

# 2020 机器学习课程作业

(课程作业40%+闭卷考试60%)

# 总体安排

作业分为两部分(各占成绩的50%):

## 1.个人作业：四选一

在“北航数据工作站”比赛网站上，提交结果，进行**成绩排名**。

提交**技术报告**和**可复现代码**，详细介绍使用的技术方案（不限于模型原理、模型结构、调参训练过程、中间结果、**比赛工作站成绩及排名截图**等）。

## 2.团队**(最多5人)**作业：二选一

在“北航数据工作站”比赛网站上，提交结果，进行**成绩排名**。

提交**技术报告**和**可复现代码**，详细介绍使用的技术方案（不限于模型原理、模型结构、调参训练过程、中间结果、比赛工作站成绩截图等）。

技术报告开头注明**每位成员在团队作业中分工和贡献**。

团队作业课堂展示环节：ppt讲解5分钟，老师提问3分钟。（根据展示情况加0-5分）

# 个人作业（四选一）

每道题目至少使用两种机器学习算法，  
鼓励尝试多种解决方案！

# 1. 球员能力评测

通过分析球员的数据，可以对球员能力进行很好的鉴定与评测。你的任务是根据球员的各项信息和技术表现值来预测该球员的综合能力。

训练集中共有7900条样本，测试集中有2500条样本。每条样本代表一位球员，数据中每个球员有60项属性。



birth\_date 生日。格式为月/日/年。  
height\_cm 身高（厘米）。  
weight\_kg 体重（公斤）。  
nationality 国籍。已被编码。  
potential 球员的潜力。数值变量。  
pac 球员速度。数值变量。  
sho 射门（能力值）。数值变量。  
pas 传球（能力值）。数值变量。  
dri 带球（能力值）。数值变量。  
def 防守（能力值）。数值变量。  
phy 身体对抗（能力值）。数值变量。  
skill\_moves 技巧动作。数值变量。  
weak\_foot 非惯用脚的能力值。数值变量。  
work\_rate\_att 球员进攻的倾向。分类变量，Low, Medium, High。  
work\_rate\_def 球员防守的倾向。分类变量，Low, Medium, High。  
preferred\_foot 惯用脚。1表示右脚、2表示左脚。  
crossing 传中（能力值）。数值变量。  
finishing 完成射门（能力值）。数值变量。  
heading\_accuracy 头球精度（能力值）。数值变量。  
short\_passing 短传（能力值）。数值变量。  
volleys 凌空球（能力值）。数值变量。  
dribbling 盘带（能力值）。数值变量。  
curve 弧线（能力值）。数值变量。  
free\_kick\_accuracy 定位球精度（能力值）。数值变量。  
long\_passing 长传（能力值）。数值变量。  
ball\_control 控球（能力值）。数值变量。  
acceleration 加速度（能力值）。数值变量。  
sprint\_speed 冲刺速度（能力值）。数值变量。  
agility 灵活性（能力值）。数值变量。  
reactions 反应（能力值）。数值变量。  
balance 身体协调（能力值）。数值变量。

shot\_power 射门力量（能力值）。数值变量。  
jumping 弹跳（能力值）。数值变量。  
stamina 体能（能力值）。数值变量。  
strength 力量（能力值）。数值变量。  
long\_shots 远射（能力值）。数值变量。  
aggression 侵略性（能力值）。数值变量。  
interceptions 拦截（能力值）。数值变量。  
positioning 位置感（能力值）。数值变量。  
vision 视野（能力值）。数值变量。  
penalties 罚点球（能力值）。数值变量。  
marking 卡位（能力值）。数值变量。  
standing\_tackle 断球（能力值）。数值变量。  
sliding\_tackle 铲球（能力值）。数值变量。  
gk\_diving 门将扑救（能力值）。数值变量。  
gk\_handling 门将控球（能力值）。数值变量。  
gk\_kicking 门将开球（能力值）。数值变量。  
gk\_positioning 门将位置感（能力值）。数值变量。  
gk\_reflexes 门将反应（能力值）。数值变量。  
rw 球员在右边锋位置的能力值。数值变量。  
rb 球员在右后卫位置的能力值。数值变量。  
st 球员在射手位置的能力值。数值变量。  
lw 球员在左边锋位置的能力值。数值变量。  
cf 球员在锋线位置的能力值。数值变量。  
cam 球员在前腰位置的能力值。数值变量。  
cm 球员在中场位置的能力值。数值变量。  
cdm 球员在后腰位置的能力值。数值变量。  
cb 球员在中后卫的能力值。数值变量。  
lb 球员在左后卫的能力值。数值变量。  
gk 球员在守门员的能力值。数值变量。  
y 该球员的综合能力值。这是要被预测的数值。

