深度学习及典型病例分析

## 机器学习简介

###### 什么是人工智能

（1）机器学习是一种实现人工智能的方法从数据中心寻找规律、建立关系，根据建立的关系去解决问题。

（2）应用场景：数据挖掘、计算机视觉、自然语言处理、证券分析、

机器人…….

（3）目前机器学习是实现人工智能的主流方法

图示

描述已自动生成

###### 基本框架

将训练数据喂给计算机，计算机自动求解数据关系，在新的数据上做出预测以及建议。

###### 主要类别

* 1. 监督学习

——训练数据包括正确的结果（标签-label）

图示

描述已自动生成

1. 无监督学习

——训练数据不包括正确结果

图示

描述已自动生成

1. 半监督学习

——训练数据包括少量正确的结果

卡通人物

中度可信度描述已自动生成

1. 强化学习

——根据结果收获的奖惩进行学习，优化结果

图示

描述已自动生成

文本, 信件

描述已自动生成

## 分类问题（Classification）

###### 分类：根基已知样本的某些特征，判断一个新的样本属于那种已知的样本类

基本框架：

图片包含 文本

描述已自动生成

###### 分类方法

逻辑回归：建立一个逻辑回归方程，判别类别

KNN近邻模型：中心点 附近 距离判别类别

决策树：树枝分叉

神经网络：模仿人类的神经元

###### 分类任务和回归任务的明显区别

分类目标：判别类别

模型输出：非连续型标签（pass/failed/1/4/5)

回归目标：建立函数关系

模型输出：连续性数值（0~10000）

线性回归拟合程度虽然好，但是有极大的局限性

问题：样本量变大以后，准确率下降

## 逻辑回归简介

###### 逻辑回归求解

根据训练样本，寻找类别边界

图示

描述已自动生成

根据训练样本，找到系数

图示, 示意图

描述已自动生成

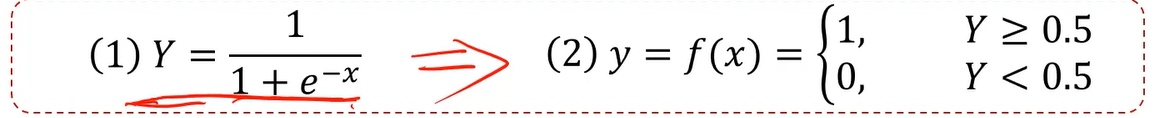
图示

描述已自动生成

图示, 文本

描述已自动生成

###### 逻辑回归



用于解决分类问题的一种模型。根据数据特征或属性，计算其归属于某一类别的概率P(x)，根据概率数值判断其所属类别。主要应用场景：二分类问题。

文本

中度可信度描述已自动生成

1. 当分类任务变得更为复杂

图表, 散点图

描述已自动生成

图示, 示意图

描述已自动生成

图片包含 图表

描述已自动生成

###### KNN Analysis （k近邻分类）监督学习

KNN算法又称k近邻分类(k-nearest neighbor classification)算法。它是根据不同特征值之间的距离来进行分类的一种简单的机器学习方法，它是一种简单但是懒惰的算法。他的训练数据都是有标签的数据，即训练的数据都有自己的类别。KNN算法主要应用领域是对未知事物进行分类，即判断未知事物属于哪一类，判断思想是，基于欧几里得定理，判断未知事物的特征和哪一类已知事物的的特征最接近。它也可以用于回归，通过找出一个样本的k个最近邻居，将这些邻居的属性的平均值赋给该样本，就可以得到该样本的属性。最简单的机器学习算法之一。

KNN分类图

图表, 散点图

描述已自动生成

图表, 散点图

描述已自动生成地图

中度可信度描述已自动生成

二、KNN算法的思想和步骤  
2.1、KNN算法的思想  
  KNN算法用于分类的核心思想是：存在一个样本数据集合，也称训练样本集，并且样本集中每个数据都存在标签，即我们知道样本集中每个数据与其所属分类的关系。输入没有标签的新数据后，将新数据的每个特征与样本集中数据对应的特征进行比较，然后算法提取样本集中特征最相似数据(最近邻)的分类标签。一般来说，我们只选择样本数据集中前k个最相似的数据，这就是k近邻算法中k的出处(通常k<20)。最后，我们选择k个最相似数据中出现次数最多的分类，作为新数据的分类。  
  KNN算法用于回归的核心思想是：跟上面一样，找到近邻的k个样本，然后取平均值作为未知样本的值，对其进行预测。  
2.2、KNN算法的步骤  
算法步骤如下：  
  1）算距离：给定未知对象，计算它与训练集中的每个对象的距离；  
   2）找近邻：圈定距离最近的k个训练对象，作为未知对象的近邻；  
  3）做分类：在这k个近邻中出线次数最多的类别就是测试对象的预测类别。

