本课程学习要求：

1. 知识准备：

线性代数
多元分析
基本概率推理（概率不等式，马氏过程等）
基本数理统计（多元回归，多元正态分布，中心极限定理等）
优化(凸优化)
编程能力：R　　多数统计软件包在该环境下运行
Matlab 优化和稀疏矩阵处理能力优异
*C/Python等，需要时候非结构化数据预处理

2. 要花费大量时间。如果没有大量的时间和精力花费在这门课上，请退课。

因为
(1) 需要大量的时间整理数据、编程、调试程序、分析结果。不愿意分析实际数据的同学请退课。

(2) 需要大量的时间阅读相关文献。该课会讲很多机器学习的方法。但是很多方法只给出参考文献和简单介绍。剩下的需要自己去读文章，一些文章需要在课堂讨论。 如果没有时间读文章，请退课。否则本课程学完后，你什么也学不到。仅会抱怨老师上课没有教到位。 我们只是领进门。

3. 本学期大概15周课程如下安排

(1) 每次上课3个小时，前两个小时老师讲解。后一个小时，学生做报告。每次报告有加分。
不定期作业，包含很多projects;
期末 final project，没有考试。
考核分为平时成绩和期末论文。

(2) 上课内容包括：

A. 有监督学习 (Supersized Learning)

回归分析 (Regression)

判别分析 (Classification)

B. 无监督学习 (Unsupervised Learning)

聚类 (Clustering), Density Estimation, Matrix Factorization (last term)

C. 半监督学习 (Semi-supervised Learning)

含有缺失数据

D. 在线学习 (online learning or recursive methods)
序列数据

E. 网络数据分析
social network

(3) 15次课程安排：

project 贯穿本学期，第一次由姚远讲讲什么是机器学习，以及机器学习能干什么。然后介绍一个实际问题。

在课程中，我们会以如下正在研究的实际问题为线索，围绕这些问题的讲统计学习方法：

(1) 计算广告中的问题；
(2) 统计排序中的问题；

(3) 蛋白质结构的问题；
(4) Twitter, 新浪微博数据分析问题；

上学期我们是讲理论方法，然后找实际数据验证方法；这学期我们是从问题出发，然后讲方法及理论。

基本教材：Elements of Statistical Learning, 2nd Ed, Hastie, Tibshirani, and Friedman. 我们不一定完全按照教材讲，根据需要改动次序，但大致包含如下内容。

1. 回归分析 + 判别分析 (chap 3,4)
2. Bootstrap,subsampling, cross validation
3. SVM (chap 5,6,12)
4. boosting (chap 10)
5. Random forest, Bagging (chap8,9, 15)
6. Graphical Models (chap 17)
7. 一些算法和理论分析。 Bragman iteration, Lasso 存在的问题等
8. Unsupervised learning -- PCA
9. Unsupervised learning -- Spectral clustering
10. Unsupervised learning -- embedding
11. neural networks and deep learning (chap 11)
12. social networks

这些课程的顺序可以调整。而且每一个内容可能不止一次课就可以讲完。