## 2. 贷款资格审查

[illegible]

在信贷风控领域，随着大数据、计算机集群技术、网络技术和人工智能的发展，越来越多的金融机构将传统的策略风控手段转向依赖机器学习模型等量化手段。信贷环节中的审批、预警、催收以及营销等诸多场景也适合机器学习模型的应用。

训练集数据包括对每个申请贷款人采集的36条信息(Q1~Q36)，信息已经被编码为36维特征向量，我们的任务是根据这36条信息预测该贷款人申请成功的概率。训练集17万条，测试集3万条。

### 3.学生成绩预测

gender	race/ethnicity	parental level of education	lunch	test preparation course	math score	reading score	writing score
female	group D	some high school	standard	none	77	84	85
female	group C	some college	free/reduced	none	44	64	57
male	group D	master's degree	standard	completed	91	91	92
male	group C	some high school	standard	completed	75	62	61
female	group C	some college	free/reduced	completed	78	90	87
female	group B	some college	standard	none	81	88	90
male	group E	master's degree	standard	none	71	68	64
female	group D	master's degree	standard	none	74	77	86
female	group D	high school	free/reduced	completed	55	66	70
female	group B	some high school	free/reduced	none	20	35	39
female	group B	high school	standard	completed	51	61	63
female	group D	some college	standard	none	68	64	67
female	group C	some college	free/reduced	none	57	75	67
female	group B	high school	free/reduced	none	65	85	84
male	group C	associate's degree	standard	none	89	91	85
female	group B	some college	free/reduced	none	61	72	64
female	group D	associate's degree	standard	completed	81	85	90
female	group B	some college	standard	none	64	71	72
female	group C	associate's degree	free/reduced	none	77	82	83
male	group D	some college	standard	none	75	67	73

给定学生的信息（包括性别、种族、父母受教育水平、午餐标准、是否参与备考课程），预测其数学、阅读、写作成绩。

训练集给定了5000名学生的信息和成绩，测试集包含1000名学生的信息，你需要提交他们的成绩。

评估标准为：对所有学生的预测结果和真实成绩计算平均平方差，差值越低越好。



## 4. 汽车保险预测

	A	B	C	D	E	F	G	H	I	J	K	L
1	id	Gender	Age	Driving_Li	Region_Co	Previously	Vehicle_Ac	Vehicle_D	Annual_Pr	Policy_Sale	Vintage	Response
2	1	Male	44	1	28	0	> 2 Years	Yes	40454	26	217	1
3	2	Male	76	1	3	0	1-2 Year	No	33536	26	183	0
4	3	Male	47	1	28	0	> 2 Years	Yes	38294	26	27	1
5	4	Male	21	1	11	1	< 1 Year	No	28619	152	203	0
6	5	Female	29	1	41	1	< 1 Year	No	27496	152	39	0
7	6	Female	24	1	33	0	< 1 Year	Yes	2630	160	176	0
8	7	Male	23	1	11	0	< 1 Year	Yes	23367	152	249	0
9	8	Female	56	1	28	0	1-2 Year	Yes	32031	26	72	1
10	9	Female	24	1	3	1	< 1 Year	No	27619	152	28	0
11	10	Female	32	1	6	1	< 1 Year	No	28771	152	80	0
12	11	Female	47	1	35	0	1-2 Year	Yes	47576	124	46	1
13	12	Female	24	1	50	1	< 1 Year	No	48699	152	289	0
14	13	Female	41	1	15	1	1-2 Year	No	31409	14	221	0
15	14	Male	76	1	28	0	1-2 Year	Yes	36770	13	15	0
16	15	Male	71	1	28	1	1-2 Year	No	46818	30	58	0
17	16	Male	37	1	6	0	1-2 Year	Yes	2630	156	147	1
18	17	Female	25	1	45	0	< 1 Year	Yes	26218	160	256	0
19	18	Female	25	1	35	1	< 1 Year	No	46622	152	299	0
20	19	Male	42	1	28	0	1-2 Year	Yes	33667	124	158	0
21	20	Female	60	1	33	0	1-2 Year	Yes	32363	124	102	1
22	21	Male	65	1	28	0	1-2 Year	Yes	41184	124	116	0
23	22	Male	49	1	28	0	1-2 Year	Yes	50791	124	177	0
24	23	Male	23	1	50	1	< 1 Year	No	45283	152	232	0
25	24	Male	44	1	28	0	1-2 Year	Yes	41852	163	60	0
26	25	Male	34	1	15	1	1-2 Year	No	38111	152	180	0
27	26	Female	21	1	28	1	< 1 Year	No	61964	152	72	0
28	27	Female	51	1	28	0	1-2 Year	Yes	38241	124	40	1

