Part1 深度学习基础 第4章 机器学习基础

一、机器学习的四个分支

(1) 监督学习：人工标注，有标签，标签学习

(2) 无监督学习：无目标，无标签，对数据内在特征挖掘，找到样本间的关系，降维、聚类和离散点检测

(3) 自监督学习：监督信息不是人工标注的，而是自动在数据中构造和挖掘自身的监督信息，发现数据它自己有什么信息，自动编码器

(4) 强化学习：智能体agent，奖励，高分，环境，游戏

https://blog.csdn.net/sdu_hao/article/details/104515917

https://zhuanlan.zhihu.com/p/96748604

二、ML评估

评估即衡量模型的泛化能力，过拟合是最大问题!

2.1 ML目的

get a good generalize and performace model，未知数据也可以表现很好，泛化能力很强

2.2 数据集的划分

(1) train， val and test dataset

(2) 作用

train dataset：training model

validation dateset：validation model，调节参数，反馈信号

test dataset：evalute model or prediction，全新未知数据，评估衡量

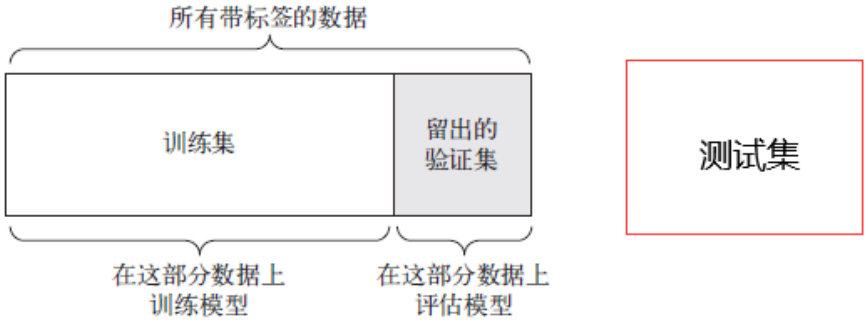
2.3 信息泄露 information leak

如果将数据集只划分为训练集和测试集，就会有信息泄露的可能，因为在训练集上对模型进行训练后，通过测试集测试后会有标准来评判模型的优劣，这时候会去调整模型的参数，在调整后再次训练，这时候测试集的数据已经参与到了模型的训练当中，就会根据测试集的反馈来调整，这样模型的数据在训练集和测试集上都是见过的，容易造成记忆，过拟合

造成这一现象的关键在于**信息泄露**（information leak）。每次基于模型在验证集上的性能来调节模型超参数，都会有一些关于验证数据的信息泄露到模型中。如果对每个参数只调节一次，那么泄露的信息很少，验证集仍然可以可靠地评估模型。但如果你多次重复这一过程（运行一次实验，在验证集上评估，然后据此修改模型），那么将会有越来越多的关于验证集的信息泄露到模型中。

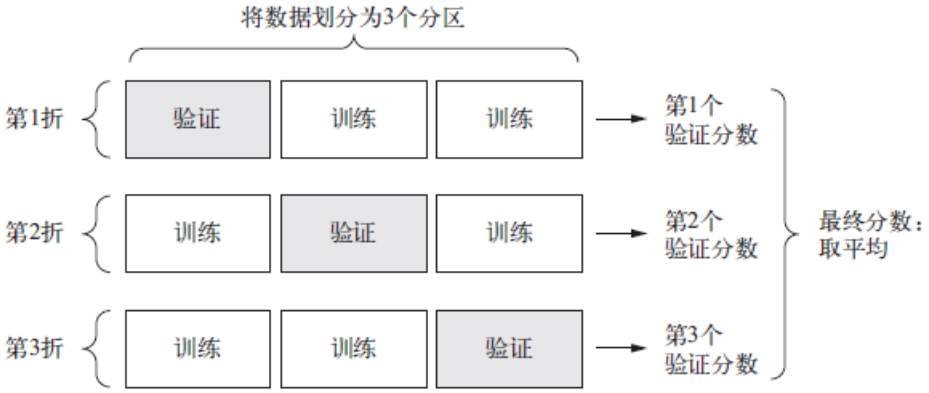
为了避免信息泄露，将数据划分为3部分

2.4 可用数据很少怎么办



(1) 留出法：原训练集一部分参与训练，一部分留出来作为验证集，但如果数据本身就很少，划分后，这样val和test的数据太少了，无法从统计学上代表数据

(2) K折交叉验证



分成K个分区，K-1个训练，剩下1个验证

https://www.zhihu.com/question/29350545?sort=created

2.5 注意事项

(1) 随机打乱顺序，假如mnist数据是按类别从0到9依次排序送入网络，神经网络就会发现这个规律，就会记忆这个顺序，当新的数据到的时候，是根据顺序来预测的，因为网络根本没有挖掘数字本身的特征

