以下是笔者的成绩排名：

排名	成员名称	最高得分	最高分提交时间
11	hacker	1760	2023年5月25日 11:45
12	HN-王宇航	1760	2023年5月25日 11:52
13	dhaia	1720	2023年5月25日 11:40
14	曹东	1680	2023年5月25日 11:42
15	w147865932	1660	2023年5月25日 11:55
16	GRD-李之赞	1650	2023年5月25日 11:42
17	GD-tan91	1640	2023年5月25日 11:55
18	doit-zs	1600	2023年5月25日 11:55
19	GRD-Reznov	1580	2023年5月25日 11:55
20	N13Saxion	1560	2023年5月25日 11:55

本次的数据思维赛总共有两道题，分别是数据预测和命名实体识别，以下是作者个人的 writeup：

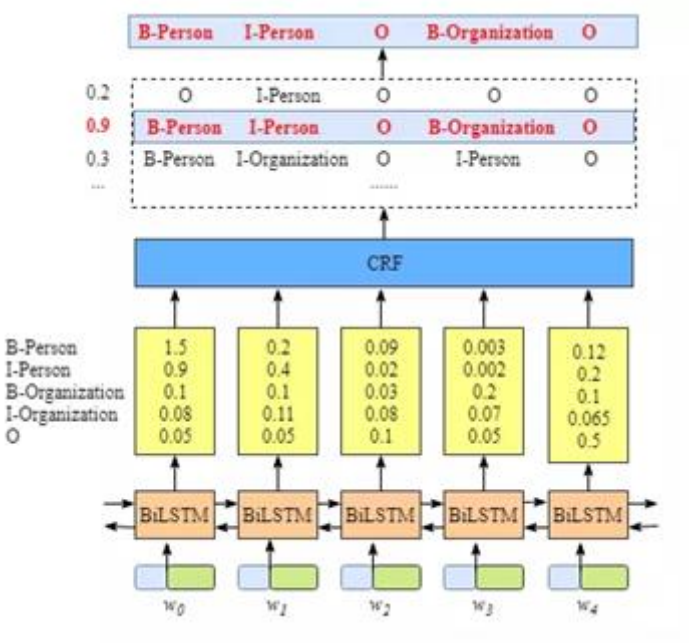
对于第一道题家电能源预测，题目的训练集给了 28 个维度，测试集的数据给了 18 个维度，然后让我们根据测试集的 18 个维度预测其另外 8 个维度的数据。对于这种数据预测的任务，如果数据是符合时间序列特征的，那么 LSTM 模型是首选。但是这道题目的 18 个维度中不包含时间特征，所以只能用暴力的全连接层拟合。

```
=================================================================
 dense (Dense)               (None, 34)                612

 dense_1 (Dense)             (None, 68)                2380

 dense_2 (Dense)             (None, 136)               9384

 dense_3 (Dense)             (None, 272)               37264

 dense_4 (Dense)             (None, 546)               149058

 dense_5 (Dense)             (None, 1092)              597324

 dense_6 (Dense)             (None, 546)               596778

 dense_7 (Dense)             (None, 272)               148784

 dense_8 (Dense)             (None, 136)               37128

 dense_9 (Dense)             (None, 68)                9316

 dense_10 (Dense)            (None, 34)                2346

 dense_11 (Dense)            (None, 17)                595

 dense_12 (Dense)            (None, 8)                 144

=================================================================
Total params: 1,591,113
Trainable params: 1,591,113
```

这是我采用的网络结构，采用的思想是每层全连接层参数都是前者的两倍，目的是为了提取更深层次的特征，所有的激活函数使用的都是 relu。当数据特征提到 1092 维度时，进行降维，每次降维的全连接层参数都是前者的二分之一，这样不至于出现特征消失的情况，由于题目的评估标准使用的是 MSE 平方损失函数，因而训练过程中采用的损失函数是 mse。