某保险公司希望向其现有的疾病保险客户推广汽车商业保险，希望你能根据客户的信息预测推广的成功概率。

训练集包含300K用户数据的信息（性别、年龄、是否持有驾照、是否经历过汽车损毁、汽车购入时间、所在地区、持有疾病保险的年金等），以及该用户是否对汽车保险感兴趣。

测试集（80K）仅包括用户信息但不包括该用户是否对汽车保险的兴趣，你需要进行预测。

# 团队作业（三选一）

鼓励尝试多种解决方案！



# 一. 问题摘要生成

---

宝宝为什么总是吐舌头啊？

Why does my baby always stick his tongue out ?

*Question*

---

我家宝宝出生快满四个月了，这几天我忽然发现宝宝总是吐舌头，而且口水也很多，那么这到底是咋回事啊？

My baby is almost four months old. In these few days, I suddenly found that my baby always stick his tongue out and has a lot of saliva. So what is this?

*Description*

---

文本摘要任务是自然语言处理领域的重要任务，将一篇文章的关键信息进行总结，可以帮助人们快速了解文档的内容。在社区问答中，有的问题非常冗长，有的在题干中没有具体描述问题，这些都不利于回答者被吸引到问题中来。比如“懂神经网络的大虾请来回答”“遇到这种事情我该怎么办”都不是好的题干，它们或许很重要，然而不容易吸引到人们的关注。这时候，社区就有必要对于这些问题描述重新进行问题摘要生成。

你的任务是，给定一个问题的详细描述（description），生成这个问题的题干。

## 二. 动漫Face检测



近年来，伴随着动漫产业的迅猛发展，动漫视频呈现出爆炸性增长。而实现对这些动漫视频智能理解的第一步就是需要检测和识别出这些视频里面的动漫人物身份信息。目前的人脸检测算法对于真实人脸的检测性能已经非常好了，那么对于动漫角色Face的检测会有怎样的不同呢？

