

神经网络与深度学习

[本页PDF](#)

机器学习

机器学习三要素：模型、学习准则、优化

常见三类机器学习问题：监督学习、无监督学习、强化学习

随机梯度下降法：每次迭代只选择一个样本来更新网络参数

- 输入训练集、验证集、学习率 α
- 随机初始化
- 循环直到模型在验证集的错误率不再下降（或小于某一个阈值）
 - 对训练集样本进行随机排序
 - 遍历每一个样本
 - 根据选取的样本更新参数
- 输出模型参数 θ

机器学习模型的几个评价指标

预测结果为 \hat{y} ，真实结果为 y

- 准确率：预测正确的样本占总样本比例

$$\mathcal{A} = \frac{1}{N} \sum_{n=1}^N I(y^{(n)} = \hat{y}^{(n)})$$

- 错误率：预测错误的样本占总样本比例

$$\begin{aligned}\mathcal{E} &= 1 - \mathcal{A} \\ &= \frac{1}{N} \sum_{n=1}^N I(y^{(n)} \neq \hat{y}^{(n)})\end{aligned}$$

- 精确率：预测为正例占真正是正例的比例

$$P = \frac{TP}{TP + FP}$$

- 召回率：真是正例占被模型预测为正例的比例

$$R = \frac{TP}{TP + FN}$$

精确率和召回率详见[统计学习方法](#)

前馈神经网络

反向传播算法执行过程

- 输入训练集，验证集，学习率，正则化系数等

- 随机初始化网络参数
- 循环直到模型在验证集的错误率不再下降（或小于某一个阈值）
 - 对训练集样本进行重排序
 - 遍历每一个样本
 - 前馈计算每一层的净输入和激活值
 - 反向传播计算每一层梯度
 - 更新参数
- 输出模型参数

CNN

RNN

长程依赖问题：循环神经网络在时间维度上非常深，包括梯度消失和梯度爆炸，可以使用**循环边改为线性依赖关系和增加非线性**来改进。

- 梯度消失：网络层数很深且导数小于1，经过每一层网络后导数不断衰减，造成梯度衰减
 - 改进模型
- 梯度爆炸：网络层数很深且导数大于1，经过每一层网络后导数不断被放大，造成梯度爆炸
 - 使用权重衰减或梯度裁剪

网络优化与正则化

神经网络优化的改善方法有**网络优化和网络正则化**两大类

小批量梯度下降：每次选取K个样本对模型参数进行更新，影响其的主要因素有

- 批量大小K
- 学习率
- 梯度估计

参数初始化方法

- 预训练初始化
- 随机初始化
 - 将参数都初始化为0导致第一遍前向计算时中间值都相同，导致反向传播权重更新也相同，导致神经元没有区分性，称为**对称权重现象**
- 固定值初始化

网络正则化方法

- L1、L2正则化
- 权重衰减
- 早停
- Dropout
- 数据增强
- 标签平滑

不能在循环神经网络中的循环连接上直接应用丢弃法：这样会损害循环网络在时间维度上记忆能力，导致信息流断裂，模型无法理解长序列依赖关系

L1、L2正则化目标函数

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^N L(y^{(n)}, f(x^{(n)}; \theta)) + \lambda l_p(\theta)$$

神经网络优化的改善方法

- 使用更有效的优化算法来提高梯度下降优化方法的效率和稳定性，比如动态学习率调整、梯度估计修正
- 使用更好的参数初始化方法、数据预处理方法来提高优化效率
- 修改网络结构来得到更好的优化地形，比如使用ReLU激活函数、残差连接、逐层归一化等
- 使用更好的超参数优化方法

RMSProp

维护梯度平方的累积变量 G_t ，梯度为 g_t

$$G_t = \beta G_{t-1} + (1 - \beta) g_t^2$$
$$\theta_t = \theta_{t-1} - \frac{\alpha}{\sqrt{G_t} + \epsilon} g_t$$

Momentum

维护一个速度变量 v_t

$$v_t = \beta v_{t-1} + (1 - \beta) g_t$$
$$\theta_t = \theta_{t-1} - \alpha v_t$$

Adam算法

- 结合了RMSProp和动量法

计算移动平均（一阶矩和二阶矩）

$$M_t = \beta_1 M_{t-1} + (1 - \beta_1) g_t$$
$$G_t = \beta_2 G_{t-1} + (1 - \beta_2) g_t \odot g_t$$