最终的损失函数值降到了 2.8。

对于第二道题目，题目的本质其实是命名实体识别，目前已经有许多的优秀模型可以做命名实体识别的任务，常见的有 BiLstm+CRF，BERT，BERT+CRF 等等。笔者这边使用的是较为经典但效果不错的 BiLstm+CRF 模型。



对于 BiLstm 模型，它已经是老生常谈的模型了，这边不再做详细的介绍，主要来讲讲 CRF 条件随机场模型。

在介绍 CRF 条件随机场模型之前，先以一个具体的分词案例来介绍一下 HMM 隐马尔科夫模型。

这边首先定义几个标识符 B,M,E,S。

B 表示词语的首个字，M 表示词语的中间字，E 表示词语的最后一个字，S 表示单独的字。

举个例子：我喜欢打篮球，分词之后为：我 喜欢 打篮球。这边“我”就是独立的字，标识符为 S，“喜欢”是一个词语，表示符为 BE，“打篮球”是一个词组，标识符为 BME。所以整句句子分词之后的结果就是 SBEBME。

然后介绍一下隐马尔科夫模型的三个重要概念，初始矩阵，状态转移矩阵和发射矩阵。

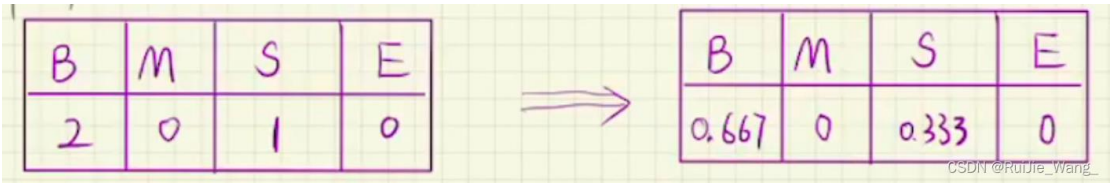
