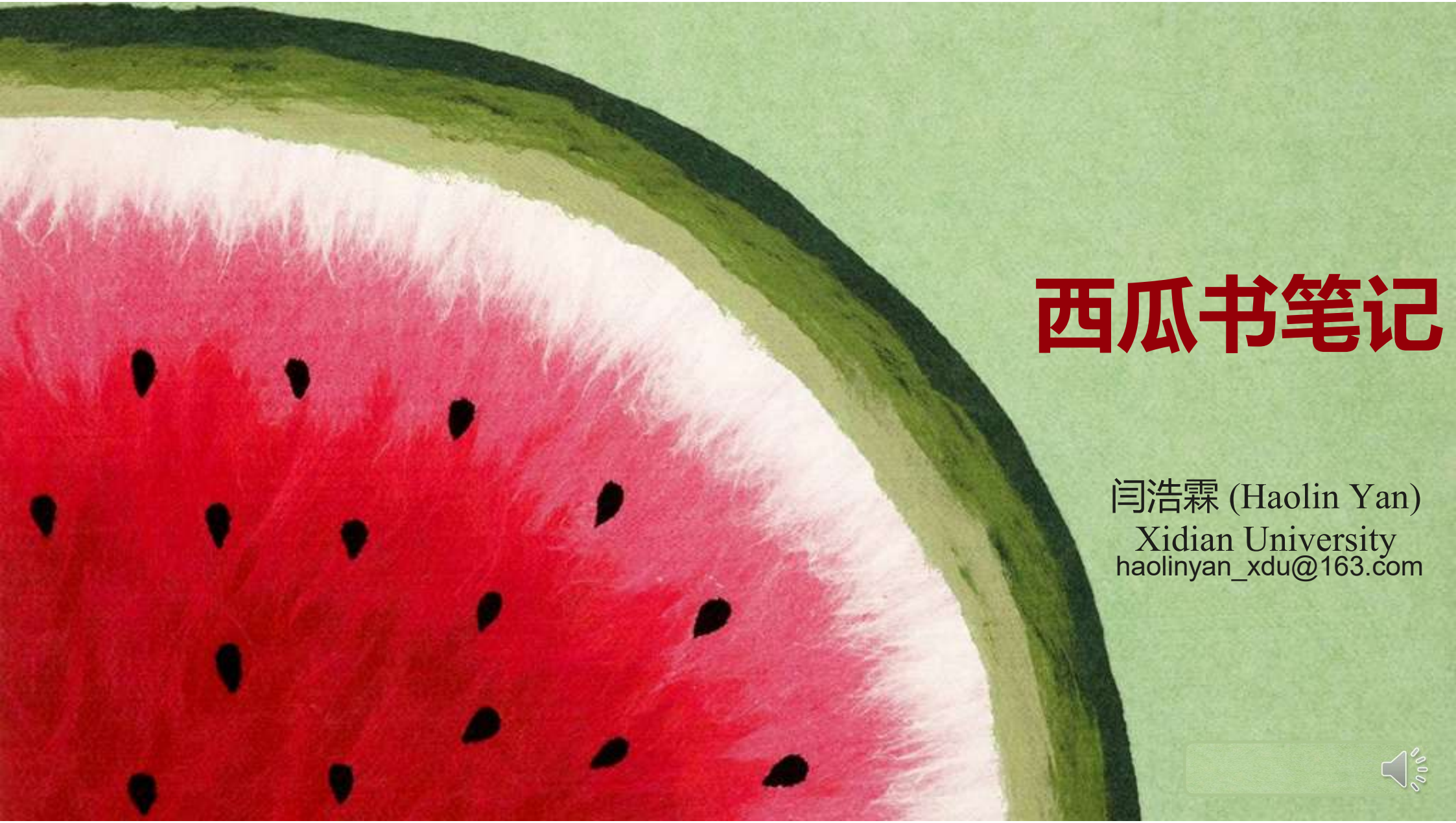
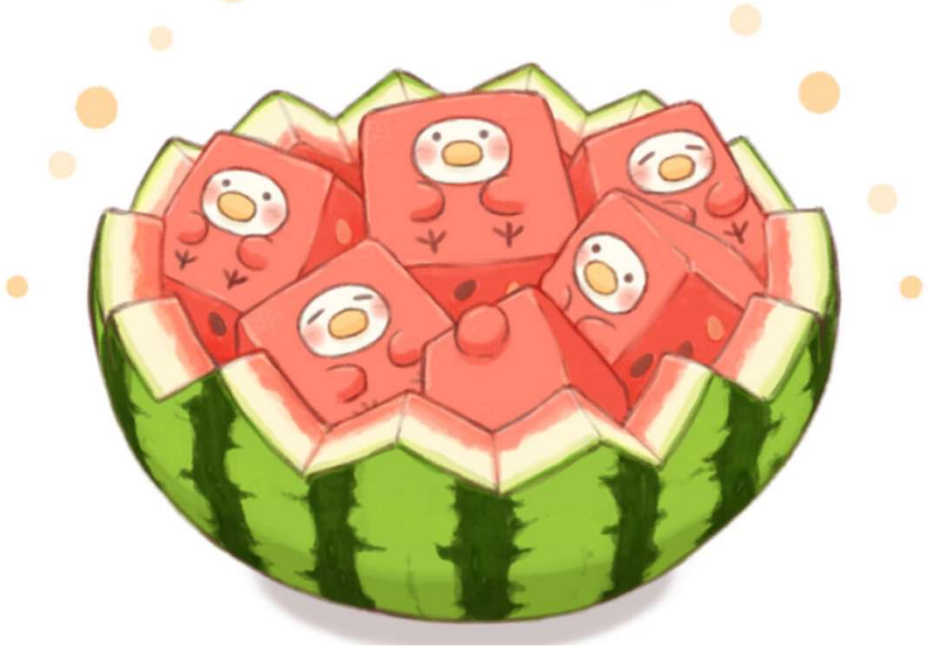