(2) 有的不能打乱，例如时间序列，因为这个序列本身是有先后顺序的，要保证语义前后相关一致，而且打乱后也会造成时间泄露

(3) 确保测试集是全新未知的，没有和其他的有任何交集数据，一方面防止数据污染，导致网络性能下降，另一方面如果数据有重复，这样你可能是在掺杂着部分训练数据上评估，这样是不准确的

三、数据预处理、特征工程

3.1 预处理

预处理后的数据对神经网络更适合，帮助网络更好的学习，更快的收敛

例如向量化、标准化、放缩、缺失值处理

其中标准化最好使得特征均值为0，方差为1

公式：$$x' = \frac{x - \mu}{\sigma}$$

放缩到[0,1] or [-1,1] or [-0.5, 0.5]区间范围内是比较好的

缺失值：只要0没有意义，不代表什么，就没有影响，放到网络中也会忽略

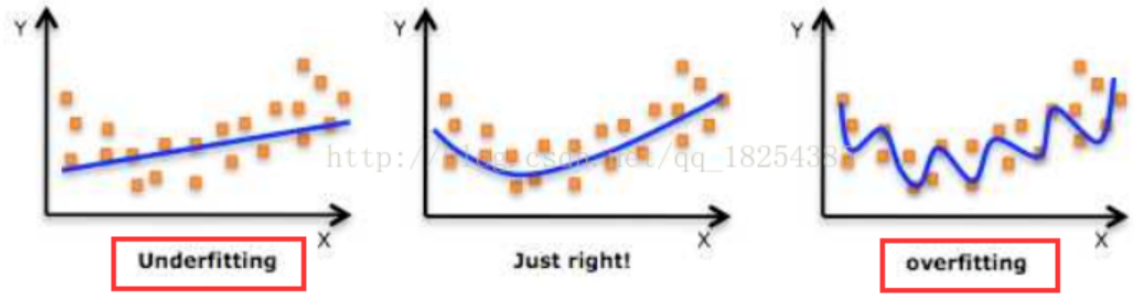
3.2 特征工程

数据编码变换

多数情况下，一个机器学习模型无法从完全任意的数据中进行学习，呈现给模型的数据应该便于模型进行学习

现代deep learning不太需要特征工程了，因为神经网络可以自动提取有用特征

四、过拟合和欠拟合



https://www.cnblogs.com/zhhfan/p/10476761.html

4.1 欠拟合

模型的表达能力不强，无法很好的表达数据，表现训练集上效果都不好，更不用说测试集

神经网络还有提升的空间，还可以再增加复杂度

4.2 过拟合

模型的表达能力太强，有足够表达数据的能力，表现在训练集上效果非常好，但就是因为模型很复杂，表达太好，将数据当中的噪声也学会了，网络会极度逼进训练数据当中的每个值，当新的数据来的时候，测试集的数据是全新的，表现的很差，泛化能力差

深度学习当中最容易出现，也是相对比较麻烦的是过拟合

4.3 解决过拟合

最理想的状态是刚好在欠拟合和过拟合的界线上，在容量不足和容量过大的界线

网络应该具有足够多的参数，防止欠拟合，但网络不应该有很多的 参数，导致过拟合

(1) 减少网络容量，减少参数的数量

(2) 添加正则化L1 or L2

奥卡姆剃刀原理：如果一件事情有两种解释，那么最可能正确的解释就是最简单的那个，即假设更少的那个。

即越复杂的模型更容易过拟合，参数越少越好

正则化是强制让模型的权值取较小的值，限制模型的复杂度，使得权值的分布更加规则，它的做法是在损失函数中加入成本

在最小化损失函数的时候也会同时最小化成本，即让成本/参数大规模的逼近于0，参数变得稀疏

公式：

- L_0-范数：$\|\vec{x}\|_0 = \#(i),\ \text{with}\ i \neq 0;$
- L_1-范数：$\|\vec{x}\|_1 = \sum_{i=1}^{d}|x_i|;$
- L_2-范数：$\|\vec{x}\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2};$
- L_p-范数：$\|\vec{x}\|_p = \left(\sum_{i=1}^{d} x_i^p\right)^{1/p};$

