**成绩：\_\_\_\_\_\_\_**

****

**课程名称：** 与时俱进,学习知识

**班级：** XXXX级专业名称X班

**姓名：**  云湖不归

**学号：**  XXXXXXXXXXX

# 机器学习的视频

## 什么是机器学习

所谓的”**机器学习**”, 顾名思义, 就是让”机”去学习, 这个”机器”可以是计算机.不同于人类的语言, 计算机没有嘴巴, 不能”叙述”, 但是计算机懂得**识别**, 它能认出”**0**”和”**1**” , 这就是它”**说话**”的基础. 也是它学习的最基本的元素.

例如: 现在有一张南瓜的图片, 计算机对图片在一定范围的标号,从而识别整个图片的不同像素, 然后记住这些像素的位置; 如果此时再有10万张不同南瓜的图片, 将其交给它, 重复之前就按的操作, 让其找到某种规律,来认定,**什么是南瓜**. 这种规律就是一种模型. 而这整个过程就是机器学习.

说白了, 我有目的的给机器数据, 机器自己总结规律, 这就是机器学习.

## 常用的机器学习模型

**邻近分类之 ()**

所谓的也是用来分类的一种算法, 具体逻辑如下:

1. 划定一片空间, 也就是定义空间

2. 定义距离公式,例如; 欧式公式 , 用来计算某个

点与周围点的距离.

3. 录入样本点信息, 也就是将数据输入进去

4. 找最近的哥样本点, 并且计算找到的最近的个样本点中, 哪个类别最多,

让后将该点归类到那一类.

相当于在一个班级(空间) , 分了几组(类), 对于整个班级再说, 某个人可能擅长写代码, 而周围的另一个同学也擅长, 那么这两人**大概率**就会在一组中, 而这个班级中的其他成员分类也是以此类推.

**注意:**

**的缺点**包括: ① 局部估算可能不符合全局 ②不能计算概率 ③对于

的取值很敏感**,一般不超过**.

**的优点**包括: 直观好理解, 不需要估算整体, 只需要看局部.

**邻近分类之 (决策树)**

所谓的决策树其实就是设置层层条件, 计算机根据这些条件去进行分类, 从而得到人们想要的分类结果. 也就是说人们设置的**判断条件**是决策树能够正确分类的关键因素, 而判断条件的好坏可以使用**熵**来进行判断, 而**熵**就是用来衡量节点内的不确定性. 如果一组分类中一致性很高, 那么熵就会很低, 反之, 如果一组分类中的一致性很低, 那么熵就会很高. 熵的计算公式如下:

那么对于当前节点, 人么可能有很多条件, 那么在使用的时候就要选择一个最好的用作分类, 也就是找到最适合的那个, 怎样才能判断使用的决策条件最适合呢?这就要使用到**熵的增益**, 即**上一层的熵减去当前一层的熵的总和**, **具体公式** :

这个算法也叫 , 但其只适用于**分类问题**.

综上, 人们需要找到**最小的熵, 熵增益**.

优势: 直观可视化, 易于追溯和倒推

劣势: ① 决策树的深浅会影响最终的分类, 导致

② 如果数据的维度很多, 那么判断条件的复杂度就会更高

**随机森林**

**为了实现 的效果,** 或者说发挥各方的**长板**优势, 使用这种算法.

**具体如下:**

说白了, 在决策树的基础上, 做数据的分类, 结果的整合.

例如: 设置三棵树, 每棵树设置三层, 将不同的数据集放入不同的树,

进行决策, 这样就可以得到三种不同的结果. 将这三种不同的结

果做整合, 就可以了.

**基本流程:**

1. 设置几棵树, 每棵树有几层

2. 随机采样, 也及时将数据分成多份

3. 输入待测样本到每棵树中, 然后对结果进行整合,

对于问题, 求**平均**.

对于问题, 求**众数**.

**优点:**

1. 模型随机性很强, 不易

2. 抗噪性强, 不易不敏感

3. 处理高维数据更快

4. 树状结构, 模型的可解释度高, 突出每个特征的重要性.

