公告

- 答疑
  - 时间：周二晚8-9PM
  - 地点：中央主楼716（进去，左手房间）
- 课件、作业和阅读材料见网络学堂
- 请在网络学堂讨论区张贴每节课相关问题、评论。

人工智能基础算法第二节
K最近邻分类器

于 国 强
清 华 大 学
2025 年 9 月 23日

本节课的安排

- 人工智能分类
- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- 训练与测试
- 以计算代替知识

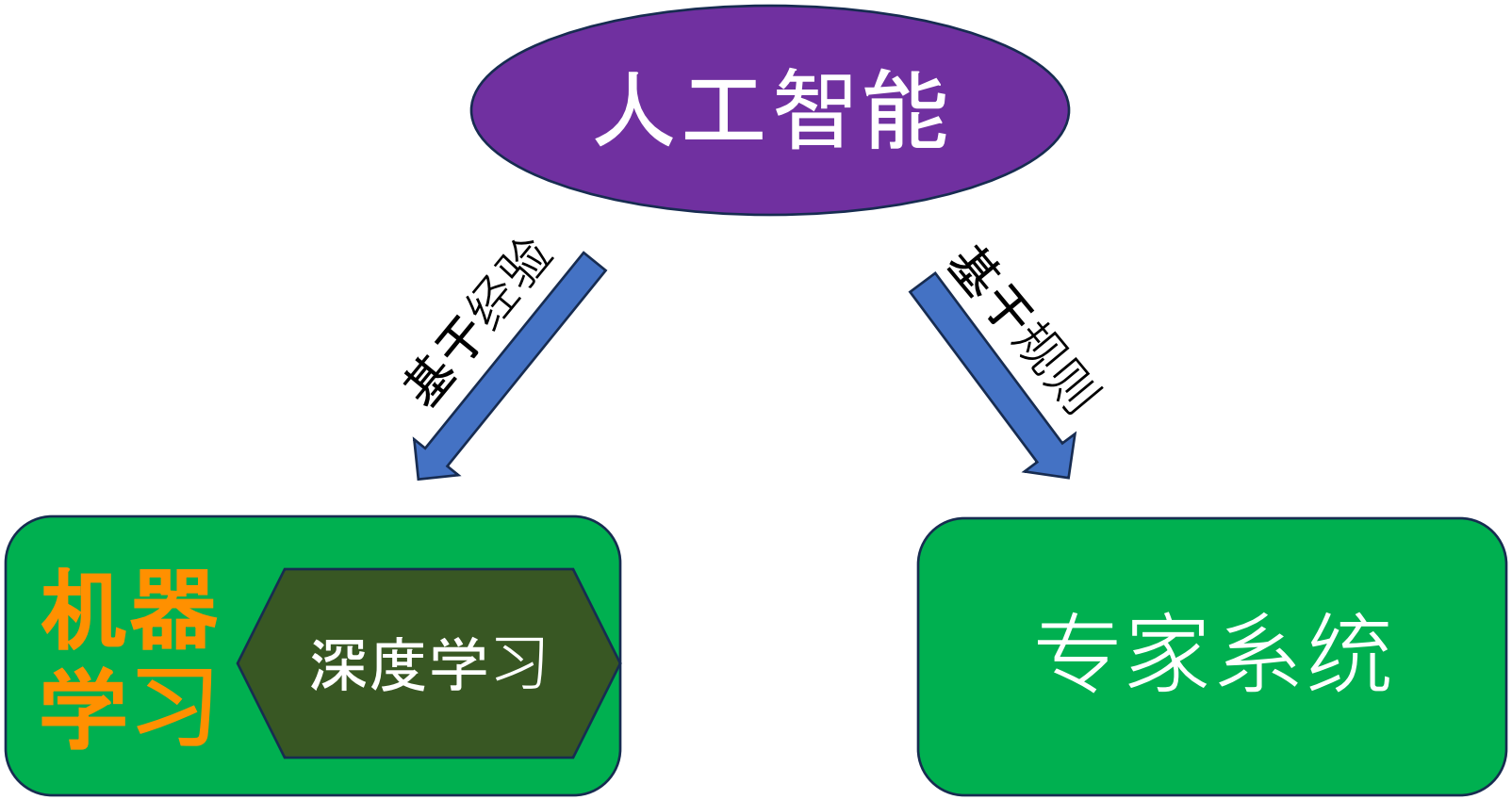
本节课的安排

- **人工智能分类**
- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- 训练与测试
- 以计算代替知识

什么是智能和人工智能？

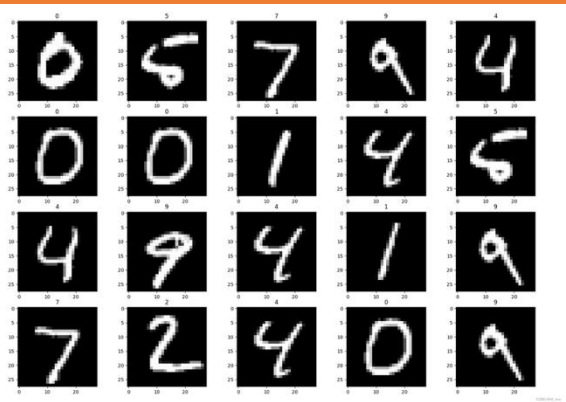
- 智能目前没有统一的定义
  - 有人认为智能是人所具有的独特特征
- 人工智能的定义似乎更容易些，但也不统一
  - 类人的行为
  - 类人的思考
- 类人的行为，也即图灵测试，这一角度为更多人接受
- 悖论是：一件事情，一旦被人彻底理解，就属于科学，而不是智能了
  - 所以，有人说人工智能可以无限接近，但永远无法达到。