初始矩阵表示的是句子的第一个字符为 BMES 的概率，举个例子，现在有三句话：

今天 天气 真 不错 。

麻辣肥牛 好吃 ！

我 喜欢 吃 好吃 的 ！

观察这三句话的第一个字符，发现有两句句字的标签为 B，一句句子为 S，所以有：



上图中的右半部分即为初始矩阵。

转移矩阵即为知道前一个字符的标识符，推断后一个字符的标识符的概率矩阵。

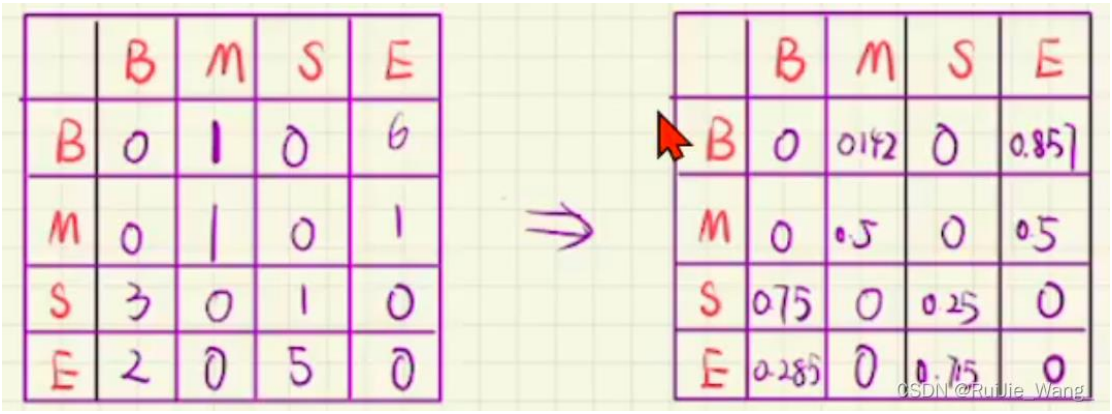
举个例子，用上面三句话做为语料，可以得知它们的标签为：

B E B E S B E S

B M M E B E S

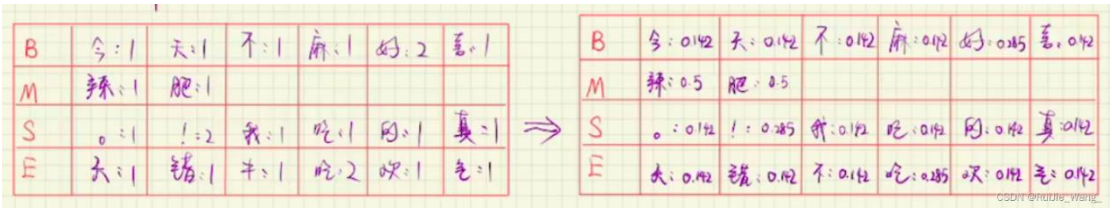
S B E S B E S S

那么统计词频可以得到：

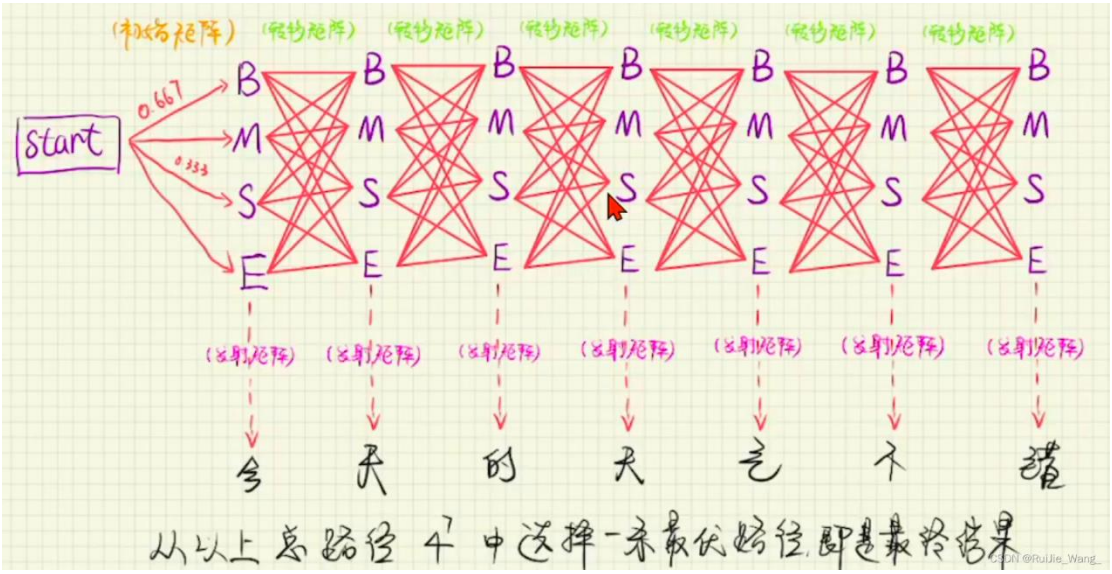


上图的右边即为状态转移矩阵。

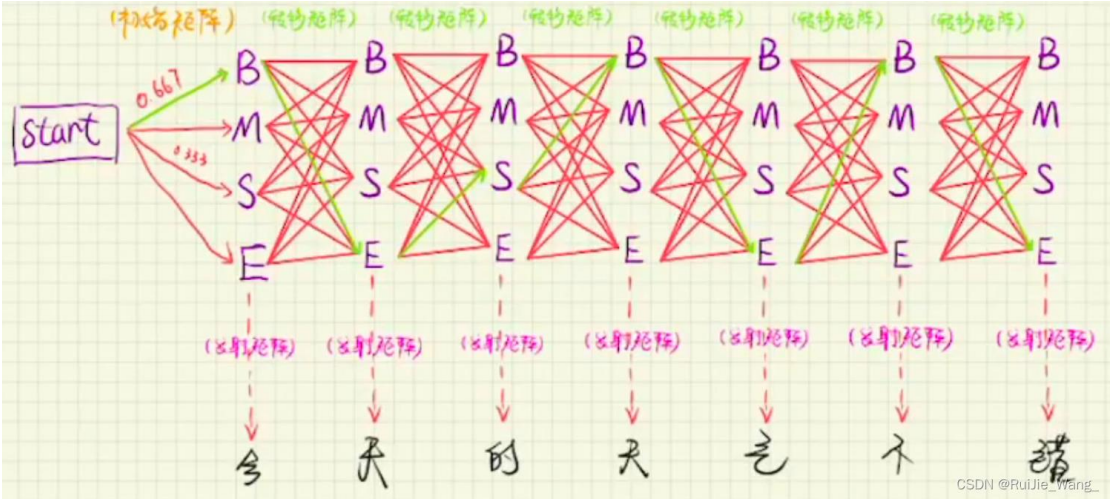
而发射矩阵就是一个字符对应为 BMES 的概率，还是以这三句话为例，那么它的发射矩阵为：



接着就可以做预测了，假设我们要将“今天的天气不错”这句句子分词，那么就可以转化为下图：



假设下图中绿色的那条为最优路径：



B	M	S	E
0.667	0	0.333	0

初始矩阵

	B	M	S	E
B	0	0.142	0	0.857
M	0	0.5	0	0.5
S	0.75	0	0.25	0
E	0.285	0	0.715	0

转移矩阵

B	今：0.142	天：0.142	不：0.142	所：0.142	好：0.285	气：0.142	
M	错：0.5	肥：0.5					
S	。：0.142	！：0.285	你：0.142	呢：0.142	的：0.142	真：0.142	
E	天：0.142	我：0.142	不：0.142	吃：0.285	饭：0.142	气：0.142	

发射矩阵

那么有：

$$0.667 \times 0.142 \times [0.857 \times 0.142] \times [0.75 \times 0.142] \times [0.25 \times 0.142] \times [0.857 \times 0.142] \times [0.285 \times 0.142] \times [0.857 \times 0.142]$$