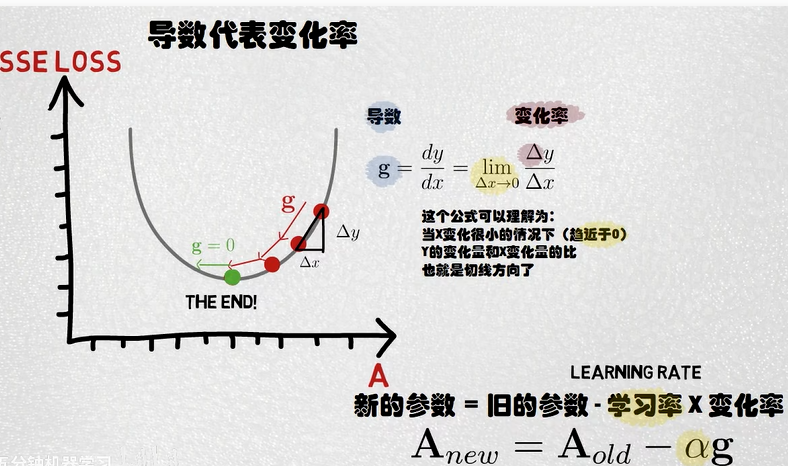
**缺点:**

模型无法处理困难的问题, 也就是难处理的数据没法正确处理.

**(梯度下降法)**

这种算法就是用来**加快计算速度**的, 结合**损失函数**来判定结果

图解:



**优点:** 更新求解快, 与高度相关

**缺点**: 1. 只能知道导数方向, 无法知道到底还有多远才能得到我们想要的结果

2. 只能保证局部最优, 无法保证全局最优.

**(监督学习)**

也叫概念学习, 控制训练集, 给出输出的结果, 达到预期的结果, 最后验证.

**具体步骤如下:**

1. 确定训练集的类型

2. 搜集训练集

3. 确定输入特征的表示法

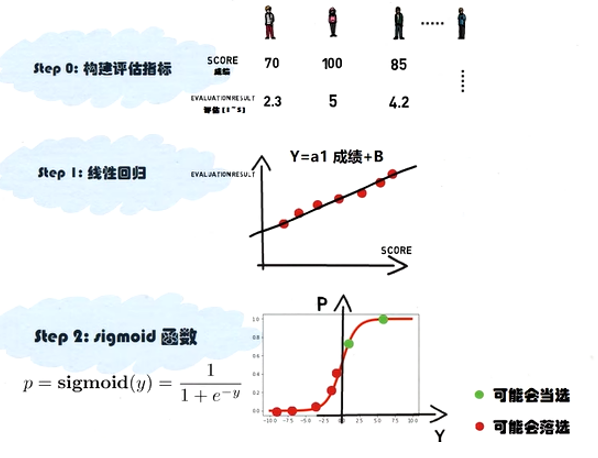
4. 确定学习的函数和对应的数据机构

5. 完成设计, 进行调优.

**回归模型之 (逻辑回归)**

其重要作用就是根据人们自己选定的参数, 对这个参数作线性回归, 如果运气好的话那会很快的看到一种趋势, 但是为了使曲线上的点更密集, 较好的处理异常值, 所以还需要使用平滑函数, 例如: .

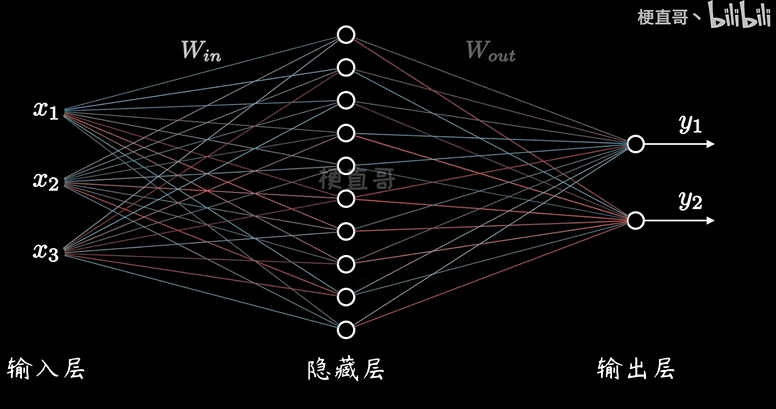
**具体流程如下方所示:**



对于逻辑回归的进度进行评估需要使用交叉熵函数:

我们最终得到的是介于之间的数值, 也就是说逻辑回归更倾向于概率, 例如:人们评估事件的好坏等.

