链接：https://pan.baidu.com/s/1bmlTu\_rzTLZ7-S43mQmXkg

密码：2fv0

课程大纲：

第一课：机器学习的数学基础1 - 数学分析

1. 机器学习的一般方法和横向比较  
2. 数学是有用的：以SVD为例  
3. 机器学习的角度看数学  
4. 复习数学分析  
5. 直观解释常数e  
6. 导数/梯度  
7. 随机梯度下降  
8. Taylor展式的落地应用  
9. gini系数  
10. 凸函数  
11. Jensen不等式  
12. 组合数与信息熵的关系

第二课：机器学习的数学基础2 - 概率论与贝叶斯先验

1. 概率论基础  
2. 古典概型  
3. 贝叶斯公式  
4. 先验分布/后验分布/共轭分布  
5. 常见概率分布  
6. 泊松分布和指数分布的物理意义  
7. 协方差(矩阵)和相关系数  
8. 独立和不相关  
9. 大数定律和中心极限定理的实践意义  
10. 深刻理解最大似然估计MLE和最大后验估计MAP  
11. 过拟合的数学原理与解决方案

第三课：机器学习的数学基础3 - 矩阵和线性代数

1. 线性代数在数学科学中的地位  
2. 马尔科夫模型  
3. 矩阵乘法的直观表达  
4. 状态转移矩阵  
5. 矩阵和向量组  
6. 特征向量的思考和实践计算  
7. QR分解  
8. 对称阵、正交阵、正定阵  
9. 数据白化及其应用  
10. 向量对向量求导  
11. 标量对向量求导  
12. 标量对矩阵求导

第四课：Python基础1 - Python及其数学库

1. 解释器Python2.7与IDE：Anaconda/Pycharm  
2. Python基础：列表/元组/字典/类/文件  
3. Taylor展式的代码实现  
4. numpy/scipy/matplotlib/panda的介绍和典型使用  
5. 多元高斯分布  
6. 泊松分布、幂律分布  
7. 典型图像处理  
8. 蝴蝶效应  
9. 分形与可视化

第五课：Python基础2 - 机器学习库

1. scikit-learn的介绍和典型使用  
2. 损失函数的绘制  
3. 多种数学曲线  
4. 多项式拟合  
5. 快速傅里叶变换FFT  
6. 奇异值分解SVD  
7. Soble/Prewitt/Laplacian算子与卷积网络  
8. 卷积与(指数)移动平均线  
9. 股票数据分析

第六课：Python基础3 - 数据清洗和特征选择

1. 实际生产问题中算法和特征的关系  
2. 股票数据的特征提取和应用  
3. 一致性检验  
4. 缺失数据的处理  
5. 环境数据异常检测和分析  
6. 模糊数据查询和数据校正方法、算法、应用  
7. 朴素贝叶斯用于鸢尾花数据  
8. GaussianNB/MultinomialNB/BernoulliNB  
9. 朴素贝叶斯用于18000+篇/Sogou新闻文本的分类

第七课： 回归

1. 线性回归  
2. Logistic/Softmax回归  
3. 广义线性回归  
4. L1/L2正则化  
5. Ridge与LASSO  
6. Elastic Net  
7. 梯度下降算法：BGD与SGD  
8. 特征选择与过拟合

第八课：Logistic回归

1. Sigmoid函数的直观解释  
2. Softmax回归的概念源头  
3. Logistic/Softmax回归  
4. 最大熵模型  
5. K-L散度  
6. 损失函数  
7. Softmax回归的实现与调参

第九课：回归实践

1. 机器学习sklearn库介绍  
2. 线性回归代码实现和调参  
3. Softmax回归代码实现和调参  
4. Ridge回归/LASSO/Elastic Net  
5. Logistic/Softmax回归  
6. 广告投入与销售额回归分析  
7. 鸢尾花数据集的分类  
8. 交叉验证  
9. 数据可视化