与人工智能相关的几个概念

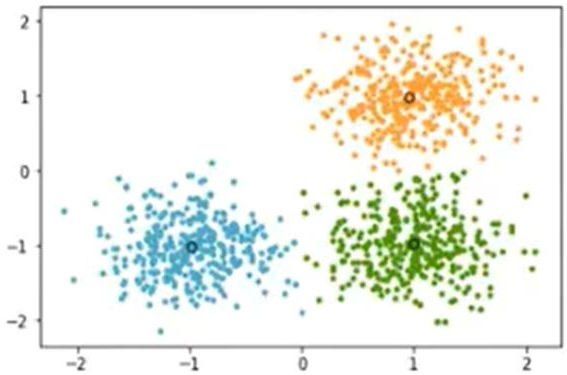


人工智能算法分类

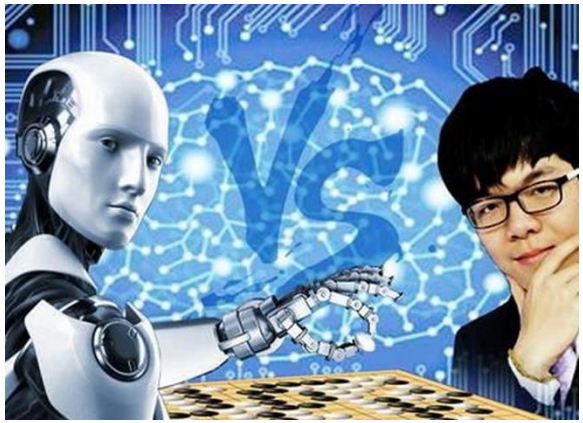
- 有监督学习
  - 有标签



- 无监督学习
  - 无标签



- 强化学习
  - 目标是动态的



- 生成式学习
  - 产生符合给定要求的样本

本节课的安排

- 人工智能分类
- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- 训练与测试
- 以计算代替知识

引言

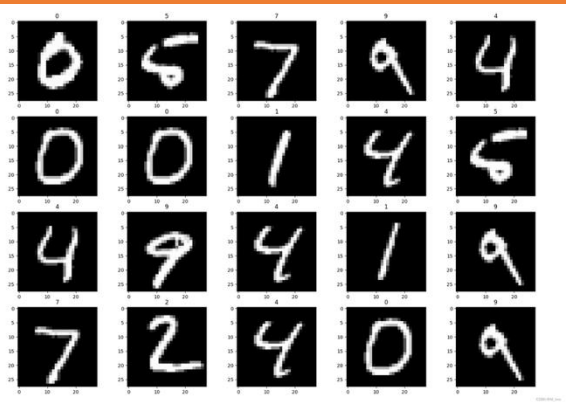
- 学习目的:
  - 学习最近邻分类器这一经典算法
    - 了解概念
    - 熟悉技术细节
    - 理论分析
  - 学习人工智能里若干重要概念
    - 分类器
    - 性能评价
    - 训练与测试
    - 以计算代替知识

本节课的安排

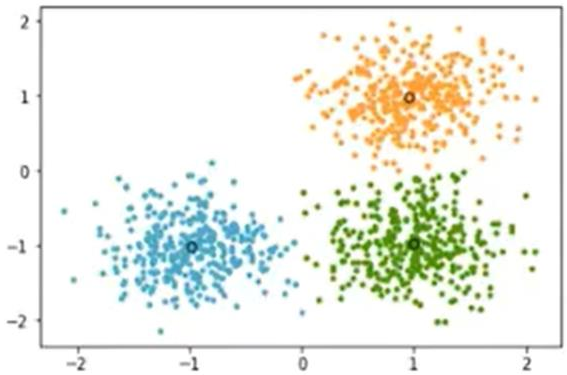
- 引言
- **分类器**
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- 训练与测试
- 以计算代替知识