今 天 的 天 气 不 错

将上面的所有路径的值都计算一遍，取最大路径那条作为输出。

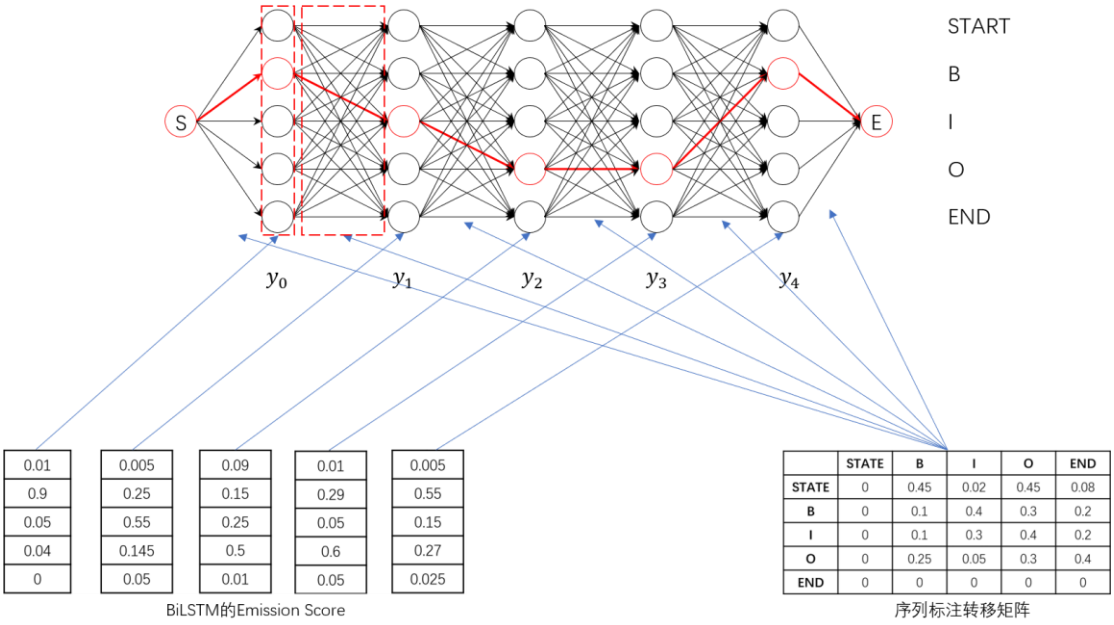
当然，这边还可以采取维特比算法等优化算法，这边就不再介绍。

介绍完隐马尔科夫模型，接着来讲一下条件随机场模型：

CRF 将 BiLSTM 的输出结果当作发射矩阵作为输入，输出符合标注转移约束条件

的、最大可能的预测标注序列。

CRF 其实是一种特殊的隐马尔科夫模型，只不过里面的状态转移矩阵不是基于词频统计，而是通过反向传播梯度下降的方法获得。里面的发射矩阵也不是基于词频，而是用 BiLstm 模型的输出取代。



	STATE	B	I	O	END
STATE	0	0.45	0.02	0.45	0.08
B	0	0.1	0.4	0.3	0.2
I	0	0.1	0.3	0.4	0.2
O	0	0.25	0.05	0.3	0.4
END	0	0	0	0	0

然后介绍一下本题中的网络安全空间的命名实体识别：

首先观察一下数据集，发现题目已经将句子分好了词，并且给每个词加上了标签，一共有 24 类标签，但 0~23 数字所对应的含义并没有告诉我们。

id	sentence_i	words	tag
0	0	Google	7
1	0	V8	0
2	0	before	3
3	0	3.14.5.3	4
4	0	as	0
5	0	used	0
6	0	in	0
7	0	Google	7
8	0	Chrome	5
9	0	before	3
10	0	24.0.1312.5	4
11	0	allows	1
12	0	remote	1
13	0	attackers	2
14	0	to	0
15	0	cause	0
16	0	a	0
17	0	denial	1
18	0	of	2
19	0	service	2
20	0	or	0
21	0	possibly	0
22	0	have	0
23	0	unspecifie	0
24	0	other	0
25	0	impact	0
26	0	via	0
27	0	crafted	0

由于题目的要求是运用模型将测试集的所有数据打上标签，将标签文件上传系统，系统将自动评分并进行排名。而测试集中又包含一些训练集中不存在的词，所以我们要用<UNK>字符来表示这些词，观察数据集，会发现存在 3.14.5.3 等版本号类似的词，所以我们规定用<NUM>标签来表示这些词。然后就可以进行命名实体识别了，这边将所的句子进行切分操作，规定句子的最大长度为 50，模型的结构如下：

```
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_ids (InputLayer)         [(None, 50)]         0           []

 embedding (Embedding)          (None, 50, 256)      1773312     ['input_ids[0][0]']

 bidirectional (Bidirectional)  (None, 50, 256)      394240      ['embedding[0][0]']

 bidirectional_1 (Bidirectional  (None, 50, 128)     164352      ['bidirectional[0][0]']
 )

 dense (Dense)                  (None, 50, 24)       3096        ['bidirectional_1[0][0]']

 target_ids (InputLayer)        [(None, 50)]         0           []

 input_lens (InputLayer)        [(None,)]            0           []

 crf (CRF)                      ()                   576         ['dense[0][0]',
                                                                  'target_ids[0][0]',
                                                                  'input_lens[0][0]']

==================================================================================================
Total params: 2,335,576
Trainable params: 2,335,576
Non-trainable params: 0
__________________________________________________________________________________________________
```

最终的损失值函数降到了 22。