Lecture 1: Course logistics, homework 0

STATS 202: Data mining and analysis

Guenther Walther
September 25, 2017

(Slide credits: Sergio Bacallado)
▶ **Videos:** Every lecture will be recorded by SCPD.
Videos: Every lecture will be recorded by SCPD.

Email policy: Please use the Piazza site for most questions. For administrative issues that only concern you, email the course staff mailing list:

stats202-aut1718-staff@lists.stanford.edu
Videos: Every lecture will be recorded by SCPD.

Email policy: Please use the Piazza site for most questions. For administrative issues that only concern you, email the course staff mailing list:

stats202-aut1718-staff@lists.stanford.edu

Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.
Prediction challenges

In a typical prediction challenge, you are given a training set of labeled datapoints.

As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.
Prediction challenges

In a typical prediction challenge, you are given a **training** set of labeled datapoints.

As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.

You are also given a **test** set of unlabeled datapoints.
In a typical prediction challenge, you are given a training set of labeled datapoints.

As an example, the MNIST dataset is a library of handwritten digits, labeled from 0 to 9.

![MNIST dataset example]

You are also given a test set of unlabeled datapoints.

Your job is to assign a label (digit) to each image in the test set as accurately as possible.
Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system. The prize was $1 million.
The Netflix prize

Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system.

The prize was $1 million.

Only 1% of the 100 million possible user-movie rating pairs were revealed.
The Netflix prize

Netflix popularized prediction challenges by organizing an open, blind contest to improve its recommendation system.

The prize was $1 million.

The challenge was to predict the unseen ratings.
Kaggle

Company founded in 2010.

Business model:

▶ Organize prediction competitions hosted online.
▶ Offer companies consulting services from Kaggle “masters”.

Kaggle-in-class is a competition engine offered to degree-granting institutions for free. Stats 202 was the first class to use it!

This quarter you will team up with up to three students and participate in a Kaggle challenge TBD.