人工智能算法分类

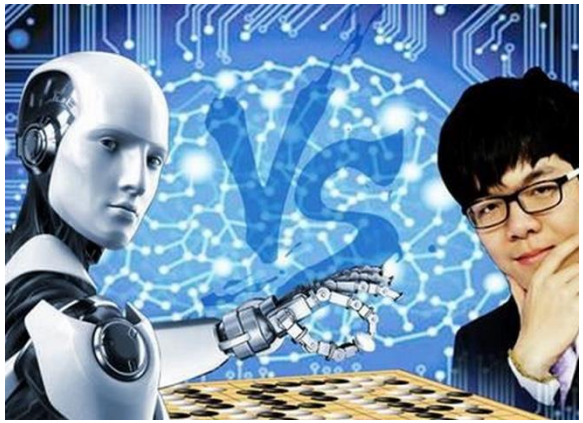
- **有监督学习**
  - 有标签



- 无监督学习
  - 无标签



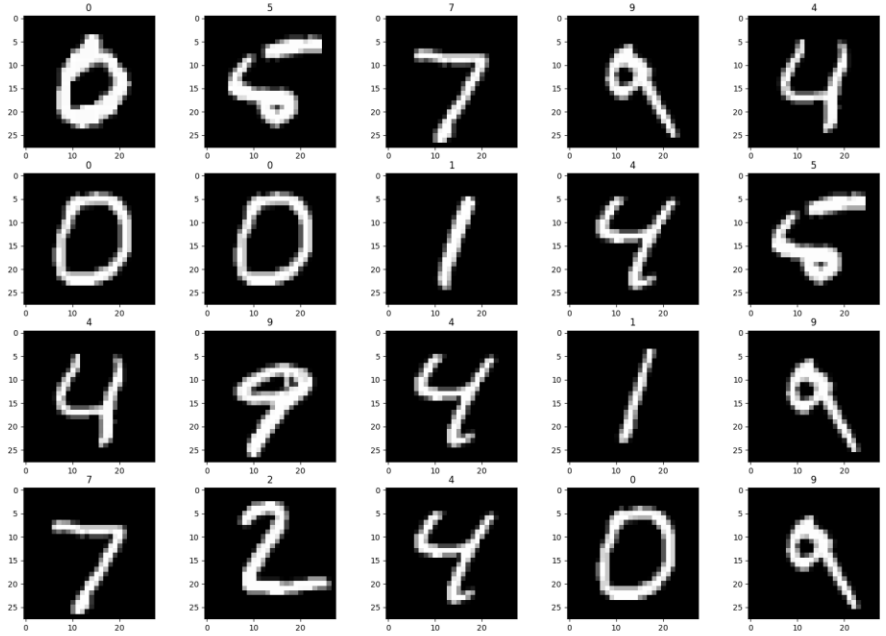
- 强化学习
  - 目标是动态的



- 生成式学习
  - 产生符合给定要求的样本

分类器的例子

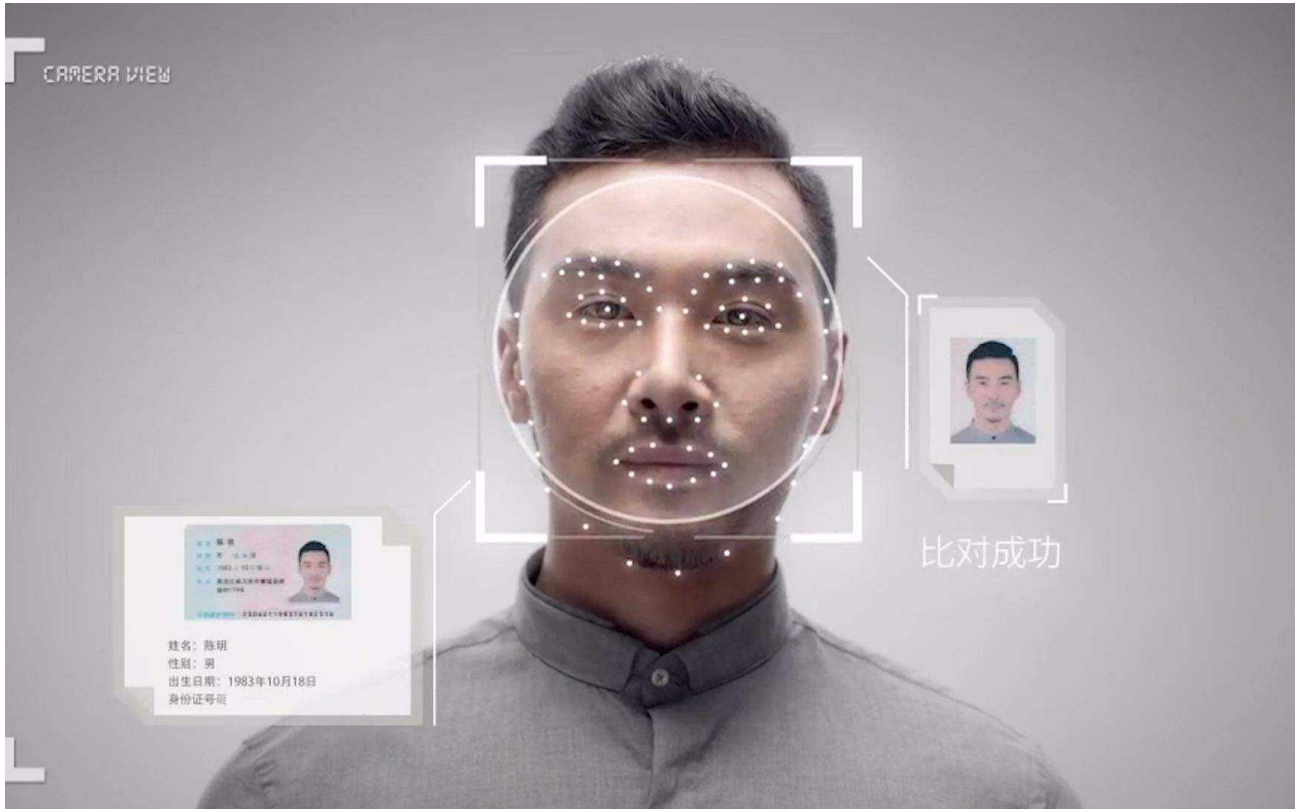
- 构建分类器是一种有监督学习方式
  - 分类器一般称为xx识别
- 手写数字识别
  - 给一张图，判断它的分类（0，1，2，…, 9)