西瓜书笔记

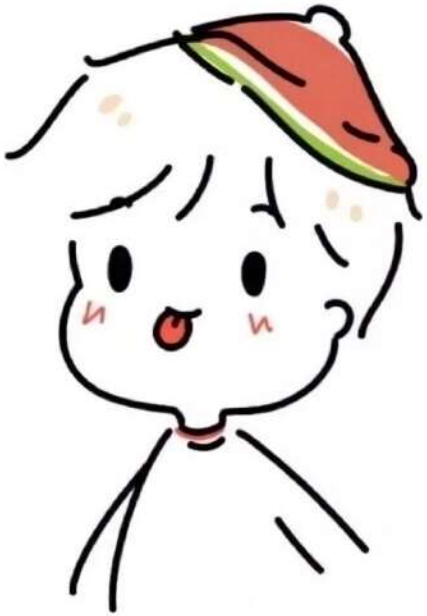
闫浩霖 (Haolin Yan)
Xidian University
haolinyan_xdu@163.com

西瓜书读书笔记

机器学习



关于吃瓜的我...



闫浩霖

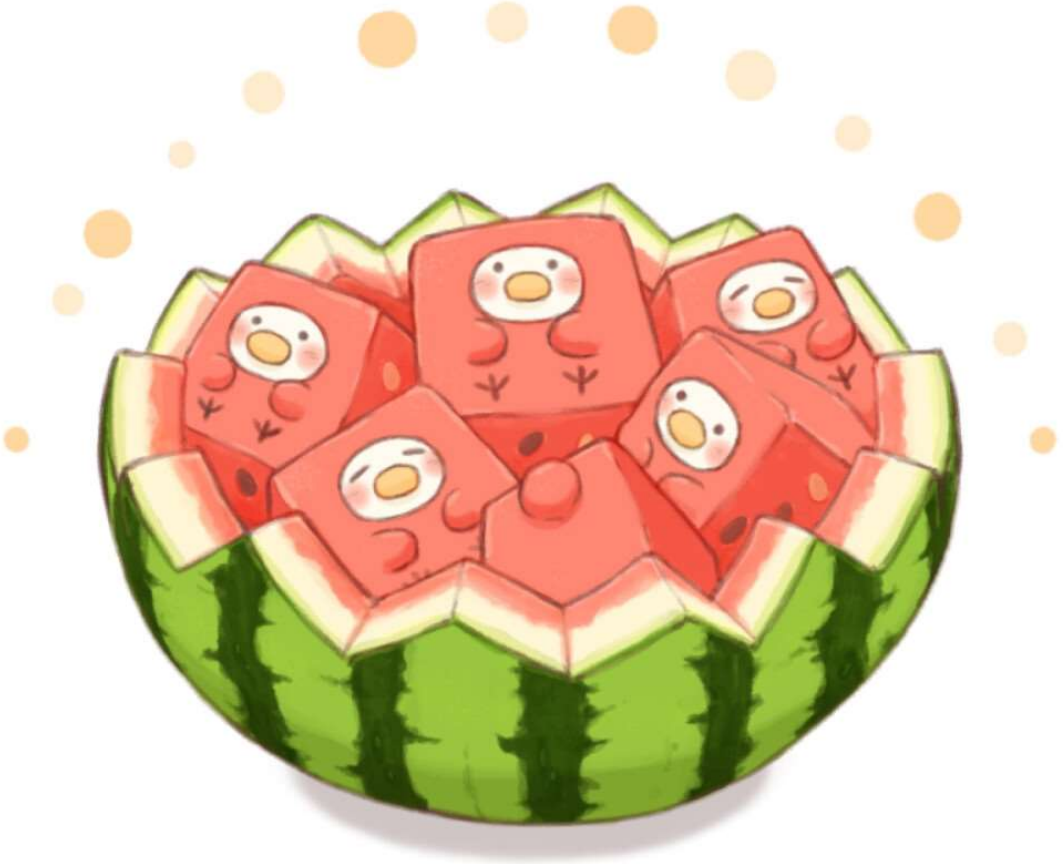
Haolin Yan

西安电子科技大学 人工智能学院

haolinyan_xdu@163.com

擅长技能:

- 深度学习基础
- pytorch/Paddle
- Object Detection
- Relation Extraction
- Image caption



第一章 绪论
Chapter 1 Intro

机器学习

机器学习是指算法根据**数据**产生**模型**，模型可以完成对数据的**分析与理解**，机器学习就是在研究这样的**学习算法**

序
能，
验
性
于
行

在，因此，机器学习所研究的主要内容，是关于在计算机上从数据中产生“模型”(model) 的算法，即“学习算法”(learning algorithm)．有了学习算法，我们把经验数据提供给它，它就能基于这些数据产生模型；在面对新的情况时(例如看到一个没剖开的西瓜)，模型会给我们提供相应的判断(例如好瓜)．如果说计算机科学是研究关于“算法”的学问，那么类似的，可以说机器学习是研究关于“学习算法”的学问．

基本术语

样本(sample)
特征(feature)
特征向量(feature vector)
特征空间(feature space)
数据集(dataset)
训练集(training set)
测试集(test set)
泛化(generation)
独立同分布(independent identically distribute)
有监督学习与无监督(supervised learning and unsupervised)
分类(classification) 回归(regression)
聚类(clustering)
归纳偏好(inductive bias)

专业人士要说内行话！

No Free Lunch

为简单起见, 假设样本空间 $\mathcal{X}$ 和假设空间 $\mathcal{H}$ 都是离散的. 令 $P(h|X, \mathfrak{L}_a)$ 代表算法 $\mathfrak{L}_a$ 基于训练数据 X 产生假设 h 的概率, 再令 f 代表我们希望学习的真实目标函数. $\mathfrak{L}_a$ 的“训练集外误差”, 即 $\mathfrak{L}_a$ 在训练集之外的所有样本上的误差为

$$E_{ote}(\mathfrak{L}_a|X, f) = \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\, \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))\, P(h \mid X, \mathfrak{L}_a)\ , \qquad (1.1)$$

其中 $\mathbb{I}(\cdot)$ 是指示函数, 若 $\cdot$ 为真则取值 1, 否则取值 0.

No Free Lunch

