

预训练学习报告（1）

汪可予 4.22

1、机器学习理论基础

大一下学期阅读了李航的统计学习方法，当时有不少没有消化的地方，在大二寒假时，配着斯坦福 CS229 的讲义把统计学习方法又过了一遍。

统计学习方法第 1 章介绍了机器学习的基础知识，下面是我对一些重要概念的理解：

模型：是如何做决策的问题，常常表示为“输入空间到输出空间的映射”

策略：定义好的模型的准则——常常是损失函数

算法：最小化损失函数的算法，比如梯度下降等

泛化：要理解泛化，首先要理解过拟合和欠拟合。过拟合是模型在训练集上效果好，在测试集上效果差，常常是由于模型过于复杂导致的；欠拟合是模型在训练集、测试集效果都不好，这常常是因为选择的模型过于简单或者不合适导致的。我们希望我们的模型泛化能力强，即模型在训练集和测试集上的表现都很好。在测试集上的训练误差叫泛化误差

模型选择：我们如何从几个模型中选择最合适的模型？这就是模型选择问题。常用方法：交叉验证，如 k 折交叉验证，留一交叉验证等。在《Speech and Language Processing》中也有介绍这个，不过它哪里验证集名称叫“development set”，我当时纠结了很久才看懂。我在实践中发现交叉验证方法的选择是有很多问题需要思考的，需要考虑数据集的特点（有偏性、数据量等），不过这已经超出本节介绍的理论基础范畴了。

统计学习方法第 2 章介绍了感知机模型，这是神经网络的基础之一。它是一个简单的二分类模型，可以解决线性可分问题。同时这章介绍了简单的优化算法。

二、python 学习

大一暑期完成了 python 基础语法的学习以及 Sklearn 的基本调包使用。把基础模型打了一遍代码，跑过一些简单的数据集，熟悉了基本的数据处理、机器学习解决问题的流程。基础的原理和常用的代码整理成博客，方便日后复习/使用。

https://blog.csdn.net/hhhenjoy/category_11329710.html?spm=1001.2014.3001.5482



机器学习实战 代码笔记
111
关注数: 0 文章数: 8
[管理文章](#)

```
kmeans.cluster_centers_# 每个实例到中心点距离 (数据类)kmeans.transform(X)# 训练实例的标签到kmeans.labels_中心点初始化方法如果事先知道了...
```

原创 2021-09-22 15:08:46 · 37 阅读 · 0 评论

【机器学习实战】Ch 8: 降维

降维的主要方法是将高维空间的点投影到低维空间上流形学习d维流形是n维空间（且d<n）的一部分，局部类似于d维超平面。许多降维算法通过训练实例所在的流形进行建模工作，称为流形学习。它依赖于流形假设：大多数现实世界的离群数据都集中在低维流形。隐式假设：如果用流形的低维空间来...

原创 2021-09-22 09:41:40 · 43 阅读 · 0 评论

【机器学习实战】Ch 7: 集成学习和随机森林

投票分类器假设已经训练好了一些分类器，每个分类器准确率为80%。这时，要创建一个更好的分类器，最简单的办法是聚合每个分类器的预测。然后将得票最多的结果作为预测类别。这种大多数投票分类器被称为投票分类器。如下用Sklearn创建一个投票分类器，由三种不同的分类器...

原创 2021-09-19 17:16:35 · 71 阅读 · 0 评论

【机器学习实战】Ch 6: 决策树

训练决策树Sklearn使用分类和回归树算法来训练决策树工作原理：使用单个特征k和阈值k将训练集分为两个子集，通过选择（k, b）最小化成本函数搜索最佳子集。并重复该过程，直到到达最大深度或达到减少不纯度的分割，停止递归。其他一些超参数也可以控制停止条件（稍后叙述）from...

原创 2021-09-19 12:28:05 · 41 阅读 · 0 评论

【机器学习实战】Ch 5: 支持向量机

线性SVM分类SVM分类器在类之间拟合可能的最宽间隔软间隔分类，超参数C越小，越容易欠拟合，间隔越宽拟合越多，但泛化效果可能更好；C越大，越容易过拟合，间隔越窄拟合越少，泛化效果更差。# detect virginica irisimport numpy as npfrom sklearn import datasetsfrom sklearn.pipeline import...

原创 2021-09-19 09:36:21 · 55 阅读 · 0 评论

三、深度学习基础和 pytorch 学习

大二上学期阅读了邱锡鹏老师的《神经网络与深度学习》，学习了前馈神经网络、卷积神经网络、循环神经网络、注意力机制等等。第一遍阅读略过了数学推导部分，除了大一暑假的时候曾手推过前馈神经网络的反向传播算法，其它的网络反向传播等等数学推导都还没看。邱老师还有其他一些书籍在网络优化一块儿有部分内容还未学懂，以后会抽时间去细看。学习了深度基础的基础知识后，我试着从零用 python 实现了一些简单的神经网络，也学习了用 pytorch 实现神经网络的一些模块。（参考资料李沐《动手深度学习》电子书+视频）。也有写过一些博客，记录着当时幼稚的想法和代码

<https://www.cnblogs.com/kyfishing/category/2096817.html>

本月我仔细阅读了一遍 pytorch 的官方文档，尤其是 Text 部分。部分内容有点吃力，计划在后期实战巩固。

下面是我挑了一些神经网络的重要概念理解：

神经元：神经网络的基本单位，常常是线性感知机

正向传播、反向传播和计算图：正向传播是由输入计算输出的过程，反向传播用于计算梯度，更新参数，计算图可以用于可视化正向传播和反向传播的过程

激活函数：保证每一层的计算是非线性的，这样神经网络层的堆叠才是有意义的

权重衰减、暂退法：用于防止过拟合的方法

卷积神经网络：多用于视觉领域，卷积层和池化层是其中比价重要的概念

循环神经网络、注意力机制：多用于文本领域，不过像 transformer 在近年来视觉也有广泛应用

四、面向 NLP 的深度学习、transformer

大二上学期读了《Speech and Language Processing》，在 6、7、8、9 章对面向 NLP 的深度学习有了初步的了解。在下学期调王然学姐当时他们类别推断的模型有了更深入的了解。寒假学习了 transformer 的基础架构原理。

这是当时读《Speech and Language Processing》的部分笔记：

<https://www.cnblogs.com/kyfishing/category/2050393.html>

五、未来计划：阅读 transformer 源码，学习 transformer 使用

综上，完整的学习路线如下。（不同基础的同学可以适当跳过某些步骤）