语音识别

- 把一段语音识别成对应的文字

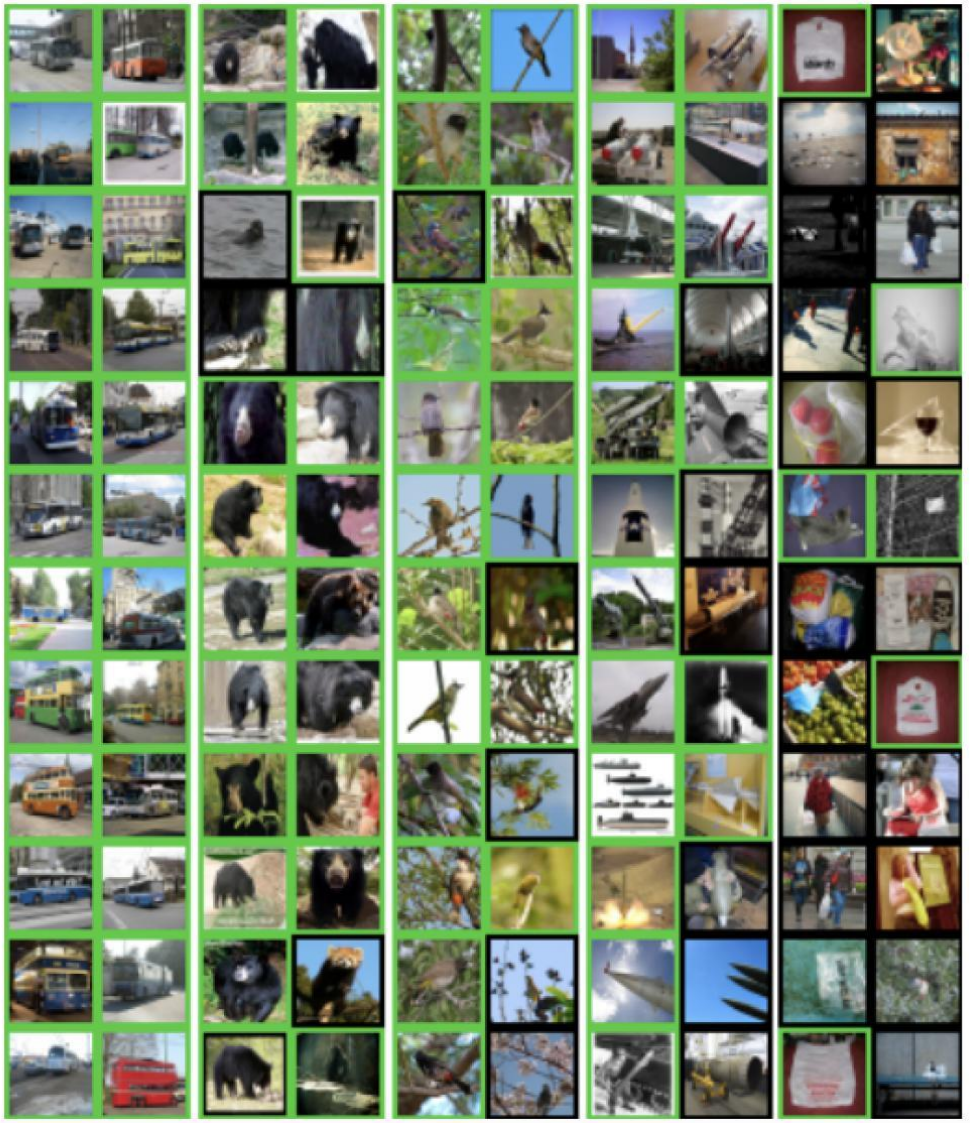
人脸识别



- 根据图像，找到数据库里存储的人的信息

图像识别

- 识别图像里包含什么。
  - ImageNet数据集里有1500万张图片，2万2千个类别。



关于分类器或类别识别

- 定义
  - 给定输入特征(Feature Variable)，输出类别或标签
  - 类别一般是离散的，而且类别之间没有直接的数量关系。
  - 譬如说，我们不能说类别A比类别B比大。
- 类别个数
  - 二分类，多分类
  - 例子？
- 属于有监督学习，但不是识别问题
  - 例子？

识别问题对人工智能发展的影响

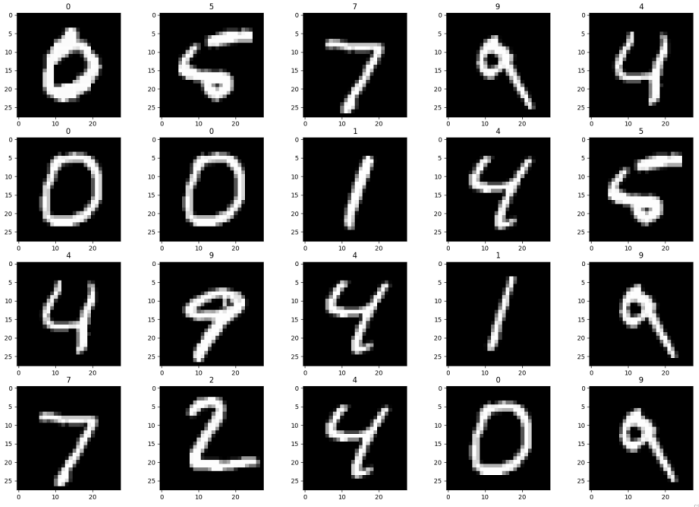
- 识别问题是人工智能技术的验金石
  - 新技术先要拿数字识别，语音识别，人脸识别，图像识别，试一试
- 人工智能发展史上的标志性进步
  - 2010年，深度学习技术极大的提高了语音识别的准确率
  - 2012年，深度学习技术在ImageNet上获得巨大的成功
  - 2014年，机器人脸识别能力突破人眼水平
  - 语音识别和人脸识别已融入日常生活中

