If you prefer not to stay in the classroom, you can attend this lecture remotely using this Zoom id:

9314 9999 627

# STAT 154

Introduction to statistical learning
(Modern statistical prediction and machine learning)

**Aug 26, 2021**
**Instructor: Song Mei**
**GSI: Ryan Theisen**

# Syllabus

# Introduction to Statistical Learning

- What is statistical learning?
- Why do we need statistical learning?
- How to **learn** statistical learning?

# What and why?

# Example 1: Default rate prediction

- Default rate prediction: if I (the bank) lend money to this person, what is the probability that they is not able to pay it back (default).

- Features (x variables): {nationality, professions, employment history, total income, credit history (credit cards), gender and race (these are sensitive and disputable features) … } of a particular person.

- Prediction (y variable): whether default (binary) / the probability of default (in [0, 1] interval).

- Data: millions of (x, y) pairs. SL is to learn from these data such that:

- Now comes a new customer, predict his default rate (0.01?, 0.001?, 0.000000001?)

- Why? Bank will only lend money to those people with default rate less than a small threshold. Otherwise, they will not be profitable.

# Example 2: Click through rate prediction

- Click through rate prediction: if we (google/facebook, e.g.) show this advertisement (of Adidas, e.g.) to this user (of a website), whether they will click it.
- Features (x variables): {search history (google), video history (youtube), consumption history (amazon), like history (facebook)} of a particular person.
- Prediction (y variable): whether click (binary) / the probability of click (in [0, 1] interval) of a particular ads.
- Data: billions of (x, y) pairs. SL is to learn from these data such that:
- Now comes a new user, predict his CTR (0.01?, 0.001?, 0.000000001?)
- Why? Google will show ads to those people with a high CTR in general, because Advertisers (Adidas, e.g.) will pay google only when a user click the ads.

# Example 3: Disease prediction

- Disease prediction: (Biotech companies, 23andMe e.g.) Given the genes of a person, predict the probability of a particular disease.
- Features (x variables): {all the genes} of a particular person.
- Prediction (y variable): the probability of getting a particular disease.
- Data: millions of (x, y) pairs. SL is to learn from these data such that:
- Now comes a new user, predict his probability of getting a disease (0.01?, 0.001?, 0.000000001?)
- Why? The gene sequencing technology is fast growing becoming increasingly cheaper. This is a service that customers will pay for and now affordable to normal people.

# Example 4: Image/video annotation

- Image annotation: (Tesla) Given the online video recorded by the camera, find out all the vehicles/pedestrians/traffic lights, etc.

- Features (x variables): An image/a video.

- Prediction (y variable): whether there is a vehicle/location of the vehicle.

- Data: trillions of (x, y) pairs. SL is to learn from these data such that:

- Given the camera data, figure out the road situation near the car.

- Why? The essential step to autonomous driving is to understand the current situation. Then the system can make decisions based on the situation.

# Example 5: Solve PDEs

- Solve PDEs: Given an operator L and the right hand side f, solve u such that L u = f. Example: Heat equation, schrodinger equation.

- Features (x variables): right hand side f.

- Prediction (y variables): solution u.

- Data: trillions of (x, y) pairs (obtained by simple simulation). SL is to learn from these data such that:

- Given a new right hand side f, solve u.

- Why? The data driven approach can be much faster (in some sense) than traditional PDE solvers.

Many more other examples....

“Statistics is the discipline that concerns the collection, organization, **analysis**, interpretation, and presentation of **data**.”

“Machine learning is the study of **computer algorithms** that can improve automatically through experience and by the use of **data**.”

*- Wikipedia*

# Many concepts with similar meanings

- Statistics
- Machine learning
- Statistical learning
- Artificial intelligence
- Data science

# This course

- **Basic topics: (70%)**
- Tasks: Regression. Classification. Dimension reduction. Clustering.
- Algorithms: Solving linear systems. Gradient descent. Newton's method. Power iteration for eigenvalue problems. EM algorithms.
- Others: Kernel methods. Regularization. Sample splitting. Resampling methods. Cross validation.
- **Advanced topics: (30%)**
- Statistical learning theory and optimization theory.
- Bagging and Boosting. Tree based models. Neural networks. Bayesian models.
- Online learning. Bandit problems.

# Prerequisite

- Multi-variate Calculus
- Linear algebra
- Basic probability and statistics
- Would be good if you have used R or Python.

# Next week

- Labs (Monday): Overview of programming languages including R and Python.
- Next lecture (Tuesday): Overview of multivariate calculus, linear algebra, and statistics.
- First topic (Thursday): Linear regression.