1. CNN优缺点：

优点：共享卷积核，对高维数据处理无压力；

无需手动选取特征，训练好权重，即得特征；

深层次的网络抽取图像信息丰富，表达效果好。

缺点：需要调参，需要大量样本，训练最好要GPU；

物理含义不明确。

1. VC维

VC维的直观定义:

* 对一个指示函数集，如果存在h个样本能够被函数集中的函数按所有可能的2h种形式分开，则称函数集能够把h个样本打散。
* 函数集的VC维就是它能打散的最大样本数目h。
* 若对任意数目的样本都有函数能将它们打散，则函数集的VC维是无穷大。

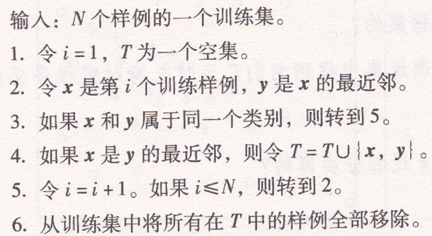
d维空间超平面的VC维是d+1。

1. What are Tomek Links, are how do we identify them in the training set? Why does the procedure sometimes have to be repeated?（什么是Tomek链接，我们如何在训练集中识别它们？ 为什么程序有时需要重复？）

一个样例对，x和y，如果同时满足如下3个条件，就说形成了一个托梅克连接：

1. x是y的最近邻
2. y是x的最近邻
3. x和y的类别不同

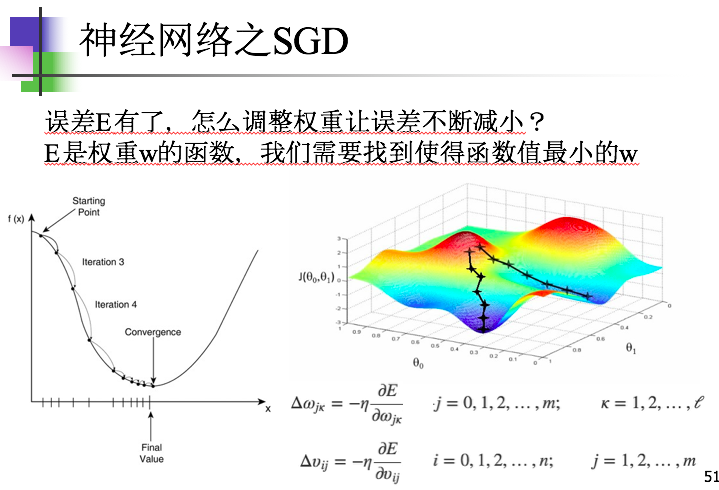
注意：有时候，移除托梅克连接后，会产生新的托梅克，所以需要重复操作，直至结束。



当其样本分布不平衡时，比如其中一类样本过大（实例数量过多）占主导的时候，新的未知实例容易被归类为这个主导样本，因为这类样本实例的数量过大，但这个新的未知实例实际并木接近目标样本

1. 什么是局部极小值问题？为什么它会成为训练中的一个问题？如何降低陷入一个局部极小值的危险（用具体方法或例子说明）

答：比如，训练中如果过早陷入局部最小值，算法就得到了一个不理想的结果，影响最终效果。



答：为梯度更新法则加一个冲量项。冲量有时可以带动梯度下降过程，冲过狭窄的局部最下值。

# （bp神经网络算法陷入了局部最小值

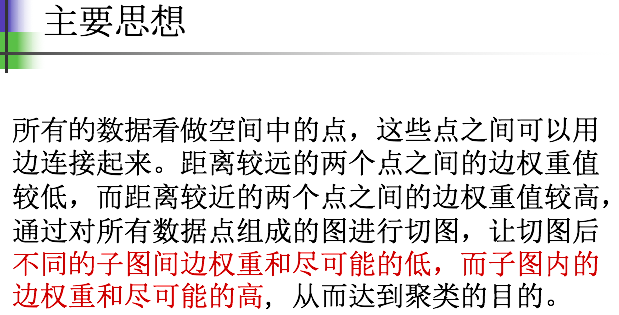
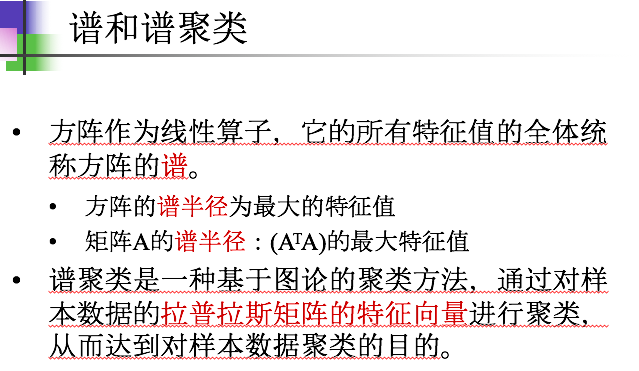
1. 使用改进的BP算法，增加动量项，或者自适应学习率。
2. 和别的优化算法组合，例如遗传算法优化初始权值，提前锁定全局最优。
3. 重新训练，每次训练的结果都是不同的，下一次的训练很有可能不会陷入局部极小。
4. 更改学习函数、训练函数再试。）

BP算法基本想法：信号正向传播和误差反向传播

信号正向传播：正向传播时，输入样本从输入层传入，经各隐层逐层处理后，传向输出层。若输出层的实际输出与期望的输出不符，则转入误差的反向传播阶段。

误差反向传播：将输出以某种形式通过隐层向输入层逐层反传，并将误差摊给各层的所有单元，从而获得各层的误差信号，此误差信号即作为修正各单元权值的依据

1. 谱聚类算法流程



以Ncut总结谱聚类算法流程

输入：样本集D=(x1,x2,...,xn)，

相似矩阵的生成方式, 降维后的维度k1,

聚类方法，聚类后的维度k2

输出： 簇划分C(c1,c2,...ck2)

1) 根据输入的相似矩阵的生成方式构建样本的相似矩阵S

2）根据相似矩阵S构建邻接矩阵W，构建度矩阵D

3）计算出拉普拉斯矩阵L

4）构建标准化后的拉普拉斯矩阵D−1/2LD−1/2

5）计算D−1/2LD−1/2最小的k1个特征值所各自对应的特征向量f

6) 将特征向量f标准化，最终组成n×k1维的特征矩阵F

7）对F中的每一行作为一个k1维的样本，共n个样本，用输入的聚类方法进行聚类，聚类维数为k2。

8）得到簇划分C(c1,c2,...ck2).

优点：1）谱聚类只需要数据之间的相似度矩阵，因此对于处理稀疏数据的聚类很有效。这点传统聚类算法比如K-Means很难做到

2）由于使用了降维，因此在处理高维数据聚类时的复杂度比传统聚类算法好。

缺点：1）如果最终聚类的维度非常高，则由于降维的幅度不够，谱聚类的运行速度和最后的聚类效果均不好。

2) 聚类效果依赖于相似矩阵，不同的相似矩阵得到的最终聚类效果可能很不同。