机器学习总结

机器学习概念

机器学习所研究的主要内容是关于在计算机上从“数据 ”中产生“模型” 的算法，即关于“算法” 的学问。

机器学习是使机器具备智能的过程。

通过机器学习的算法研究及其与具体问题的恰当结合，获得合适的模型。

机器学习具备三个条件：

1. 系统中可能存在模式

2. 这种模式不是一般解析手段可以猜测得到的

3. 数据可以获取

一般步骤：输入、算法、输出、评价

任务：

分类、回归、数据生成、结构化预测、知识获取→相关性知识

研究内容：有监督学习、无监督学习、半监督学习、强化学习

数据

数据分类

训练集：参数

验证集

- 超参数
- 模型选择
- 模型偏好
- 正则化

测试集：模型泛化能力

如何划分数据集？

- 预先手工划分
- 交叉验证
- 采样（放回采样，不放回采用）

预处理

如何处理噪声数据

- 计算机和人工检查结合

  计算机检测可疑数据，然后对它们进行人工判断、效率较低

- 回归：通过让数据适应回归函数来平滑数据
- 聚类：监测并且去除孤立点
- 分箱：按照属性值划分子区间

处理缺失项，可行的方法：

- 人工填写空缺值：工作量大
- 使用默认值
- 在所有样本上，使用属性的平均值填充空缺值
- 使用与给定元组属同一类的所有样本的平均值填充空值

数据规模不足如何处理？

- 研究小规模数据处理模型
- 数据扩展：生成伪样本 GAN VAE
- 标注新数据
- 迁移学习

数据集不均衡如何处理？（类别不平衡问题）

- 欠采样
- 过采样
- 预测函数修正（缩放策略）

模型

什么是模型？

- 解决问题方法
- 抽象函数表达式
- 解决问题方法的形式化描述

过拟合、欠拟合如何判断及处理

过拟合：模型学习能力太强，以至于将训练集单个样本自身的特点都能捕捉到，并将其认为是"一般规律"，过分依赖训练数据

过拟合模型表现为在训练集上具有高方差和低偏差

过拟合往往能较好地学习训练集数据的性质，而在测试集上的性能较差

欠拟合：模型学习能力较弱，**对于训练样本的一般性质尚未学好，**未能学习训练数据中的关系

欠拟合模型表现为在训练集上具有低方差和高偏差

欠拟合在训练集和测试集上的性能都较差

解决模型过拟合的方法

- 正则化：L1 和 L2
- Bayes 方法

- 数据扩增，即增加训练数据样本 增加训练数据量（5~10）
- Dropout：每次随机忽略隐层的某些节点
- Early stopping：一种迭代次数截断的方法来防止过拟合

解决模型欠拟合方法：

- 增加模型复杂度
- 添加其他特征项
- 添加多项式特征
- 减少正则化的程度

CNN

CNN 是一种深层神经网络模型，包含了一个由卷积层和子采样层构成的特征抽取器，可以直接处理输入的数据，适用于处理图像任务。

特殊性：相邻神经元间的连接是非全连接；同一层某些神经元之间的连接的权重是共享的。

三大特性：

- 局部感受野

每个神经元不和上一层的所有神经元相连，能够减少模型中的参数

- 权值共享
- 下采样

使用 pooling 减少每层的样本数，进一步减少参数数量，同时提高模型的鲁棒性。

一个卷积神经网络有若干卷积层、池化层、全连接层、输入层、输出层等基本构件组成。

卷积层：使用卷积核进行特征提取，每个卷积核同输入数据进行卷积运算，形成新的特征“图”

池化层：用于特征降维，压缩数据和参数的数量，减少过拟合、提高

模型的容错性。

归一化层：加速训练，提高精度

切分层：学习多套参数，更强特征描述能力

融合层：对单独学习的分支进行融合，构建高效而精简的特征组合

额外功能：

- 非线性激励：卷积是线性运算，增加非线性描述能力
- 降维：特征图稀疏，减少数据运算量，保持精度
- 归一化：特征的 scale 保持一致
- 区域分割：不同区域进行独立学习
- 区域融合：对分开的区域合并，方便信息融合
- 增维：增加图片生成或探测任务中空间信息

