机器学习 讨论题

讨论小组： 2班 第二小组 （序号22~41）

学习任务： 完成对应小组编号的讨论题

Intro T2.2 监督学习与无监督学习的区别和联系，那么有没有半监督学习呢？监督学习、无监督学习、半监督学习对数据的要求各是怎样的？

答： 监督学习和无监督学习的区别：

1.原理不同 监督学习是指利用一组已知类别的样本调整分类器的参数，使其达到所要求性能的过程。无监督学习指根据类别未知(没有被标记)的训练样本解决模式识别中的各种问题的过程。

2.算法不同 监督学习的算法是通过分析已知类别的训练数据产生的。无监督学习的算法主要有主成分分析方法、等距映射方法、局部线性嵌入方法、拉普拉斯特征映射方法、黑塞局部线性嵌入方法和局部切空间排列方法等。

3.适用条件不同 监督学习适用于样本数据已知的情况。非监督学习适用于无类别信息的情况。 二者的联系： 有训练样本则考虑采用有监督学习方法；无训练样本，则一定不能用有监督学习方法。但是，现实问题中，即使没有训练样本，我们也能够凭借自己的双眼，从待分类的数据中，人工标注一些样本，并把它们作为训练样本，这样的话，可以把条件改善，用监督学习方法来做。

有没有半监督学习： 半监督学习是模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。半监督学习使用大量的未标记数据，以及同时使用标记数据，来进行模式识别工作。当使用半监督学习时，将会要求尽量少的人员来从事工作，同时，又能够带来比较高的准确性，因此，半监督学习正越来越受到人们的重视。

三者对数据的要求：
监督学习：
训练集要求包括 输入和输出，也可以说特征和目标，目标是自己标注的。即在已知的模型中，寻找最优解。
无监督学习：
输入数据没被标记，即无法预测样本标签。
半监督学习：
分两类：
1.直推SSL只处理样本空间内给定的训练数据，利用训练数据中有类标签的样本和无类标签的样例进行训练，预测训练数据中无类标签的样例的类标签。
2.归纳SSL处理整个样本空间中所有给定和未知的样例，同时利用训练数据中有类标签的样本和无类标签的样例，以及未知的测试样例一起进行训练，不仅预测训练数据中无类标签的样例的类标签，更主要的是预测未知的测试样例的类标签。

Perceptron T3-1 神经元的组成是怎样的?

答： 由胞体，枝蔓，突触和轴突组成。感知器由一组突触或连接链路，一个加法器和一个激活函数单元组成。

Perceptron T3-2 单个神经元可以用来做什么?

答： 可以做二分类器，把空间用一个超平面分成两半，解决线性可分问题。

Perceptron T3-3 线性可分问题是什么问题?

答： 线性可分指的是可以用一个线性函数将两类样本分开,比如在二维空间中的直线，三位空间中的平面以及高维空间中的线性函数。

Perceptron T3-4 从数据角度谈机器学习的挑战都有哪些?

答： 机器学习在数据方面的挑战有：1.数据形式具有多样性 2.数据质量参差不齐 3.数据维度大，含有过多无关属性。

Perceptron T3-5 感知器学习的梯度下降法，其基本要素有哪些?

答： 学习率、梯度、迭代次数。

MLP T4-1 请问反向传播算法是有监督的吗?

答： 是。有监督学习：通过已有的训练样本去训练得到一个最优模型，再利用这个模型将所有的输入映射为相应的输出，对输出进行简单的判断从而实现预测和分类的目的，也就具有了对未知数据进行预测和分类的能力。反向传播算法需要对训练集进行训练得到训练模型，因此是有监督的学习算法。

MLP T4-2 反向传播算法的目标函数是什么?

答： 目标函数是经验风险和结构风险最小化的最终优化函数。即对历史数据的拟合较好且结构较简单。反向传播算法的目标函数是损失函数对所有权重的偏导数，通过梯度下降法和其他更加高级的优化算法对损失函数进行最小化。反向传播算法，实质上就是使用链式法则求解每一层的中间变量（传播误差）δ(l)，并利用δ(l)计算损失函数J对每一层的权重矩阵W(l)中的每一个权重Wji的偏导数。

MLP T4-3 classification 与regression 这两类问题有什么不同？为什么不能用解决regression问题的做法来解决classification问题?