三、KNN算法的核心知识  
3.1、距离或相似度的衡量  
  在KNN算法中常使用欧氏距离、曼哈顿距离和夹角余弦来计算距离从而来衡量各个对象之间的非相似度。在实际中使用哪一种衡量方法需要具体情况具体分析。对于关系型数据，常使用欧氏距离；对于文本分类来说，使用夹角余弦(cosine)来计算相似度就比欧式(Euclidean)距离更合适。  
3.2、k值的选取  
  在KNN算法中k的选取非常重要，KNN分类的准备率对K值很敏感。不同的值有可能带来不同的结果。如果K选大了的话，可能求出来的k最近邻集合可能包含了太多隶属于其它类别的样本点，不具有代表性，最极端的就是k取训练集的大小，此时无论输入实例是什么，都只是简单的预测它属于在训练实例中最多的类，模型过于简单，忽略了训练实例中大量有用信息。如果K选小了的话，结果对噪音样本点很敏感。在实际中，一般采用交叉验证（一部分样本做训练集，一部分做测试集）或者依靠经验的方法来选取k值。k值初始时取一个比较小的数值，之后不断来调整K值的大小来使得样本分类最优，最优时的K值即为所选值。k值一般为奇数。有一个经验规则：k一般低于训练样本数的平方根。  
3.3、类别的判定  
  投票决定：少数服从多数，在k个近邻中哪个类别的点最多就分为哪类。  
  加权投票法：根据距离的远近，对近邻的投票进行加权，距离越近则权重越大（权重为距离平方的倒数）。这是考虑到各个对象的相似度有悬殊，不同距离的样本有可能对未知样本产生的影响不同。  
3.4、k个邻近样本的选取  
  在KKN算法中，整个样本集中的每一个样本都要与待测样本的进行距离的计算，然后在其中取k个最近邻。但这带来了巨大的距离计算量，这也就是懒惰算法所带来的计算成本。改进方案有两个：一个是对样本集进行组织与整理，分群分层，尽可能将计算压缩到在接近测试样本邻域的小范围内，避免盲目地与训练样本集中每个样本进行距离计算。另一个就是在原有样本集中挑选出对分类计算有效的样本，使样本总数合理地减少，以同时达到既减少计算量，又减少存储量的双重效果。KD树方法采用的就是第一个思路，压缩近邻算法采用的是第二个思路

三、KNN算法的核心知识  
3.1、距离或相似度的衡量  
  在KNN算法中常使用欧氏距离、曼哈顿距离和夹角余弦来计算距离从而来衡量各个对象之间的非相似度。在实际中使用哪一种衡量方法需要具体情况具体分析。对于关系型数据，常使用欧氏距离；对于文本分类来说，使用夹角余弦(cosine)来计算相似度就比欧式(Euclidean)距离更合适。  
3.2、k值的选取  
  在KNN算法中k的选取非常重要，KNN分类的准备率对K值很敏感。不同的值有可能带来不同的结果。如果K选大了的话，可能求出来的k最近邻集合可能包含了太多隶属于其它类别的样本点，不具有代表性，最极端的就是k取训练集的大小，此时无论输入实例是什么，都只是简单的预测它属于在训练实例中最多的类，模型过于简单，忽略了训练实例中大量有用信息。如果K选小了的话，结果对噪音样本点很敏感。在实际中，一般采用交叉验证（一部分样本做训练集，一部分做测试集）或者依靠经验的方法来选取k值。k值初始时取一个比较小的数值，之后不断来调整K值的大小来使得样本分类最优，最优时的K值即为所选值。k值一般为奇数。有一个经验规则：k一般低于训练样本数的平方根。  
3.3、类别的判定  
  投票决定：少数服从多数，在k个近邻中哪个类别的点最多就分为哪类。  
  加权投票法：根据距离的远近，对近邻的投票进行加权，距离越近则权重越大（权重为距离平方的倒数）。这是考虑到各个对象的相似度有悬殊，不同距离的样本有可能对未知样本产生的影响不同。  
3.4、k个邻近样本的选取  
  在KKN算法中，整个样本集中的每一个样本都要与待测样本的进行距离的计算，然后在其中取k个最近邻。但这带来了巨大的距离计算量，这也就是懒惰算法所带来的计算成本。改进方案有两个：一个是对样本集进行组织与整理，分群分层，尽可能将计算压缩到在接近测试样本邻域的小范围内，避免盲目地与训练样本集中每个样本进行距离计算。另一个就是在原有样本集中挑选出对分类计算有效的样本，使样本总数合理地减少，以同时达到既减少计算量，又减少存储量的双重效果。KD树方法采用的就是第一个思路，压缩近邻算法采用的是第二个思路。

