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. 本课程与“统计学习”一课为姊妹课，统计学习主要覆盖有监督学习(Supervised Learning)部分，本课程则试图覆盖一些非监督学习的方面并给出一个统一的数学系统。

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。本课程没有期末考试，但有很多的作业和projects, 需要实际投入。

(2) 需要大量的时间阅读相关文献。该课会讲很多方法。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。后一个小时讨论或报

告，学生做报告有加分。

(2) 不定期作业（大约每周weekly），包含很多miniprojects; 期末final project，没有考试。考核分为平时成绩和projects 的综合评分。

(3) 在projects 中，鼓励同学组成小组，因为数据科学家首要的能力是多学科合作(Collaborations),每组不超过3个人, 大家共享结果，但是在footnote/acknowledgement中要说明每个人的贡献。

(4) 我们将组织student workshop，用peer review的方式打分选出优秀的project工作。

(5) 上课内容包括：

I. Geometric aspects of data
A. Geometric embedding of data:
    Euclidean Embedding (MDS + extensions)
    Hilbertian Embedding (Kernel methods)
B. Adaptive embedding and matrix factorization
    Dictionary learning (basis, frames, nonnegative, topic)
    Hierarchical Embedding (DeepNN)
C. Dimensionality Reduction
    Linear adaptive DR and PCA
    Random projections
    Nonlinear DR and manifold learning
D. Curse of Dimensionality and High dimensional statistics
    Failure of mean and covariance estimators when p>n
    Sparse representation
    Random matrix theory
II. Topological aspects of data
    Clustering on data graphs
    High order homology on data complexes

Connecting geometry and topology: Hodge Theory

Spectral clustering and graph Laplacian

Statistical Ranking and graph Helmholtzian

Game Theory

Localizing the holes: Sensor Network coverage

III. Topics in algebraic methods

Tensor and Mixture models

Sum-of-Squares and MDS

IV. Algorithmic aspects

Convex and non-convex optimization

online or stochastic gradient methods

V. Statistical and Computational Complexity

Statistical stability and decision theory

Stability and computational complexity in real machines

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；

(5) 以及其它一些大家感兴趣或者新提出的问题；

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。