答： 1.回归问题的应用场景（预测的结果是连续的，例如预测明天的温度，23，24，25度）

回归问题通常是用来预测一个值，如预测房价、未来的天气情况等等，例如一个产品的实际价格为500元，通过回归分析预测值为499元，我们认为这是一个比较好的回归分析。一个比较常见的回归算法是线性回归算法（LR）。另外，回归分析用在神经网络上，其最上层是不需要加上softmax函数的，而是直接对前一层累加即可。回归是对真实值的一种逼近预测。

2.分类问题的应用场景（预测的结果是离散的，例如预测明天天气-阴，晴，雨）

分类问题是用于将事物打上一个标签，通常结果为离散值。例如判断一幅图片上的动物是一只猫还是一只狗，分类通常是建立在回归之上，分类的最后一层通常要使用softmax函数进行判断其所属类别。分类并没有逼近的概念，最终正确结果只有一个，错误的就是错误的，不会有相近的概念。最常见的分类方法是逻辑回归，或者叫逻辑分类。

为什么不能用回归解决分类问题
大多数情况下分类输出离散数据，回归输出连续数据，分类寻找决策边界，回归寻找最优拟合，两者目的不一样，分类用精度，混淆矩阵FPR，TPR等参数评价，回归用SSE评价

Perf-T-3 所谓泛化能力，你认为指的是什么？在评价一个学习机的泛化能力上，你认为存在哪些可能的困难和问题?

答： 泛化能力通俗来讲就是指学习到的模型对未知数据的预测能力。在实际情况中，我们通常通过测试误差来评价学习方法的泛化能力。

可能存在的困难和问题 1.训练数据的数量不足 大部分机器学习算法需要大量的数据才能正常工作。训练数据不足可能会影响泛化能力的评估

2.训练数据不具代表性 为了很好地实现泛化，训练数据一定要非常有代表性。使用不具代表性的训练集训练出来的模型不可能做出准确的预估。

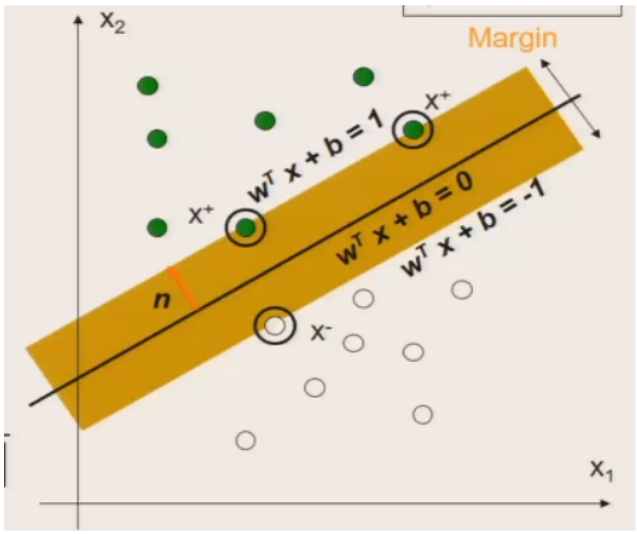
3.低质量数据 如果训练集满是错误、异常值和噪声（例如，低质量的测量产生的数据），系统将更难检测到底层模式

4,训练次数过多。拟合了训练数据中的噪声和训练样例中没有代表性的特征，泛化能力评估不准确。

SVM T-2 什么叫支持向量？它是怎样得到的？它在分类问题上的重要性表现在哪里？

Q：什么叫支持向量？它是怎样得到的？他在分类问题上的重要性表现在哪些方面？

A： 支持向量机算法的目标是在 N 维空间中找到一个对数据点进行明确分类的超平面。而支持向量就是更接近超平面并影响超平面的位置和方向的数据点。



即上图中黑色圆圈所圈出的部分。

决策边界上的支持向量点满足：

$$\mathbf{w}^T\mathbf{x}^+ + b = 1$$
$$\mathbf{w}^T\mathbf{x}^- + b = -1$$

支持向量的重要性：

1. 通过支持向量，可以实现最大化分类器的边距，而最大化分类边际的思想是SVM方法的核心；
2. 在SVM分类决策中起决定作用的是支持向量。
3. 拥有高维样本空间的数据也能用SVM，这是因为数据集的复杂度只取决于支持向量而不是数据集的维度，这在某种意义上避免了"维数灾难"（如下式所示）

$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \sum_{i \in \mathrm{SV}} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b$$

4. 支持向量决定了SVM算法具有较好的“鲁棒”性。

Kmeans-T-2 K-means聚类是一个优化问题吗?其目标函数是怎样的?优化过程(比如梯度下降)所得到的就是所谓的k-means 算法吗?

答：K-means聚类是一个优化问题；

其目标函数是误差的平方和(Sum of the Squared Erro.SSE)，我们计算每个数据点的误差，即它到最近质心的欧氏距离，然后计算误差的平方和。SSE形式地定义如下:

$$\mathrm{SSE}=\sum_{i=1}^{k}\sum_{x\epsilon C_i} dist(c_i,x)^2;$$

优化过程所得到的不一定就是所谓的k-means算法，因为最终得到的结果可能并不是k个簇，并且可能得到的是局部最优解而不是全局最优解。

SOM-T-2 SOM是通过怎样的机制实现拓扑保持的?

答： SOM只包含输入层和输出层，由于没有中间的隐藏层，所以SOM映射之后的输出保持了输入数据原有的拓扑结构，SOM 网络能将任意维输入模式在输出层映射成一维或二维图形,并保持其拓扑结构不变;网络通过对输入模式的反复学习可以使权重向量空间与输入模式的概率分布趋于一致。