你的任务是，通过给定的动漫Face检测数据集来设计一个针对动漫角色的Face检测模型。

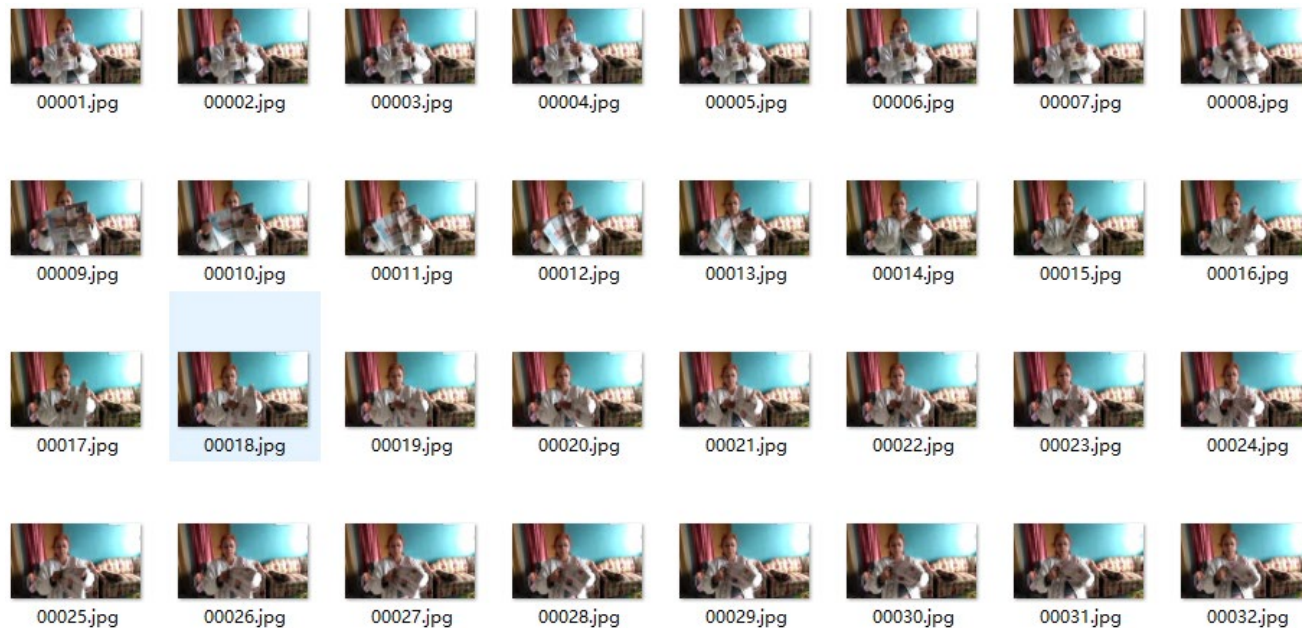
### 三. 视频动作识别

数据集包括174类不同的动作视频， 一共10000个视频，其中8000个用于训练， 2000个用于测试。

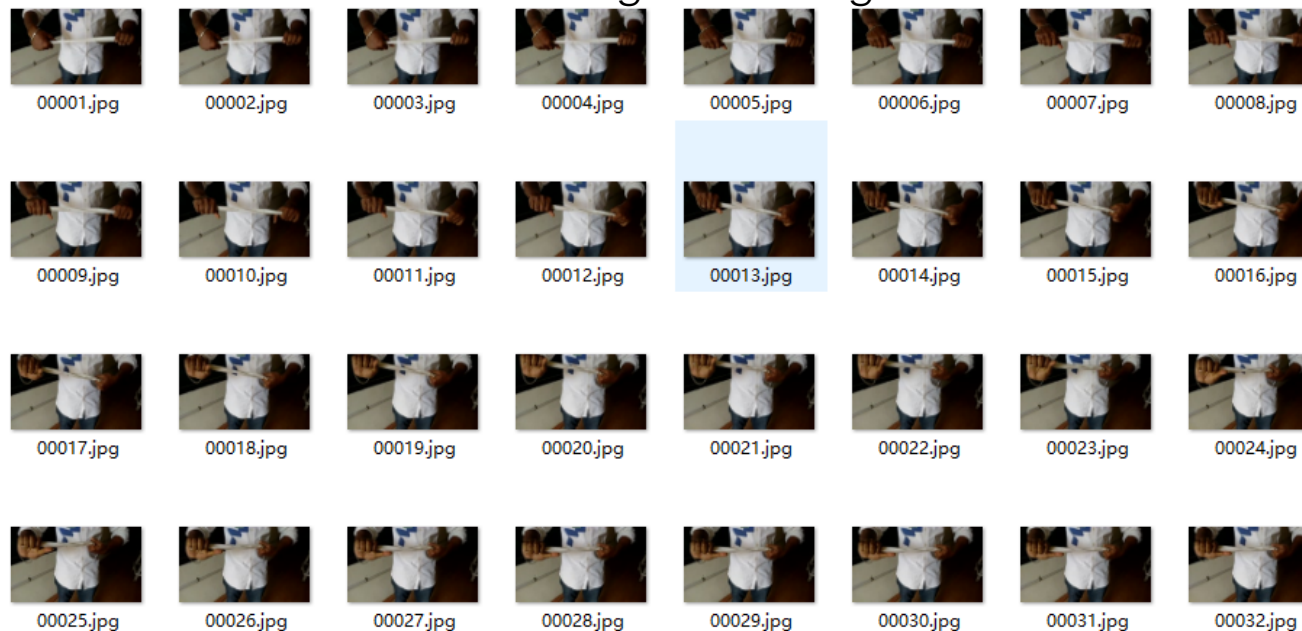
已经为大家对视频抽完帧，每个子文件夹代表一个视频，子文件夹下是视频抽帧后得到的图片。数据集部分样例和标注如右图所示。