考虑二分类问题, 且真实目标函数可以是任何函数 $\mathcal{X} \mapsto \{0,1\}$, 函数空间为 $\{0,1\}^{|\mathcal{X}|}$. 对所有可能的 f 按均匀分布对误差求和, 有

$$
\begin{aligned}
\sum_f E_{ote}(\mathfrak{L}_a|X,f) &= \sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\ \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))\ P(h \mid X, \mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a) \frac{1}{2} 2^{|\mathcal{X}|} \\
&= \frac{1}{2} 2^{|\mathcal{X}|} \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h \mid X, \mathfrak{L}_a)
\end{aligned}
$$

No Free Lunch

[解析]：第 1 步到第 2 步：

$$\begin{aligned}
&\sum_f \sum_h \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x})\mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))P(h|X,\mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_f \sum_h \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))P(h|X,\mathfrak{L}_a) \\
&= \sum_{\boldsymbol{x}\in\mathcal{X}-X} P(\boldsymbol{x}) \sum_h P(h|X,\mathfrak{L}_a) \sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x}))
\end{aligned}$$

第 2 步到第 3 步：首先要知道此时我们对 f 的假设是任何能将样本映射到 0,1 的函数且服从均匀分布，也就是说不止一个 f 且每个 f 出现的概率相等，例如样本空间只有两个样本时：$\mathcal{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2\}, |\mathcal{X}| = 2$，那么所有的真实目标函数 f 为：

$$\begin{aligned}
f_1 &: f_1(\boldsymbol{x}_1) = 0, f_1(\boldsymbol{x}_2) = 0; \\
f_2 &: f_2(\boldsymbol{x}_1) = 0, f_2(\boldsymbol{x}_2) = 1; \\
f_3 &: f_3(\boldsymbol{x}_1) = 1, f_3(\boldsymbol{x}_2) = 0; \\
f_4 &: f_4(\boldsymbol{x}_1) = 1, f_4(\boldsymbol{x}_2) = 1;
\end{aligned}$$