**回归模型之 (循环神经网络)介绍**



简单来说, 就是让这种神经网络本身具有了**一定的记忆能力**, 这种记忆能力主要是在隐藏层上, 那么怎样记性记忆呢? 我觉得可以这样理解, 早上我吃了两个肉饼, 一个鸡蛋, 如果在我吃中午饭的时候, 我的脑海里能想起”包子”, “馒头”, 那不就是记忆吗? 而体现在数学公式上就是, 最开始的隐藏层用到的公式为

而则是:

也就是说 与 被关联起来了, 也就是所谓的记忆, 到了仍然记得

需要注意的是 采用了误差反向传播和梯度下降来计算权重, 只是由于有了所谓的记忆, 所以每次输入的数据不同了而已.

以上表示对于该种神经网络的理解, 对于具体的使用范围来说: 善于处理, 例如: 图片转文字, 音频, 翻译等.

**浅层神经网络**

也就是只有一个**隐藏层**的神经网络, 整个神经网络包括:

输入层() , 隐藏层() , 输出层()

具体的示意图如下:

输出层

隐藏层

输入层

上方给出的就是三个样本, 当个隐藏层的神经网络, 也就是浅层神经网络.

**注意: ①** 在计算层数的时候, 输入层不算, 所以上方的浅层神经网络只有隐藏

层和输出层这两层.

**②** **隐藏层:** 包括 和

以上方的浅层神经网络为例具体公式如下:

其中表示权重, 表示偏置项

为了方便计算, 写成矩阵的形式:

**,**

特别特别注意**维度**, 以上图为例:

为 , ,的维度与之一致.

为 ,

为

也就是说 **行表示神经元的个数, 列表示输入的样本数目.**

在浅层神经元中,

1. **权重在初始化的时候, 不能全为零, 否则隐藏层的多个神经元就没啥意**

**义了.权重一般都是都是随机的, 比如通过函数来产生,** 并且设

置的时 候最这样梯度下降的比较快.

2. 常用激活函数包括**: 函数, 函数 , 函数,**

**函数,** 而在浅层神经网络中, 一般使用过的都是非线性的激活函数, 因为

如果投入的是线性的激活函数, 最终得到的记过也是线性的

## 术语介绍

**样本数据划分**:

**训练集:** 说白了就是用来训练的, 也就是我们输入到模型中的数据集

**测试集:** 在模型训练好了之后, 进行测试的. 就如同一名学生完成了一学期的课程,

通过期末考试, 来看看这个学生学的怎样.

**验证集:** 也就是将**训练集**作为样本, 划分出新的训练集和测试集, **新的训练集**就是

**验证集**, 用来挑选最优的参数

**模型性能评价标准**:

**泛化能力:** 说白了就是检验一个学生的学习相同领域, 新东西的哪里, 比如一个只

会Fortran的人, 让其去学Python , 这两种语言的基础语法差不多, 如\

果这个人学的很好, 那就是泛化能力强, 反之, 就是泛化能力弱.

**过拟合:** 说白了就像一个故步自封的人, 只会在自己领域搞得好, 遇到外界新的

知识, 手足无措.

**稳定性:** 输入计算机的数据发生变化, 其输出是否会发生大幅度变化

**可解释性:** 说白了就是人们要知道, 比如,模型是怎么跑出那些所谓的最优的

参数的, 怎么找到的, 对于我们气象人来说, 用什么理论去解释, 怎么解释, 也就是说我们要能用具体的理论去解释跑出来的结果. 而不知道非常糊涂的说, **只要结果好就行**.

## 气象领域AI面临的挑战

1. 现在能得到的观测数据不够丰富, 样本数据不够多, 所以容易过拟合同样是

因为数据量少, 所以机器学习得到的模型不稳定, 所以无法真正的投

入到预报当中.

2. 气象领域理论基础非常丰富, 机器学习得到的模型在运行的时候不能违背这

些理论和客观事实.

3. 可解释性太弱, 即使模型预报的很准, 如果人们不知道为什么准确, 以及之后

该如和继续优化都是问题, 大多数情况下, 模型都是黑盒的, 或者说, 人们知

道如何设置真个程序, 设计算法, 但是对于利用气象领域的知识和理论来解释

对应的结果, 是非常差的

4. 个人觉得算力还有待提高, 量子计算机正在发展, 如果能够投入到气象上使

用, 对于预报的准确性将无疑是巨大的飞跃, 因为网格的进度, 算力的消耗也

是不容小觑的, 而模型的训练本身就是需要等性能良好的设备, 如果可以运行的更快, 那么人们将有更多的时间去调参, 优化, 改进, 深入探索.