你的任务是，任意给定一段视频序列，预测其相应的正确动作类别。

Unfolding something



Throwing something



# 比赛平台使用方法



# 北航数据工作站

基于高校群体智能促进数据科学研究发展

热门比赛

精选比赛

注册

竞赛(我的工作站)

\*用户名

BY1806\*\*\*\_张三

学号+姓名!!!

\*电子邮箱

\*\*\*\*\*

\*密码

\*\*\*\*\*

\*确认密码

\*\*\*\*\*

☒ 我接受相关隐私条例

注册

竞赛

我参加的竞赛

我举办的竞赛

我的数据集



## 贷款资格审查

竞赛组织者: 2020研究生机器学习助教

在信贷风控领域,随着大数据、计算机集群技术、网络技术和人工智能的发展,越来越多的金融机构将传统的策略风控手段转向依赖机器学习模型等量化手段。信贷环节中的审批、预警、催收以及营销等诸多场景也适合机器学习模型的应用。

十一月 15, 2020-十二月 31, 2020

1位参与者



## 足球运动员身价估计

竞赛组织者: 2020研究生机器学习助教

每个足球运动员在转会市场都有各自的价码。根据球员的各项信息和能力值可以预测该球员的市场价值。

十一月 15, 2020-十二月 31, 2020

1位参与者



## 规范口罩佩戴

Secret url: [https://localhost/competitions/74?secret\\_key=954ac57b-3341-464a-80e1-786bbb9ff7d6](https://localhost/competitions/74?secret_key=954ac57b-3341-464a-80e1-786bbb9ff7d6)

竞赛组织者: 2020研究生机器学习助教 - 当前服务器时间: 十一月 21, 2020, 2:51 p.m. 北京时间

### ► 最近的

Final test

十一月 15, 2020, 8 a.m. 北京时间

### 结束

竞赛结束

十二月 31, 2020, 午夜 北京时间

比赛细则

阶段

参与比赛

查看成绩

#### 规则综述

评分准则

比赛条款

### Welcome!

这是研究生机器学习课程的团队作业比赛项目之一: 规范口罩佩戴。你的任务是根据给出的训练数据集, 创建口罩佩戴分析的机器学习模型, 在测试集上预测人们正确佩戴了口罩的概率。参赛者需要严格按照测试集数据条目先后顺序, 给出正确佩戴了口罩的预测概率, 以submission.txt格式压缩成submission.zip文件后再提交。注意: 1.每个队除打榜之外, 最后还需要提交详细的项目技术报告和可复现代码。2.Deadline: 2020-12-31 24:00:00

比赛细则

阶段

参与比赛

查看成绩

你还没有报名参加这次竞赛

要参加这个比赛, 你必须接受它的具体条款和条件。报名后, 比赛主办方会评审你的申请并在你的申请被批准后通知你

☐ 我接受条款和条件。关于该比赛。

申请参加

获取数据

比赛附件

提交结果/查看提交

trainset有17万条记录，每条包括对每个申请贷款人采集的36条信息(Q1~Q36)，信息已经被编码为36维特征向量，贷款申请成功（label为1），贷款申请失败（label为0）。testset有3万条记录，请严格按照测试集数据条目先后顺序，根据自己训练的模型给出每条记录对应的贷款申请成功的预测概率，以submission.txt格式最终压缩成submission.zip提交。相应的数据集和提交文件样例下载链接：  
<https://bhpan.buaa.edu.cn:443/link/F2586DD5E1DCE38D0B3BE3436DA8E3FC>