偏差修正

$$\hat{M}_t = \frac{M_t}{1 - \beta_1^t}$$
$$\hat{G}_t = \frac{G_t}{1 - \beta_2^t}$$

更新参数

$$\Delta \theta_t = - \frac{\alpha}{\sqrt{\hat{G}_t} + \epsilon} \hat{M}_t$$

逐层归一化

- 批量归一化（Batch Normalization）：不同样本的同一维度上进行归一化
- 层归一化（Layer Normalization）：同一样本的不同维度进行归一化
- 权重归一化
- 局部响应归一化

无监督学习

典型的无监督学习问题分类

- 无监督特征学习
- 密度估计
- 聚类

主成分分析

- 优化函数:

第一主成分优化问题

$$\max_{w_1} w_1^T \Sigma w_1 \text{ s.t. } w_1^T w_1 = 1$$

第k主成分优化问题

\$\$

$$\begin{aligned} \max_{w_k} \quad & w_k^T \Sigma w_k \\ \text{s.t.} \quad & w_k^T w_k = 1, \quad w_k^T w_j = 0 \quad (j = 1, \dots, k-1) \end{aligned}$$

\$\$

- 求解过程: 使用拉格朗日乘子法, 对 w_1 求导, 令导数为0

线性编码

- 表达式
 - 输入样本 x 可以由基向量 A 和字典 Z 表示
 - 其中, 基向量 $A = [a_1, \dots, a_m]$, 编码向量是 $Z = [z_1, \dots, z_m]^T$

解码

$$x = \sum_{m=1}^M z_m a_m = AZ$$

编码

$$z = \sum_{m=1}^M x_m a_m = A^T x$$

稀疏编码

- 目标函数
 - 其中 ρ 是稀疏性衡量函数, η 是一个超参数, 用来控制稀疏性的强度

$$\mathcal{L}(A, Z) = \sum_{n=1}^N (\|x^{(n)} - Az^{(n)}\|^2 + \eta \rho(z^{(n)}))$$

- 训练过程

1. 固定基向量 A ，对于每个输入 $x^{(n)}$ ，计算其对应的最优编码

$$\min_{z^{(n)}} \|x^{(n)} - Az^{(n)}\|^2 + \eta\rho(z^{(n)}), \forall n \in [1, N]$$

2. 固定上一步得到的编码 $\{z^{(n)}\}_{n=1}^N$ ，计算其最优的基向量

$$\min_A \sum_{n=1}^N (\|x^{(n)} - Az^{(n)}\|^2) + \lambda \frac{1}{2} \|A\|^2$$

编码器

自编码器

- 分为编码器和解码器
- 学习目标是重构错误（最小化重构误差）
- 和主成分分析在学习目标上完全一致：最小化重构误差
 - 在特定条件下，两者是同构的
 - 自编码器是主成分分析的**非线性泛化**

稀疏自编码器

- 目标函数
 - 其中 W 是自编码器中的参数
 - 不仅要求还原输入，还需要学习到稀疏且有意义的特征表示，且引入正则化进行参数惩罚

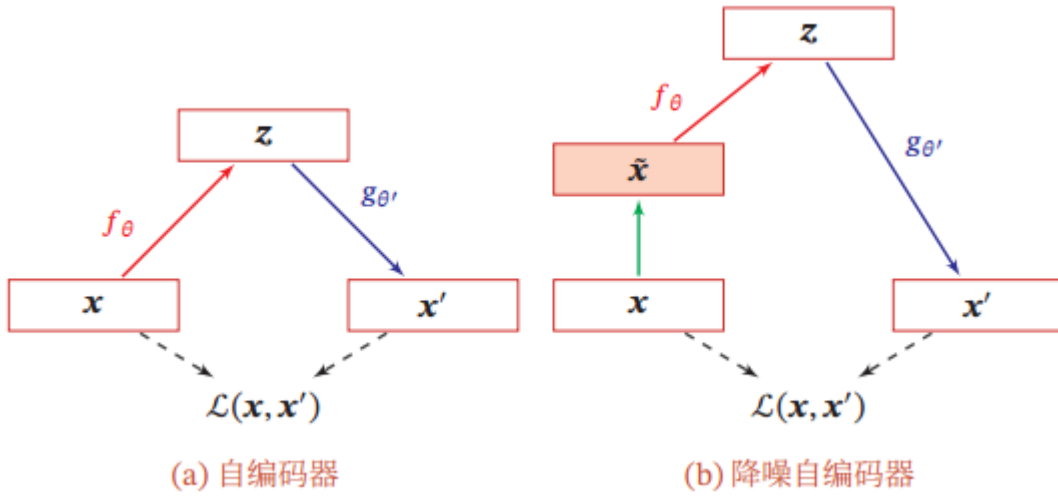
$$\mathcal{L} = \sum_{n=1}^N \|x^{(n)} - x'^{(n)}\|^2 + \eta\rho(z^{(n)}) + \lambda \|W\|^2$$