CNN 核心特点

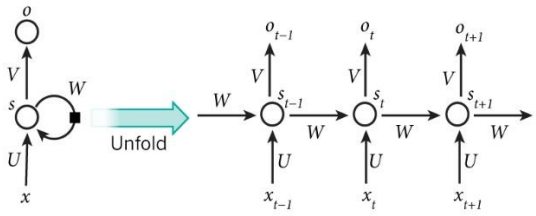
- 局部->整体，低层次的特征->组合->组成高层次特征
- 局部连接/权值共享/池化操作/多层次结构

训练的难点：

- 输入需要归一化大小
- 卷积核大小
- 激活函数的选择
- 过拟合

RNN

RNN 是一种对序列数据建模的神经网络，即一个序列当前的输出与前面的输出也有关。具体表现形式为：网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点是有链接的，并且输入不仅包括输入层的输入还有上一时刻隐藏层的输出。



训练时，超参数如何确定

超参数选择

Grid Search：在高维空间中对一定区域进行遍历

Random Search：随机在高维空间中选择若干超参数

如何选择合适的学习率

Fixed：固定学习率

Step：采用均匀降低的方式

AdaGrad：自适应学习率，只需要设定一个全局的学习率

RMSporp：在 AdaGrad 基础上，对学习率改进，每回合学习速率都有一定比例的衰减，衰减系数 r

Adam：带有 Momentum 动量项的 RMSProp，它利用梯度的一阶矩估计和二阶矩估计动态调整每个参数的学习速率。为不同的参数计算不同的自适应学习速率。

mini-batch 的选择

太小会使训练速度很慢；太大会加快训练速度，但同时会导致内存占用过高，并有可能降低准确率。

所以 32 至 256 是不错的初始值选择，尤其是 64 和 128，选择 2 的指数倍的原因是：计算机内存一般为 2 的指数倍，采用 2 进制编码。

梯度爆炸解决：梯度裁剪

梯度消失解决：

合理初始化权重

选择合适的激活函数，如 ReLu

使用 LSTM(Long Short Term Memory)或 GRU (Gated Recurrent Unit)

梯度消失的原因：在多层网络中，影响梯度大小的因素主要有两个：权重和激活函数的偏导。深层的梯度是多个激活函数偏导乘积的形式来计算，如果这些激活函数的偏导比较小（小于 1 ）或者为 0，那么梯度随时间很容易 vanishing；相反，如果这些激活函数的偏导比较大（大于 1），那么梯度很有可能就会 exploding 。因而，梯度的计算和更新非常困难。

解决方案：使用一个合适激活函数，它的梯度在一个合理的范围。

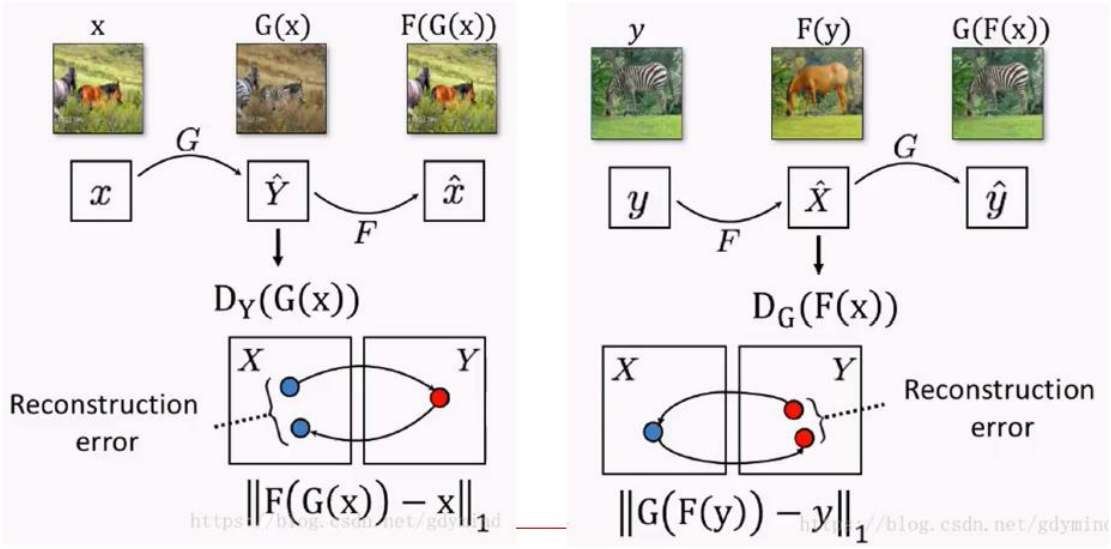
LSTM 使用 gate function，有选择的让一部分信息通过。gate 是由一个 sigmoid 单元和一个逐点乘积操作组成，sigmoid 单元输出 1 或 0，用来判断通过还是阻止，然后训练这些 gate 的组合。所以，当 gate 是