本节课的安排

- 引言
- 分类器
- **最近邻分类器**
- K最近邻分类器
- 分类器的评价
- 训练与测试
- 以计算代替知识

最近邻分类器

- 最近邻分类器 （Nearest Neighbor Algorithm）
  - 给定一组带有类别标签的样本，对每一个未知标签的样本，找到最近的样本，然后把对应样本的标签赋给该未知样本。

- 例子：给定图像  和 对应标签

0	5	7	9	4
0	0	1	4	5
4	9	4	1	9
7	2	4	0	9

- 求未知样本  的类别

最近邻分类器的直觉分析

- 直觉 -> 理性分析 ->直觉
- 最近邻分类器的效果应该不错
  - 因为，相似的样本应该拥有相似的标签
- 暗含的假设：概率分布的连续性
  - If $\mathrm{x} \to \mathrm{x}_0$, then $P(c_i|\mathrm{x}) \to P(c_i|\mathrm{x}_0)$
- 理论分析待后

最近邻分类器的技术细节

- 给定一组带有类别标签的样本，对每一个未知标签的样本，找到最近的样本，然后把对应样本的标签赋给该未知样本。
  - 针对这个定义，在实际使用时，还有什么是需要使用者确定的呢？

最近邻分类器的技术细节（1）

- 给定一组带有类别标签的样本，对每一个未知标签的样本，找到**最近**的样本，然后把对应样本的标签赋给该未知样本。
  - 最近是指，“距离最近”。

最近邻分类器的技术细节（1）

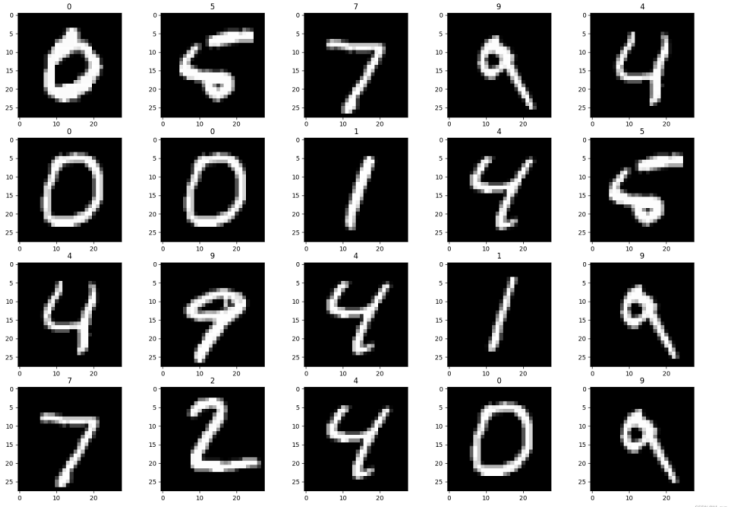
- 给定两个样本，x_1 和 x_2，记$\mathrm{v} = \mathrm{x}_1 - \mathrm{x}_2$，两者之间的距离可计算为，
  - 欧几里得距离：$\sqrt{\mathrm{v}^T\mathrm{v}} = \sqrt{\sum_i v_i^2}$
  - 曼哈顿距离：$|\mathrm{v}| = \sum_i |v_i|$
  - L_∞范数距离：$\max_i |v_i|$
  - 普遍形式：闵可夫斯基距离距离$\left(\sum_i v_i^p\right)^{\frac{1}{p}}$
  - **其他距离**?

最近邻分类器的技术细节（2）

- 最近邻分类器的计算量是巨大的
  - 假设你的数据库里有N个样本，你需要进行N次比对。
- 最近邻分类器的快速算法
  - 去掉冗余的样本
  - 如果一个样本的周围都是同一个标签，那么这个样本可以去掉。

最近邻分类器的技术细节（3）

- 变量和特征 （Feature Variable）
  - 我们并不一定基于直接给定的特征
  - 可以进行特征选择，特征变换，特征提取
  - 现在人工智能的重要方向是learn representation
- 手写数字识别示例
  - 特征选择
  - 特征变换
  - 表征学习?