第十课：决策树和随机森林

1. 熵、联合熵、条件熵、KL散度、互信息  
2. 最大似然估计与最大熵模型  
3. ID3、C4.5、CART详解  
4. 决策树的正则化  
5. 预剪枝和后剪枝  
6. Bagging  
7. 随机森林  
8. 不平衡数据集的处理  
9. 利用随机森林做特征选择  
10. 使用随机森林计算样本相似度  
11. 数据异常值检测

第十一课：随机森林实践

1. 随机森林与特征选择  
2. 决策树应用于回归  
3. 多标记的决策树回归  
4. 决策树和随机森林的可视化  
5. 葡萄酒数据集的决策树/随机森林分类  
6. 波士顿房价预测

第十二课：提升

1. 提升为什么有效  
2. 梯度提升决策树GBDT  
3. XGBoost算法详解  
4. Adaboost算法  
5. 加法模型与指数损失

第十三课：提升实践

1. Adaboost用于蘑菇数据分类  
2. Adaboost与随机森林的比较  
3. XGBoost库介绍  
4. Taylor展式与学习算法  
5. KAGGLE简介  
6. 泰坦尼克乘客存活率估计

第十四课：SVM

1. 线性可分支持向量机  
2. 软间隔的改进  
3. 损失函数的理解  
4. 核函数的原理和选择  
5. SMO算法  
6. 支持向量回归SVR

第十五课：SVM实践

1. libSVM代码库介绍  
2. 原始数据和特征提取  
3. 葡萄酒数据分类  
4. 数字图像的手写体识别  
5. SVR用于时间序列曲线预测  
6. SVM、Logistic回归、随机森林三者的横向比较

第十六课：聚类（上）

1. 各种相似度度量及其相互关系  
2. Jaccard相似度和准确率、召回率  
3. Pearson相关系数与余弦相似度  
4. K-means与K-Medoids及变种  
5. AP算法(Sci07)/LPA算法及其应用

第十七课：聚类（下）

1. 密度聚类DBSCAN/DensityPeak(Sci14)  
2. DensityPeak(Sci14)  
3. 谱聚类SC  
4. 聚类评价AMI/ARI/Silhouette  
5. LPA算法及其应用

第十八课：聚类实践

1. K-Means++算法原理和实现  
2. 向量量化VQ及图像近似  
3. 并查集的实践应用  
4. 密度聚类的代码实现  
5. 谱聚类用于图片分割

第十九课：EM算法

1. 最大似然估计  
2. Jensen不等式  
3. 朴素理解EM算法  
4. 精确推导EM算法  
5. EM算法的深入理解  
6. 混合高斯分布  
7. 主题模型pLSA

第二十课：EM算法实践

1. 多元高斯分布的EM实现  
2. 分类结果的数据可视化  
3. EM与聚类的比较  
4. Dirichlet过程EM  
5. 三维及等高线等图件的绘制  
6. 主题模型pLSA与EM算法

第二十一课：主题模型LDA

1. 贝叶斯学派的模型认识  
2. Beta分布与二项分布  
3. 共轭先验分布  
4. Dirichlet分布  
5. Laplace平滑  
6. Gibbs采样详解

第二十二课：LDA实践

1. 网络爬虫的原理和代码实现  
2. 停止词和高频词  
3. 动手自己实现LDA  
4. LDA开源包的使用和过程分析  
5. Metropolis-Hastings算法  
6. MCMC  
7. LDA与word2vec的比较  
8. TextRank算法与实践

第二十三课：隐马尔科夫模型HMM

1. 概率计算问题  
2. 前向/后向算法  
3. HMM的参数学习  
4. Baum-Welch算法详解  
5. Viterbi算法详解  
6. 隐马尔科夫模型的应用优劣比较

第二十四课：HMM实践

1. 动手自己实现HMM用于中文分词  
2. 多个语言分词开源包的使用和过程分析  
3. 文件数据格式UFT-8、Unicode  
4. 停止词和标点符号对分词的影响  
5. 前向后向算法计算概率溢出的解决方案  
6. 发现新词和分词效果分析  
7. 高斯混合模型HMM  
8. GMM-HMM用于股票数据特征提取