打开的（梯度接近于 1），梯度就不会 vanish。并且 sigmoid 不超过 1，那么梯度也不会 explode。

Cycle GAN

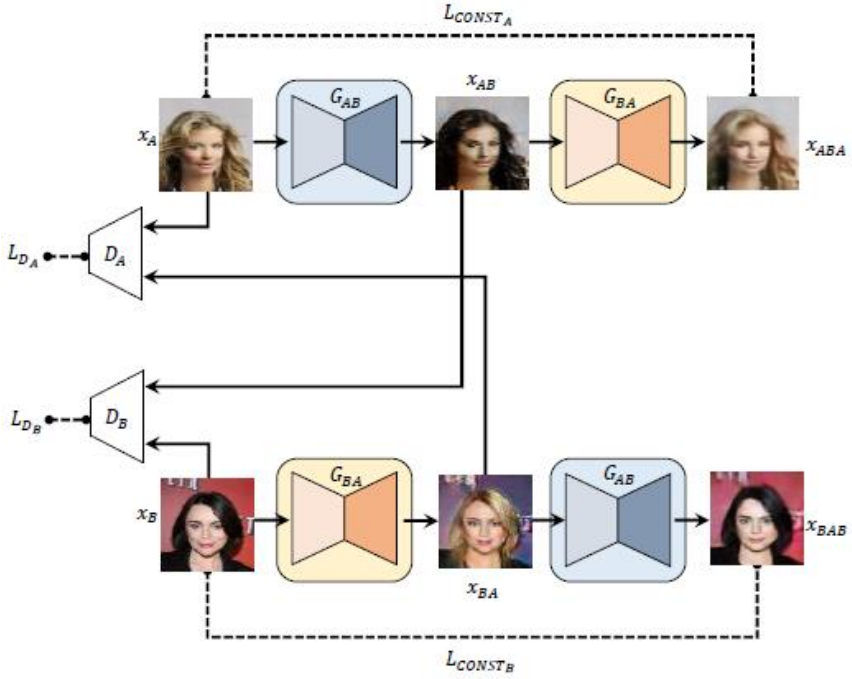
本质：两个镜像对称的 GAN，构成一个环形网络

两个 GAN 共享两个生成器，各自一个判别器



Disco GAN 发现跨域关系

用一种风格图片生成另一种风格



Variational Auto Encoder VAE

本质：为每个样本构造专属正态分布，然后采用重构

避免模型退化

提高重构精度→方差趋于 0→随机性减少→模型得到确定结果

任务 考虑因素

数据来源、规模、预处理、分类

问题类型：分类、回归

实验环境

模型选择

参数选择

性能评估

在工业应用场景中，面对一个要解决的问题（分类、回归或者结构化预测），在给出解决方案前，你会考虑哪些因素？

针对数据：

1. 该领域提供的历史数据集容量大小
2. 相关数据的可区分性
3. 相关数据是否为时序相关的
4. 数据的质量如何（有无噪声数据或者数值缺失情况）
5. 给定历史数据中是否存在私密信息
6. 数据是否符合某些已知的分布

针对应用场合：

1. 考虑项目的背景，分析希望算法的准确率高还是召回率高
2. 考虑目前的设备情况，尽量选择设备能够承受的计算复杂算法
3. 查阅文献，是否有应用于相关领域的较好的算法

准备数据

模型设计

训练细节

一般的处理过程是：

1）获取数据；

2）提取最能体现数据的特征；

3）利用算法建模；

4）将建立的模型用于预测。