最近邻分类器的直觉分析

- 直觉 -> 理性分析 ->直觉

- 最近邻分类器的效果应该不错
  - 因为，相似的样本应该拥有相似的标签

- 暗含的假设：概率分布的连续性
  - If $\mathrm{x} \to \mathrm{x}_0$, then $P(c_i|\mathrm{x}) \to P(c_i|\mathrm{x}_0)$

- **理论分析待后**

最近邻分类器的理论分析

- 贝叶斯概率
  - 也叫，后验概率，或，贝叶斯后验概率
  - 起因于贝叶斯条件概率理论
    - $P(c_i|\mathrm{x}) = \frac{P(c_i,\mathrm{x})}{P(\mathrm{x})} = \frac{P(\mathrm{x}|c_i)P(c_i)}{\sum_i P(\mathrm{x}|c_i)P(c_i)}$
- 贝叶斯决策
  - 给定样本x，我们选取c_k，使得$P(c_k|\mathrm{x}) \geq P(c_i|\mathrm{x})$, for $i = 1, \ldots K, \text{but } i \neq k,$。
  - 很显然，贝叶斯决策是错误率最小的决策。
  - 贝叶斯错误率$\leq \frac{c-1}{c}$

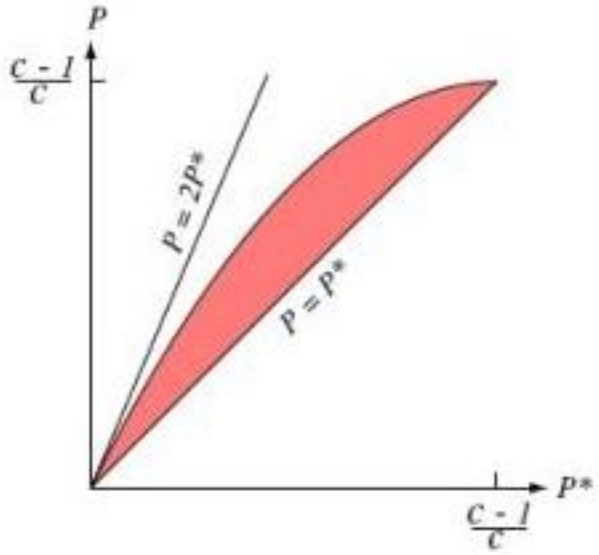
最近邻分类器的理论分析

- 最近邻分类器的错误率
  - ≥贝叶斯错误率
  - 因为贝叶斯错误率是最优的

- 最近邻分类器的错误率
  - ≤2倍贝叶斯错误率
  - 基于样本数无穷多的假设

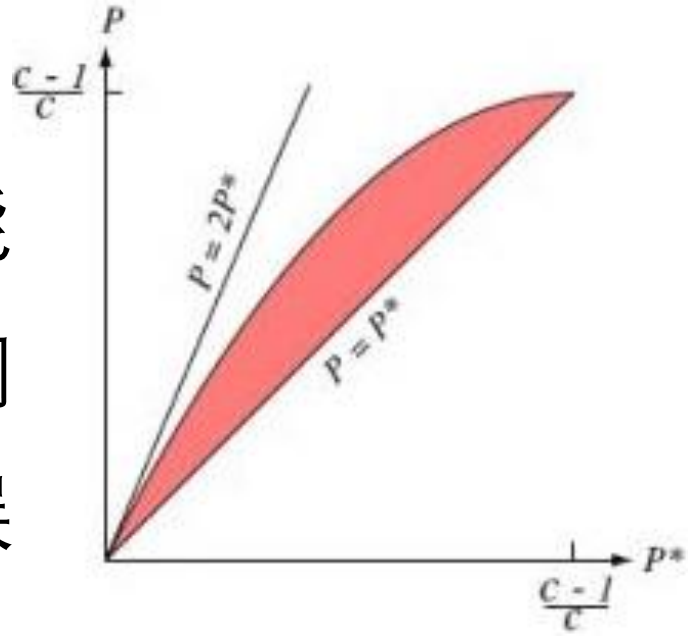
最近邻分类器的理论分析

- 贝叶斯错误率记为$P^*(e|\mathrm{x}) = 1 - \max_i P(c_i|\mathrm{x})$
- 最近邻分类器的错误率$P(e|\mathrm{x})$
  - 在样本数无穷多的假设下，
  - $P(e|\mathrm{x}) = 1 - \sum_i P^2(c_i|\mathrm{x}) \leq 1 - \left(\max_i P(c_i|\mathrm{x})\right)^2 \leq 2P^*(e|\mathrm{x})$
  - 更详细的分析可以得出，
  - $P(e|\mathrm{x}) \leq P^*(e|\mathrm{x})\left(2 - \frac{c}{c-1}P^*(e|\mathrm{x})\right)$



最近邻分类器的理论分析

- 该理论基于无穷多样本的假设。
  - 现实中样本总是有限的。
- 如果我们有无穷多样本，和能设计无限准确的分类器，我们最多能把最近邻分类器的错误率降低一半。
  - 一半的信息蕴含在最近邻中。