堆叠自编码器

- 将编码器进行逐层堆叠，训练一个深度的自编码器
- 采用**逐层训练**的方式更新参数

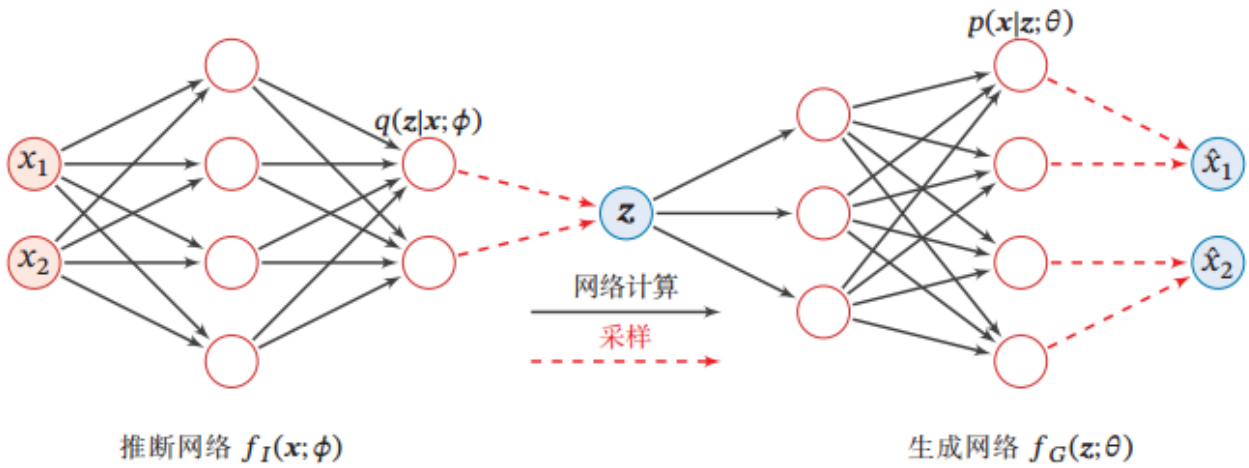
降噪自编码器

- 思想：通过引入噪声来学习更鲁棒性的数据编码，并提高模型的泛化能力
- 结构：在原始输入 x 中引入噪声，交由编码器还原
- 学习目标：要求网络能够对输入样本进行去噪，从含有噪声的样本中恢复原始数据



变分自编码器

- 结构：



- 目标函数：最小化KL散度

$$\phi^* = \arg \min_{\phi} \text{KL}(q(z|x; \phi), p(z|x; \theta))$$

模型独立的学习方法

- 集成学习
- 协同学习
- 多任务学习
- 迁移学习
- 终身学习
- 元学习

总结

:::tabs

@tab:active 第二章

机器学习三要素，常见三类机器学习问题

随机梯度下降法步骤

机器学习几个评价指标

反向传播算法执行过程是什么

什么是长程依赖问题、梯度消失、梯度爆炸，分别如何改进

神经网络优化的改善方法有哪些

什么是小批量梯度下降，影响其的主要因素有哪些

参数初始化方法有哪些，为什么神经网络不能初始化参数为0，会产生什么现象

网络正则化方法有哪些，L1、L2正则化目标函数

为什么不能在循环神经网络中的循环连接上直接应用丢弃法

神经网络优化的改善方法有哪些

几种常用的逐层归一化方法

@tab 第三章 典型的无监督学习问题可以分为哪几类

模型独立的学习方法有哪些

线性编码表达式 稀疏编码目标函数和训练过程