区别：
使用L1范数，可以使得参数稀疏化；
使用L2范数，倾向于使参数稠密地接近于0，避免过拟合。

(3) 添加dropout

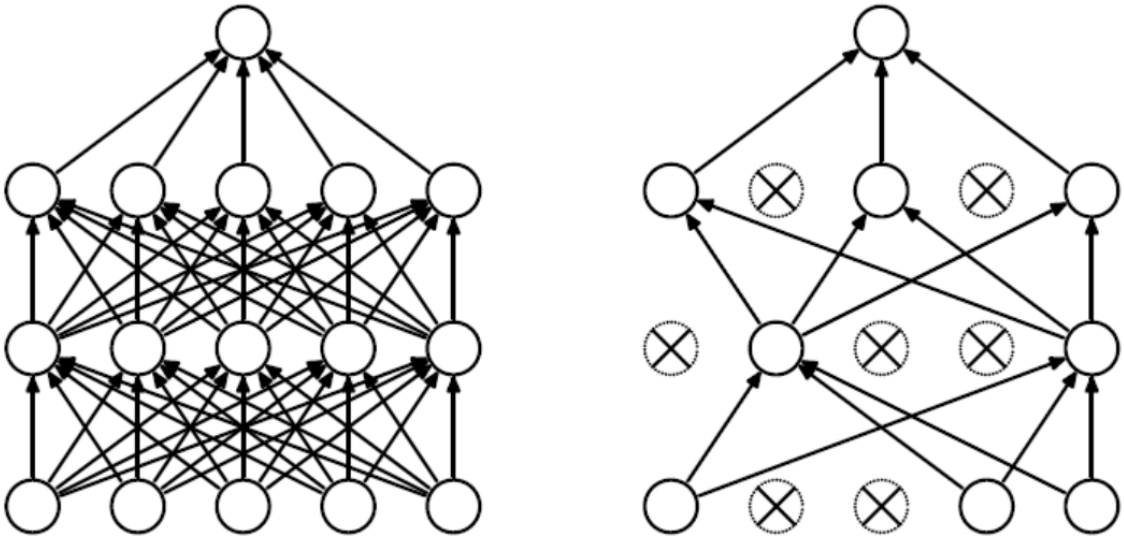


图1：使用Dropout的神经网络模型

每层如果使用了dropout，这一层会随机丢弃掉部分的神经单元，每个神经单元都有一定的概率会断开，置0，对整个网络不做出贡献

dropout为什么可以降低过拟合?

来自于银行的防欺诈机制。用他自己的话来说：“我去银行办理业务。柜员不停地换人，于是我问其中一人这是为什么。他说他不知道，但他们经常换来换去。我猜想，银行工作人员要想成功欺诈银行，他们之间要互相合作才行。这让我意识到，在每个样本中随机删除不同的部分神经元，可以阻止它们的阴谋，因此可以降低过拟合。”[1] 其核心思想是在层的输出值中引入噪声，打破不显著的偶然模式（Hinton 称之为**阴谋**）。如果没有噪声的话，网络将会记住这些偶然模式。

避免了神经元之间的依赖关系，让神经元真正去理解特征，而不是考协作记忆

也有种观点是集成学习，因为每次进行训练的时候随机丢弃的神经元是不一样的，这样就形成了不同的子网络，由多个子网络共同决策权值

(4) 添加数据量

神经网络是靠数据feed出来的，大量的数据才是真正避免过拟合的根本方法，因为网络可以见识到更多的数据，有更大的学习空间，掌握更多的数据模态

五、DL通用流程

准备数据，加载数据，数据预处理，确定类型(2分类、多分类、回归等)，选择度量指标、损失函数、优化器、激活函数等，构建网络模型，训练模型，抑制过拟合、画Loss和Acc or mae等图

特别是

1. 数据的预处理，张量一般放缩到较小的区间，有利于训练和收敛，不同特征的要做标准化
2. 模型要达到过拟合，至少要保证模型的表达能力是足够的，再去想办法抑制过拟合
3. 模型的参考基准：要比猜的准，例如mnist数据集，精度起码要大于0.1(10类中猜一类)，imdb数据集，精度起码要大于0.5(2类中猜一类)
4. 训练模型要防止数据泄露，不然会降低模型的可靠性