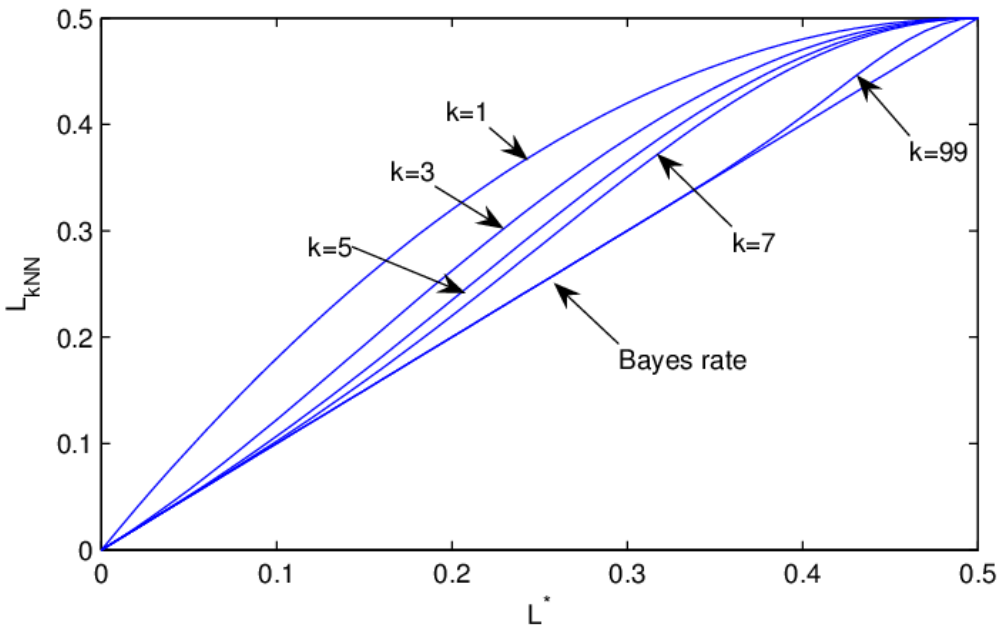


本节课的安排

- 引言
- 分类器
- 最近邻分类器
- **K最近邻分类器**
- 分类器的评价
- 训练与测试
- 以计算代替知识

K 近邻分类器

- K 近邻分类器 （KNN – K Nearest Neighbor Algorithm）
  - 给定一组带有类别标签的样本，对每一个未知标签的样本，找到最近的K个样本，然后把这些样本中最多的标签赋给该未知样本。
- K 近邻分类器的错误率
  - 假设样本数无穷多，
  - 如右图所示。



本节课的安排

- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- **分类器的评价**
- 训练与测试
- 以计算代替知识

分类器的评价

- 对于任何人工智能技术，如何评价它都是一个重要的问题
  - 评价它在实际应用中的表现
  - 根据评价效果设定重要的参数

- 评价指标往往不唯一

分类器的评价

- 对于最近邻或KNN分类器
  - 理论错误率是基于样本无穷多假设
  - 需要选取重要的参数
    - 哪种距离?
    - K的数值
- 最近邻或KNN分类器的评价指标
  - 错误率
  - 计算速度
  - 内存消耗

分类器的评价

- 对于最近邻或KNN分类器
  - 理论错误率是基于样本无穷多假设
  - 需要选取重要的参数
    - 哪种距离?
    - K的数值
- 最近邻或KNN分类器的评价指标
  - 错误率（**你能想到什么方案呢?**）
  - 计算速度
  - 内存消耗

分类器的评价

- 交叉检验评价分类器的错误率
  - 把数据集分为两部分
    - 一部分作为训练集，另一部分作为测试集
    - 比例设为多少？50% vs 50%？
  - 有没有方法让训练集大，测试集也大呢？

分类器的评价

- K 折交叉检验 （K Fold Cross Validation)
  - 把数据集分为同等大小的K份
    - 针对K份，分别测试
    - 每次把另外K-1份合起来作为训练集
    - 最终把这K次测试的结果合起来

- 问题：这种方案利用了多少样本作为训练集，多少样本作为测试集？
- 问题：这种方案的缺点是什么？

分类器的评价

- 留一法交叉检验 （LOOCV - Leave One Out Cross Validation)
  - 利用了多少样本训练，多少样本测试?
  - K折法的一个特例
  - 为什么?

本节课的安排

- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- **训练与测试**
- 以计算代替知识

训练与测试

- 训练与测试是人工智能中非常重要的两个概念
  - 需要**训练**是因为现在的人工智能大多数基于从经验中学习规律
    - 有些方法的训练是隐式的(implicit), 如最近邻方法
    - 有些非机器学习的人工智能方法不需要训练
      - 或者说训练是由创造者完成的
  - 需要**测试**是因为需要评估方法的优劣或进行技术的开发与优化
    - 严格的测试应该有第三方独立完成
    - 便捷的方式是使用交叉检验或其他直接基于手头上数据的方式

