《从自然语言处理入门机器学习》

系列精品课程

开课背景：

为了帮助零基础的学员系统性地学习机器学习，成为 BAT 的机器学习工程师，特开设本课程。

一门来自 BAT 讲师团队的机器学习课程，讲师来自百度、阿里巴巴等。

为什么要从自然语言处理入门机器学习：

机器学习必须和具体的数据类型、应用场景结合。由于文本数据处理相对于语音和视频图像要容易一点，加上互联网累积的文本数据是最丰富的，因此，目前超过半数的机器学习工程师都在做自然语言处理。

课程说明：

1）每章节安排授课时间两课时。

2）讲师课后答疑，助教督促学员完成作业

3）本课程以实践为主轴，打通概念的讲述，但并非纯实战课程，而是理论体系比较完备、又注重实战的课程

【初级班】

课程目标：

能够完成基于朴素贝叶斯的、逻辑回归的中文垃圾短信识别，掌握分词、NLP 各种工具包、常见的 linux 命令，理解线性回归、逻辑回归、朴素贝叶斯、最大熵、BP 神经网络等基本模型

课程说明：

文本的特征表示、模型评估问题，会在实战过程中介绍

适合学员：

对自然语言处理零基础、刚接触机器学习的学员

课程安排：

第一章：机器学习与自然语言处理基础

机器学习基本概念（有监督、无监督、分类、聚类、回归等）

自然语言处理简介

Python 基础

jieba 分词

Linux 基本命令

轻松一刻：聊聊行业大师

第二章：自然语言的特征表示与语言模型简介

文本的特征表示：词袋特征、tf-idf 特征、ngram 特征、词粒度等

主题模型、词嵌入简介

统计语言模型

【实战】基于语言模型的乱序句子重建实验

第三章：线性回归和梯度下降

线性回归模型

梯度下降法(BGD/SGD/online GD)

过拟合、正则化

数据抽样算法

【实战】实现线性回归模型，预测波士顿地区房价（手写，不掉包）

【实战】实现线性回归模型，预测波士顿地区房价（调包实现）

第四章：逻辑回归模型与二分类实践

Logistic 函数的由来、在神经网络中的应用

逻辑回归模型原理与训练

广义线性模型

【实战】实现基于逻辑回归的垃圾短信分类（手写，不掉包）

Python 工具包介绍(numpy scipy sklearn)

【实战】实现基于逻辑回归的垃圾短信分类（通过 sklearn 实现）

初级班期中考试

第五章：朴素贝叶斯模型与贝叶斯网络

贝叶斯概率论基础

朴素贝叶斯模型

朴素贝叶斯模型在 NLP 中的应用

生成模型与判别模型

贝叶斯网络

【实战】基于朴素贝叶斯的垃圾短信分类（手写，不掉包）

【练习】完成基于朴素贝叶斯的 kaggle 影评数据情感分析

第六章：最大熵模型原理与凸优化基础

信息熵系列概念介绍

交叉熵与机器学习模型损失的评估方法

凸优化和拉格朗日对偶、KKT 条件的证明

最大熵模型

最大熵模型在 NLP 中的应用

【实战】基于最大熵模型的文本分类

第七章：神经网络基础与初探

计算图与链式法则

神经网络基本概念

多层前馈网络和 BP 算法

【实战】基于 MLP 神经网络的垃圾短信分类

第八章：初级班练习与实战

【习题课】讲解一些关于基本概念的题目

【实战】使用初级班所学模型，解决酒店评论情感分析问题

初级班结业考试

学习资料：

统计自然语言处理基础

统计学习方法

【中级班】

课程目标：

能够完成基于更高级特征、其他传统机器学习模型、更丰富的 NLP 工具和模型训练工具完成中文垃圾短信识别，特征包括“字粒度 vs 词粒度、词袋 vs 词嵌入、主题模型 vs word2vec”，传统机器学习模型包括 knn、svm、决策树、gbdt，NLP 工具包括 fasttext、glovec、word2vec、spacy、textblob，模型训练工具包括 xgboost、libsvm、liblinear、weka、sklearn、lightgbm。

适合学员：

已修完初级班的学员

课程说明：

中级课程仍然仅限于传统机器学习模型，不会涉及到深度学习模型

相较于初级课程，中级课程内容比较饱满，课后还需要多加练习巩固。

课程安排：

第一章：KNN 与 SVM

KNN 算法

Kd 树

SVM：线性可 SVM/Soft-margin SVM/kernel SVM

[实战]基于 KNN 的垃圾短信分类

[实战]基于 SVM 分别通过 sklearn 和 libsvm 工具，解决垃圾短信分类问题

第二章：决策树与模型融合

决策树模型

模型融合(bagging/boosting)

随机森林

GBDT

【实战】在 sklearn 环境下，基于 ngram 和随机森林的垃圾短信分类

第三章：主题模型

LSA

PLSA

LDA

【实战】主题模型在垃圾短信分类中的应用

第四章：词嵌入

Distributed representation 的理论原理

Word2vec 与 Gensim

Glovec

Fasttext

【实战】基于 fasttext 的垃圾短信分类

第五章：模型训练工具介绍与实操

Weka

Liblinear

Libsvm (复习)

Xgboost

Sklearn（复习）

lightgbm

第六章：垃圾短信分类进一步实战与各种方法效果对比

【实战】使用 Xgboost 完成基于 ngram 和 GBDT 模型的垃圾短信分类

各种方法效果的对比与小节

第七章：企业实战模拟

【实战】在训练语料无标注的条件下，如何基于规则来做垃圾短信分类

第八章：强化与考试

【习题课】讲解一些关于基本概念的题目

文本分类问题一般流程的总结

下次课程预告

中级班结业考试(在线答题与实战考核结合)

【高级班】

课程目标：

基于深度学习和 tensorflow 的自然语言处理。会介绍 tensorflow 的使用、深度学习的理论、以及深度学习在自然语言处理中的应用

第一章：深层神经网络简介

深度学习简介

基本概念：激活函数 dropout 交叉熵等

第二章：Tensorflow 基本概念和架构

Tensor graph op session 等基本概念

Tensorflow 源码结构 master-worker 模式等

第三章：Tensorflow 第一步

【代码解析】基于 tensorflow 的手写数字识别（mnist）

第四章：CNN 与自然语言处理

CNN 原理

CNN 处理文本分类的过程

第五章：CNN 文本分类实战

【代码解析】Tensorflow 实例讲解：“Kaggle San Francisco Crime Description”数据集

【实战】基于 cnn 的垃圾短信分类

第六章：RNN 与自然语言处理

RNN 原理

LSTM/GRU

RNN 在自然语言处理中的应用

第七章：RNN 文本分类实战

【代码解析】Tensorflow 实例讲解

【实战】基于 rnn 的垃圾短信分类

普通 RNN vs lstm vs gru

单向 vs 双向

第八章：构建基于 tensorflow serving 的实时服务

tensorflow serving 简介

【实战】使用 tensorflow serving 搭建手写数字识别实时服务（mnist）

【实战】打造基于 Tensorflow serving 和双向 GRU 的垃圾短信识别实时服务

【复习课】

【习题课】讲解一些关于基本概念的题目

自然语言处理的一些面试题

【练习】用 RNN 模拟写诗

课程咨询：

这么饱满的课程内容，该怎么报名呢？

快来联系一下咨询老师吧 ：

QQ 咨询

微信咨询