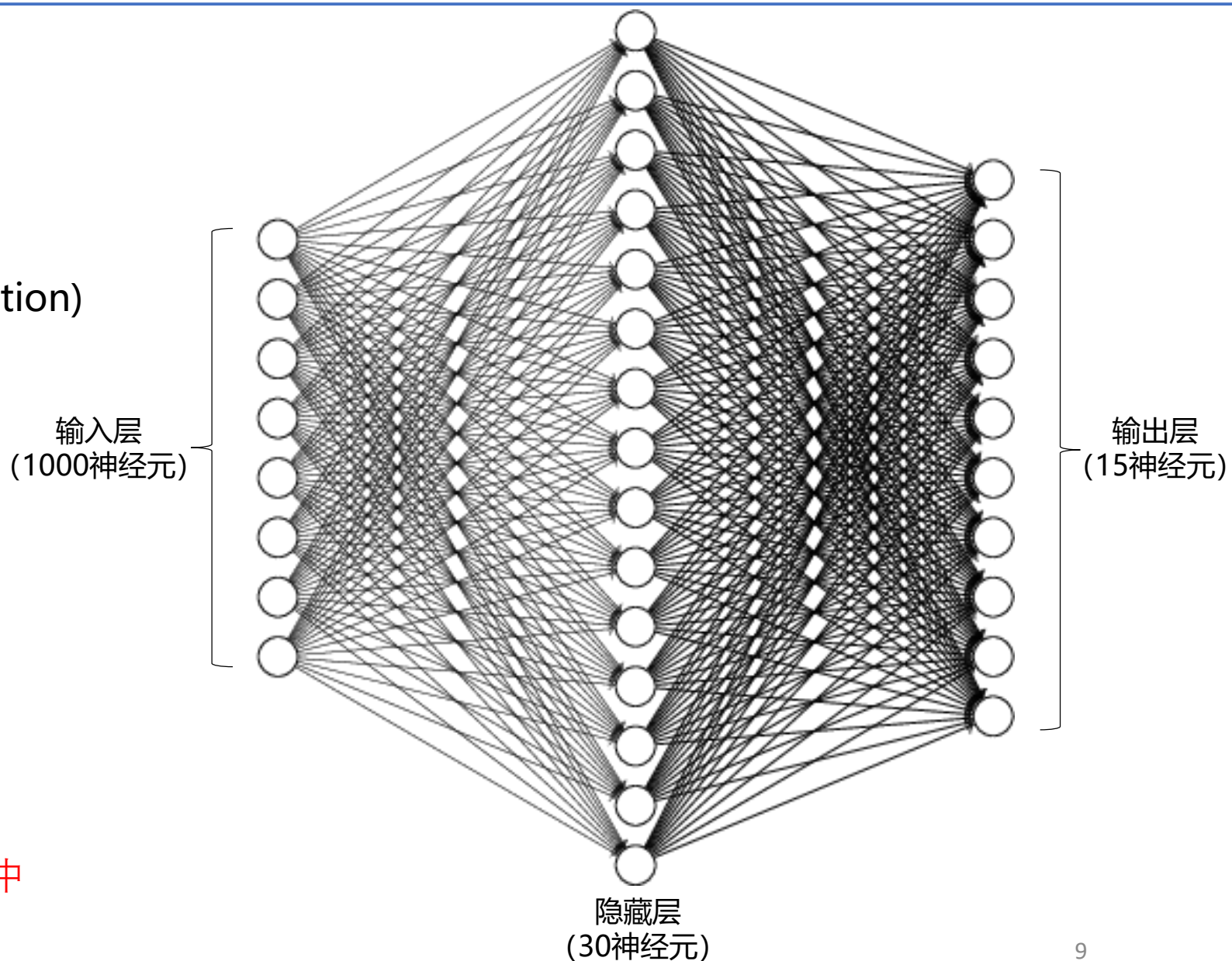
神经网络

- 超参数

- 学习率(learning rate)
- 权值初始化(Weight Initialization)
- 网络层数(Layers)
- 单层神经元数(Units)

- 随机梯度算法(SGD)

- 迭代期(Epoch)
- 批大小(mini-batch)



模型实现的代码，文件夹notebooks中

评价指标

- 本任务的评价指标包括宏观准确率(Macro Precision), 宏观召回率(Macro Recall), Average F1值。最终排名以Average F1值为基准。
- 假设我们有n个类别, $C_1, \dots, C_i, \dots, C_n$ 。

$$\text{准确率 } P_i = \frac{\text{正确预测为类别 } C_i \text{ 的样本个数}}{\text{预测为 } C_i \text{ 类的样本个数}}$$

$$\text{召回率 } R_i = \frac{\text{正确预测为类别 } C_i \text{ 的样本个数}}{\text{真实的 } C_i \text{ 类的样本个数}}$$

$$\text{Average F1} = \left(\frac{1}{n}\right) \sum_{i=1}^n \frac{2 * P_i * R_i}{P_i + R_i}$$

动手操作-一共三个步骤

https://github.com/zonghui0228/cn_med_text_class

- 下载git仓库:

```
git clone https://github.com/zonghui0228/cn_med_text_class
```

- 基于JupyterNotebook, 训练机器学习分类器模型

KNN, LR, NB, SVM 任选其一

- 基于Docker, 搭建pyramid使用界面

```
docker run -it -d -p 6543:6543 zonghui0228/cn_med_text_class
```

动手操作-步骤1， 下载github项目

- 在桌面上建立一个文件夹命名为experiment。
- 进入文件夹， 按住shift， 点击鼠标右键， 选择在此处打开Powershell Window。

- 输入命令， 从github下载实验相关文件。

```
git clone https://github.com/zonghui0228/cn_med_text_class
```

- 在powershell window中输入下述命令， 按照python依赖包。

```
pip install -r requirements.txt
```



动手操作-步骤2，训练自己的模型

- 在powershell window中输如 `jupyter notebook`，会弹出网页的编辑界面。
- 进入notebooks文件夹，选择一个模型文件如（`logistic_regression.ipynb`），点开，按照操作步骤运行，训练机器学习模型。
- 输如命令，从github下载实验相关文件。

```
git clone https://github.com/zonghui0228/cn_med_text_class
```

动手操作-步骤3，创建自己的AI网页应用

- 通过WinSCP+putty进入服务器，创建文件夹model，并将刚才训练好的三个模型文件移动到该文件夹下。

- 通过docker启动基于pyramid搭建的网页应用。

```
docker run -it -d -p 6543:6543 zonghui0228/cn_med_text_class
```

这个地方更改为任意端口数字

这是容器id，通过命令*docker ps -a*查询

- 将自己的模型文件加载进去。

```
docker cp ./ CONTAINER_ID:/home/zonghui/mynginx/myproj/myproj/views/model/mymodel
```

- 在window中打开网页，输入 `http://ip:6543/index`

这个地方是服务器的ip地址，通过命令*ifconfig*查看

这与上面的数字同步