训练时间与使用时间

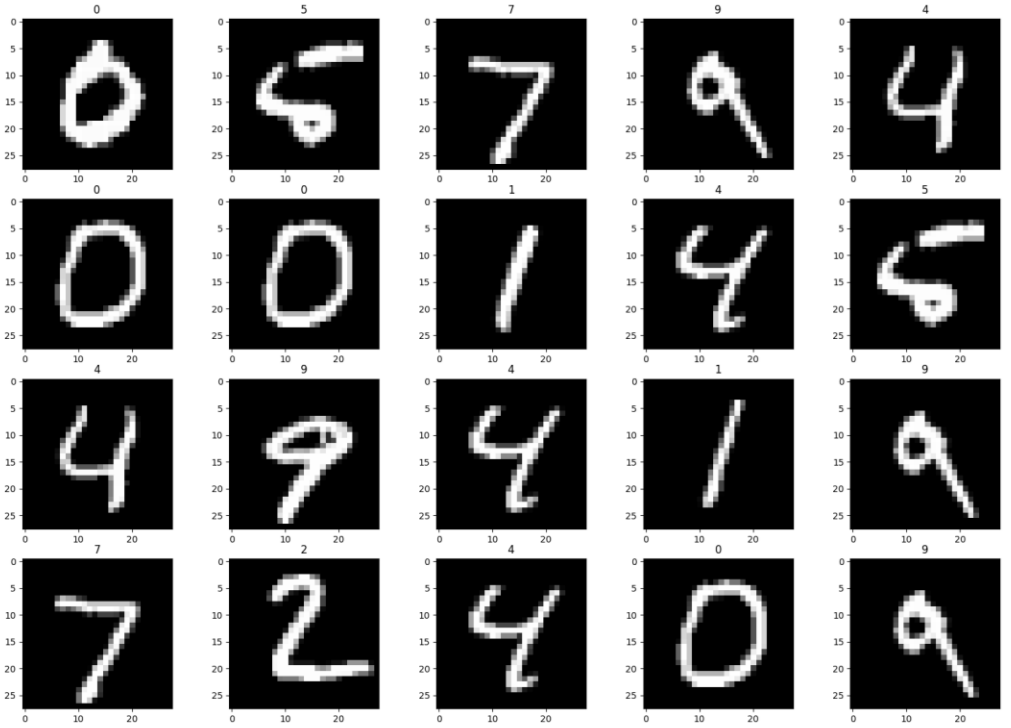
- 训练时间
  - 有时也叫学习时间 （Learning time）
- 使用时间
  - 有时也叫推断时间 （Inference time）
- 训练时间与使用时间是一组平衡
  - 一般说来训练时间越短，使用的耗时就越长
    - KNN没有训练时间，但使用时耗时很长
  - 训练时间越长，使用的耗时就越短
    - 对KNN的数据集进行预处理，可以减少使用时的时间
    - 基于深度神经网络的训练时间长，使用时耗时短
  - 关键在于找到一个较佳的平衡点，可能依赖于问题本身

本节课的安排

- 引言
- 分类器
- 最近邻分类器
- K最近邻分类器
- 分类器的评价
- 训练与测试
- **以计算代替知识**

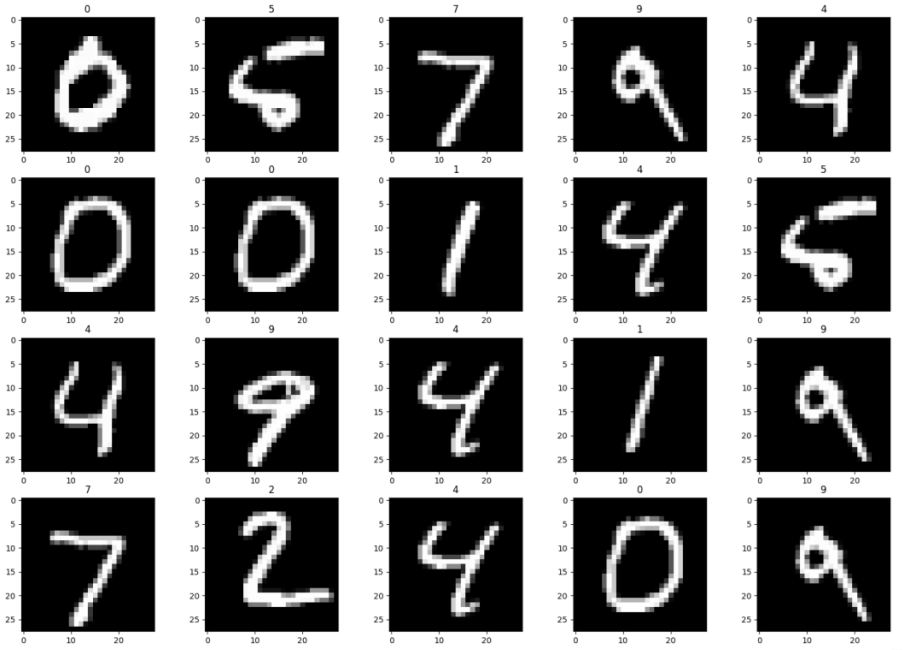
以计算代替知识

- 最近邻方法直接利用了数据
- 对于许多问题，我们有相关知识，充分利用起来有望更好的解决问题。
- 如何利用知识呢？



以计算代替知识

- 许多新方法是基于问题独特的知识或属性
  - 非常具有创新性
  - 费时间，费脑子
- 根据知识创造新样本，然后让机器去学习。
  - 譬如，对数字放缩，旋转，平移，加噪声，然后用KNN
  - 简单，粗暴，但需要大算力



Data Augmentation
数据增强

KNN算法的常见应用

- 没有
  - 没有人会说自己的技术基于KNN，因为它太简单了
- 为什么还要学习KNN算法?

总 结

- 学习目的:
  - 学习最近邻分类器这一经典算法
    - 了解概念
    - 熟悉技术细节
    - 理论分析
  - 学习人工智能里若干重要概念
    - 分类器
    - 性能评价
    - 训练与测试
    - 以计算代替知识

下一节课内容

- 线性有监督学习方法
  - 线性回归
  - 逻辑回归
  - 带约束的线性方法?

下一节课内容

- 线性有监督学习方法
  - 线性回归
  - 逻辑回归
  - 带约束的线性方法?

提 醒

- 如果你觉得不适合这门课，请抓紧退课。
  - 我们还有很多同学因为课容量，无法选这门课。

Thank you!