获取数据

比赛附件

提交结果/查看提交

Final test

阶段说明  
无

每天最大提交数 100

总计最大提交数 10000

文档上传

选择文件 未选择任何文件

可以自愿在此上传一些文件信息

上传

点击提交按钮上传一个新的提交

可以自愿在此补充一些提交信息

不要需要操作的部分！！！！

提交

只需要点击此处提交按钮上传结果！！！！

这是目前为止你的提交 (✓ 下载排行榜上的所有提交):

#	分数	文件名	提交日期	状态	✓
表中没有可用的数据					

submission.txt

submission.zip



这是目前为止你的提交 (✔️ 下载排行榜上的所有提交):

#	分数	文件名	提交日期	状态	✔️
1	---	submission.zip	11/21/2020 15:08:00	Submitting	-

没有关于提交的解释。

更新解释

下载你的提交  
浏览评分器读入信息  
浏览评分器错误日志  
浏览预测的输出日志  
浏览预测的错误日志  
从预处理阶段下载你代码的运行结果  
下载最终的评分结果

分数同步到团队

Refresh status

#	分数	文件名	提交日期	状态	✔️
1	---	submission.zip	11/21/2020 15:08:00	Running	-

没有关于提交的解释。

更新解释

下载你的提交  
浏览评分器读入信息  
浏览评分器错误日志  
浏览预测的输出日志  
浏览预测的错误日志  
从预处理阶段下载你代码的运行结果  
下载最终的评分结果

分数同步到团队

Refresh status

#	分数	文件名	提交日期	状态	✔️
1	---	submission.zip	11/21/2020 15:08:00	Finished	-

没有关于提交的解释。

更新解释

下载你的提交  
浏览评分器读入信息  
浏览评分器错误日志  
浏览预测的输出日志  
浏览预测的错误日志  
从预处理阶段下载你代码的运行结果  
下载最终的评分结果

分数同步到团队

Submit to Leaderboard

#	分数	文件名	提交日期	状态	✔️
1	1.0	submission.zip	11/21/2020 15:08:00	Finished	-

Description:

更新解释

下载你的提交  
浏览评分器读入信息  
浏览评分器错误日志  
浏览预测的输出日志  
浏览预测的错误日志  
从预处理阶段下载你代码的运行结果  
下载最终的评分结果

分数同步到团队

Submit to Leaderboard

最终结果一定要提交到排行榜上！！！！

比赛细则 阶段 参与比赛 查看成绩

Final test

阶段说明  
无

每天最大提交数 100

总计最大提交数 10000

下载排行榜中参与者的成绩 下载排行榜上的所有提交 下载所有的参与者的成绩

所有参与者的成绩

#	用户名	团队名	团队成绩
1	2020研究生机器学习助教	-	-

Results

#	用户名	登录	上次登录日期	Score ▲
1	2020研究生机器学习助教	1	11/21/20	1.000000 (1)


成绩+排名

比赛细则

阶段

参与比赛

查看成绩

团队 

规则综述

评分准则

比赛条款

## Welcome!

这是研究生机器学习课程的团队作业比赛项目之一：规范口罩佩戴。你的任务是根据给出的训练数据集，创建口罩佩戴分析的机器学习模型，在测试集上预测人们正确佩戴了口罩的概率。参赛者需要严格按照测试集数据条目先后顺序，给出正确佩戴了口罩的预测概率，以submission.txt格式压缩成submission.zip文件后再提交。注意：1.每个队除打榜之外，最后还需要提交详细的项目技术报告和可复现代码。2.Deadline：2020-12-31 24:00:00

创建新的团队

### 来自其他团队的邀请

团队	消息内容:	日期	操作
----	-------	----	----

### 给其他团队的请求

团队	消息内容:	日期	操作
----	-------	----	----

## 团队信息

Name:

Description:

Allow requests:

☐

Image:

未选择任何文件

# 作业提交时间节点

- 2020.12.31 24:00 比赛平台关闭!
- 2020.12.31 24:00 (技术报告+可复现代码)提交截止!

队长负责收齐和提交队员的个人作业和团队作业。

邮件主题：队长学号\_姓名\_作业提交 → [zyaml2020@163.com](mailto:zyaml2020@163.com)

压缩包文件命名：

队长学号\_姓名\_团队作业\_团队作业名称.zip

个人学号\_姓名\_个人作业\_个人作业名称.zip