一共 $2^{|\mathcal{X}|} = 2^2 = 4$ 个真实目标函数。所以此时通过算法 $\mathfrak{L}_a$ 学习出来的模型 $h(\boldsymbol{x})$ 对每个样本无论预测值为 0 还是 1 必然有一半的 f 与之预测值相等，例如，现在学出来的模型 $h(\boldsymbol{x})$ 对 $\boldsymbol{x}_1$ 的预测值为 1，也即 $h(\boldsymbol{x}_1) = 1$，那么有且只有 f_3 和 f_4 与 $h(\boldsymbol{x})$ 的预测值相等，也就是有且只有一半的 f 与它预测值相等，所以 $\sum_f \mathbb{I}(h(\boldsymbol{x}) \neq f(\boldsymbol{x})) = \frac{1}{2}2^{|\mathcal{X}|}$；第 3 步一直到最后显然成立。值得一提的是，在这

Homework

表 1.1　西瓜数据集

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	浊响	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

1.C(3,1) + C(3,2) + C(3,3) = 7
好瓜->(色泽=*；根蒂=*；浊响=*)

2.单个合取式的假设空间为3*4*4+1=49
对于析合范式假设空间：C(49,k)

1.1　表 1.1 中若只包含编号为 1 和 4 的两个样例, 试给出相应的版本空间.

1.2　与使用单个合取式来进行假设表示相比, 使用“析合范式”将使得假设空间具有更强的表示能力. 例如

$$\text{好瓜} \leftrightarrow \Big((\text{色泽}=*) \wedge (\text{根蒂}=\text{蜷缩}) \wedge (\text{敲声}=*)\Big) \vee \Big((\text{色泽}=\text{乌黑}) \wedge (\text{根蒂}=*) \wedge (\text{敲声}=\text{沉闷})\Big),$$

会把“(色泽=青绿) ∧ (根蒂=蜷缩) ∧ (敲声=清脆)”以及“(色泽=乌黑) ∧ (根蒂=硬挺) ∧ (敲声=沉闷)”都分类为“好瓜”. 若使用最多包含 k 个合取式的析合范式来表达表 1.1 西瓜分类问题的假设空间, 试估算共有多少种可能的假设.

Homework

即不存在训练错误为 0 的假设.

1.3 若数据包含噪声, 则假设空间中有可能不存在与所有训练样本都一致的假设. 在此情形下, 试设计一种归纳偏好用于假设选择.

1.4* 本章 1.4 节在论述“没有免费的午餐”定理时, 默认使用了“分类错误率”作为性能度量来对分类器进行评估. 若换用其他性能度量 ℓ, 则式(1.1)将改为

$$E_{ote}(\mathfrak{L}_a|X,f)=\sum_{h}\sum_{\boldsymbol{x}\in\mathcal{X}-X}P(\boldsymbol{x})\ell\left(h\left(\boldsymbol{x}\right),f\left(\boldsymbol{x}\right)\right)P\left(h\mid X,\mathfrak{L}_a\right)\ ,$$

试证明“没有免费的午餐定理”仍成立.

Homework

证明：

假设性能度量值为有限离散值，离散值空间大小为N,数据集的大小为$|x|$,因为目标函数是均匀分布的，因此可以假设存在$N^{|x|}$种f，在给定的$h(x)$下对于每一个度量值l_i都有$N^{|x|-1}$种对应的函数，因此原公式可证明：

公式 </>

$$\sum_f E(\phi|X,f) = \sum_f \sum_h \sum_x P(x) l(h(x), f(x)) P(h|X,\phi)$$

$$\sum_f E(\phi|X,f) = \sum_x \sum_h P(x) P(h|X,\phi) \sum_f l(h(x), f(x))$$

$$\sum_f E(\phi|X,f) = \sum_x \sum_h P(x) P(h|X,\phi) N^{|x|-1} \sum_i^N l_i$$

$$\sum_f E(\phi|X,f) = N^{|x|-1} \sum_i^N l_i \sum_x P(x) \sum_h P(h|X,\phi)$$

$$\sum_f E(\phi|X,f) = N^{|x|-1} \sum_i^N l_i \sum_x \sum_h P(x)$$

总误差与学习算法无关,NFl证毕。

感谢观看！

Thank you.

为成为国家新时代人工智能人才而努力！为新时代人工智能发展做贡献！勇于探索，积极进取，使命必达！

Strive to become a national artificial intelligence talent in the new era! Contribute to the development of artificial intelligence in the new era! Dare to explore and forge ahead, and the mission will be achieved!

西安电子科技大学 人工智能学院2019级学生
闫浩霖
2021 年 11 月