四、KNN算法的优缺点  
4.1、优点  
  （1） 简单，易于理解，易于实现，无需估计参数，无需训练；  
  （2） 适合对稀有事件进行分类（例如当流失率很低时，比如低于0.5%，构造流失预测模型）；  
  （3） 特别适合于多分类问题(multi-modal,对象具有多个类别标签)，例如根据基因特征来判断其功能分类，kNN比SVM的表现要好。  
4.2、缺点  
  （1） 懒惰算法，对测试样本分类时的计算量大，内存开销大；  
  （2） 可解释性较差，无法给出决策树那样的规则。  
  （3） 该算法在分类时有个主要的不足是，当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。 该算法只计算“最近的”邻居样本，某一类的样本数量很大，那么或者这类样本并不接近目标样本，或者这类样本很靠近目标样本。无论怎样，数量并不能影响运行结果。

五、改进算法  
  针对以上算法的不足，算法的改进方向主要分成了分类效率和分类效果两方面。  
  （1）分类效率：事先对样本属性进行约简，删除对分类结果影响较小的属性，快速的得出待分类样本的类别。该算法比较适用于样本容量比较大的类域的自动分类，而那些样本容量较小的类域采用这种算法比较容易产生误分。  
  （2）分类效果：采用权值的方法（和该样本距离小的邻居权值大）来改进，Han等人于2002年尝试利用贪心法，针对文件分类时做可调整权重的k最近邻居法WAkNN (weighted adjusted k nearest neighbor)，以促进分类效果；而Li等人于2004年提出由于不同分类的文件本身有数量上有差异，因此也应该依照训练集合中各种分类的文件数量，选取不同数目的最近邻居，来参与分类。

## 病例分析

|  |  |  |
| --- | --- | --- |
| 诊断结论 | | |
| 窦性心律 | 窦性心律 | 窦性心律 |
| 窦性心律失常 | 窦性心律不齐 |
| 窦性心动过速 |
| 窦性心动过缓 |
| 快速型心律失常 | 早搏 | 室性早搏 |
| 交界性早搏 |
| 房性早搏 |
| 心动过速 | 室性心动过速 |
| 室上性心动过速 |
| 扑动颤动 | 房扑 |
| 房颤 |
| 预激综合症 | WPW |
| 缓慢型心律失常 | 逸搏 | 室性逸搏 |
| 房性逸搏 |
| 交界性逸搏 |
| 房室传导阻滞 | 一度房室传导阻滞 |
| 二度房室传导阻滞 |
| 三度房室传导阻滞 |
| 室内传导阻滞 | 室内传导阻滞 |
| 其他 | 停搏 |
| 噪声过大 |

|  |  |  |
| --- | --- | --- |
| 诊断结论 | | 备注 |
| 分析参数 | 心率 |  |
| P波长度 |  |
| PR间期 |  |
| QRS间期 |  |
| QT间期 |  |
| QTc间期 |  |
| T波长度 |  |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 医生诊断 | |
| 异常\_阳性P | |
| 算法预测 | 异常(阳性P) | TP | FP |
| 正常(阴性N) | FN | TN |

* 灵敏度（Recall）= TP / ( TP + FN )

o 灵敏度的取值范围为0到1，灵敏度越高，说明假阴性（FN）比例越低，即漏诊率（有病，算法未检出）越低

* 特异度（Specificity）= TN / ( FP +TN )

o 特异度的取值范围为0到1，特异度越高，说明假阳性（FP）比例越低，即误诊率（没病，算法却告诉我有病）越低

* ROC曲线下面积（AUROC）：以Recall与Specificity为X-Y轴绘制ROC曲线（受试者工作特征曲线），曲线下面积（AUROC）用于衡量不同算法模型的整体优劣。

o AUROC的取值范围为0到1，AUROC越大，说明算法整体表现越好

**数据集一：CPSC**

中国心电竞赛开源数据 ：The China Physiological Signal Challenge 2018: Automatic identification of the rhythm/morphology abnormalities in 12-lead ECGs。

* 样本量：6877例，平均长度为30秒的12导联ECG，取I导联为单导联测试数据
* 数据来源：11家医院
* 标注方式：未知

测试结果（AUROC）

表格

描述已自动生成 图表, 折线图

描述已自动生成

**数据集二**

三甲医院心电科门诊数据（私有）

* 样本量： 测试数据集15437例
* 标注方式：从心电科医生的原始文本诊断中抽取

测试结果（Recall、Specificity、AUROC）

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 诊断结论 | 标签 | Recall | Specificity | AUROC |
| 窦性心律 | SN | 0.9996 | 0.8182 | 0.9918 |
| 窦性心律不齐 | SNA | 0.7206 | 0.9692 | 0.9664 |
| 室性早搏 | PVC | 0.9138 | 0.9973 | 0.9943 |
| 房性早搏 | PAC | 0.7905 | 0.9910 | 0.9768 |
| 房扑 | AFL | 0.9091 | 0.9998 | 0.9997 |
| 房颤 | AF | 0.8824 | 1.0000 | 0.9998 |
| 左束支传导阻滞 | LBBB | 0.9524 | 0.9996 | 0.9997 |
| 右束支传导阻滞 | RBBB | 0.7778 | 0.9980 | 0.9934 |