# 气象统计回归中对于过拟合问题的认知

1. **缺数据:** 气由于气象的资料无法做到所谓的大样本, 即在实际中气象数据还

是不够多, 所以导致气象统计上容易过拟合;

2. **非线性:** 象统计上的过拟合首先一定不是纯数学的过拟合, 它是具有物理意\义的, 而大多数物理公式都是非线性的, 气象领域的很多解都不是数学上的那种已经很精确的解了, 我们都是做了假设的, 有些即使做了假设也没有解析解, 由此观之正是由于公式的非线性等复杂原因, 导致气象回归中很多时候已发生过拟合

3. **参数问题:** 整个世界是混沌的, 一个事件往往是有多种因素造成的, 所以对于气象上的某个气象要素, 天气系统, 天气形势来说, 影响的因子往往是多样的, 人们无法准确的求解, 有时候多个因子都很重要, 但是罗列的公式有十分复杂, 既要满足数学上的约束, 也要满足显示的约束, 所以参数的复杂性, 多样性, 也是过拟合的影响因素.

我想过拟合还有很多影响因素, 但是事物是复杂的, 我们只需知道那些事主导因

素就可以了. 过拟合的评判, 可以通过过**拟合系数**来评判, 而为了避免过拟合,

可以通过正则, 等方法来弥补或者说抑制.

下面展示两个案例:

# 案例

## 浅层神经网络模型展示不稳定性

**数据介绍**:

总的样本数为的1000个点 与 的100个点, 共计1100个值.

**大样本训练集:** 的**1000** 点

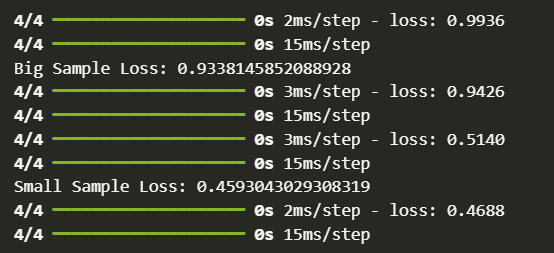
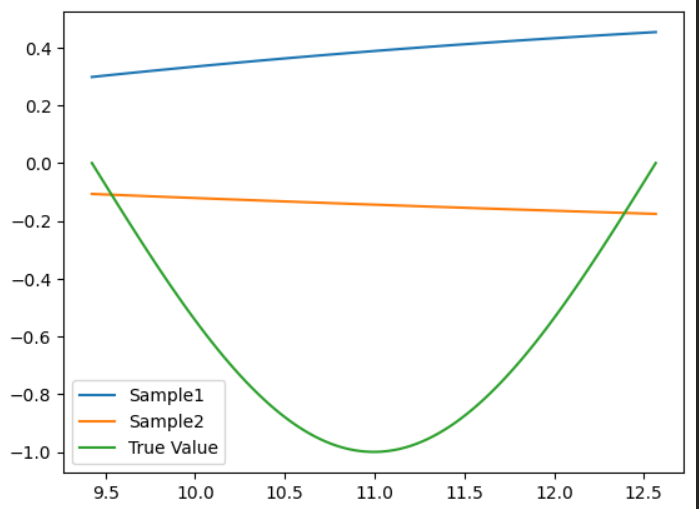
**小样本训练集:** 从的1000点中选出的**100**个点.

**测试集**为的100个点

两个样本使用的激活函数是相同的

在单个隐藏神经元, 即浅层神经网络情况下, 以均方差作为损失函数.

大样本情况下的约为 , 小样本的情况下的约为

大样本的损失, 即采取的均方差大于小样本的均方差. 两者相差较大, 并且两者的损失本就较大, 所以其不稳定性是比较明显的. 并且本人更改激活函数后, 其损失随之发生较大的变化, 由此可见, 浅层神经网络的确是不稳定的, 它受到激活函数, 网络的深度, 样本大小等多种因素的影响. 当然本次大样本, 小样本预测值都是与真实值有非常大的差距, 也许是方法设计的不对. 这也是之后要认真学习的, 即神经网络到底要怎么设置, 神经元要设计几个, 要设置多少层等等这些知识点并未十分熟练掌握.

## 回归模型展示过拟合

上图采用的数据:

总的 : 从到3 取500个点.

总的 : , 其中

训练集: 占用的总的数据

测试集: 占用的总的数据

上图是采用线性回归, 分别对于次的多项式进行拟合

从图中明天的看到, 当多项式的最高逐渐升高的时候, 线性回归的过拟合程度就越大, 可以明显的看到, 当最高次的时候, 在的时, 过拟合候尤为明显.