STATS 305A - Lecture 1

John Duchi
Scribed by Michael Howes

August 10, 2023

Outline for today

- Logistics
- Overview and motivation
- Background and linear models

1 Logistics

- Lecturer: John Duchi, four TAs Michael, Suyesh, James and one more.
- Email list: stats305a-aut2122@lists.stanford.edu
- Course webpage stanford.edu/class/stats305a
- Assessment:
  - (∼50%) 4 problem sets due fortnightly. Some combination of coding and maths. You can collaborate with each other and consult sources but do not ask the questions online on stackexchange, stack overflow, etc.
  - (∼50%) around 5 *etudes*. These are smaller assignments. No student collaboration allowed.
  - The lowest scoring PSET or etude will be dropped.
  - Gradescope will be used for submitting assessment.
- Course discussion will take place on ed-discussions.
- Recommended background linear algebra + probability + statistics + coding. If you have 3+ of these you will be fine. You might struggle if you are missing 2 or more.
- A variety of sources will be used. See the webpage. These textbooks can be accessed for free online at Springerlink. Lecture notes will also be online. You can order textbooks from Springerlink for $25.

2 Overview and motivation

2.1 Statistics

The big picture of 305A/B/C is that stats helps us make important decisions and scientific discoveries.

The point of this class (and most of statistics) is to make generalisation. This is normally done in the following steps: get data from a population, make inferences and learn things about the population, learn/make something new such as a better prediction or a new association.

The setting for most of this class will consist of

- Data in pairs (x, y) called observations or examples.
- The x's are called covariates, independent variables or input variables.
- The y's are called targets, dependent variables or labels.
- We want to do *supervised learning*. That is, we wish to predict new y's from new x's.

There are two broad approaches to supervised learning. The ML (machine learning) approach can be characterised by the following

- Find a good predictor of $y|x$ (y given x).
- Algorithmic.
- Rarely care much about the precise model that goes from x to y.

In contrast, the statistical approach is characterised by

- Find a model of (X, Y). These models are always wrong but are often insigtful.
- Model-based approach.
- Care a *lot* about uncertainty and we want to quantify the uncertainty in our estimates.

2.2 Applied statistics

We wish to summarize and display data and communicate what we have done outside statistics. We have to consider the losses and consequenes of models/predictions. For example

- If we want to evaluate a policy intervention, how do we decide which outcomes to measure? How do we work out how to measure them?
- In around 2007 John was working at YouTube and his team was given the task of "improving YouTube". They had to decide on a metric for how to measure this and choose to you "click through rate" [what proportion of the time do viewers click on a suggested video]. They developed an ML algorithm called Sibyl that dramatically increased click through rate. Fourteen years later, it is quite clear that "click through rate" was not a good choice of metric as Sibyl has resulted in many people going into rabbit holes about conspiracy theories on YouTube.
- Kathyrn Page Harden is a researcher who works on GWAS (genome-wide association studies). These studies associate gene expression (X) with outcomes linke income, IQ, level of education and school success (Y). A linear model is used to predict Y from X. From a particular person's x we can calculate their "score" $\mathbb{E}[Y|X = x]$. This raises some scary questions.
  - What if this was used in job selection or college admissions?
  - What if embroys in IVF were chosen based on the scores?

  The way in which the scores are communicated is very important. On a population level, there is a strong trend that the IQ/income/level of education do increase with these scores but on an individual level the variance within a group of people with the same score is much greater than the variance across the different categories.

In applied statistics we also have to think about how we gather the data and make good estimates and we have to talk to domain experts.

2.3 Models

Much of this course focus on the choice of model for the given data. There is a famous quote by the statistician Box "All models are wrong but some are useful." One might ask: "If all models are wrong, then what's the point?". A partial answer is that we know the following:

- Robustness: some methods may be "optimal" under the assumptions of one model but still perform "okay" when the assumptions are missing.
- We can measure our errors and validate our models. We can see if we do better than without the model.
- Models provide simplification. They can capture what's happening and give us enough understanding to make decisions.

2.4 Supervised learning/Prediction Problems

As we before we have our observations/examples (X, Y) which come in pairs. The Y is the target/dependent variable/label and the X is the covariate/input/indepent variable. In this section we could ask two questions.

- Prediction: Given a new X, can I predict the corresponding Y?
- Causality: If I intervened on X, what happens to Y?

Causal questions are more sciency and include asking if smoking cause cancer and if government spending causes economic growth. Causal questions are much harder to answer and won't be taught much in this course. Prediction is still very useful as sometime we cannot perform interventions, eg path of a hurricane.

The variables X and Y may be any one of a number of different types. Some common types for Y are $\mathbb{R}$ (real numbers), $\{0, 1\}$ or $\{1, -1\}$ (binary classification), $\{1, \ldots, k\}$ (classification), ordered lists, $\mathbb{R}^d$ or structure prediction. Some possibilities for X are $\emptyset$ (no covariates), $\{0, 1\}$ or $\{1, \ldots, k\}$ (analysis of variance (ANOVA)), $\mathbb{R}^d$ and more! In this class we will be focusing on the case when $Y \in \mathbb{R}$ but we will see each of the possibilities for X. The course 305B will look at when Y is in $\{0, 1\}$,$\{1, \ldots, k\}$ or an ordered list. Finally 305C will look at when $Y \in \mathbb{R}^d$. The tools of natural language processing (NLP) can be used for structured prediction.

3 Linear Models

An example of a linear model is

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

where $\beta_0, \beta_1 \in \mathbb{R}$ are our parameters and ϵ_i is random noise. We will often assume $\epsilon_i \overset{\text{iid}}{\sim} (0, \sigma^2)$ (that is ϵ_i are independent and indentically distributed, $\mathbb{E}[\epsilon_i] = 0$, and the variance of ϵ_i is σ^2). Other times we make the stronger assumption that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$ which states in addition that each ϵ_i is normally distributed. Depending on the context/application will we write either

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 x_i \quad \text{or} \quad \mathbb{E}[Y | X = x_i] = \beta_0 + \beta_1 x_i.$$

The parameter β_0 is called the intercept and β_1 is called a parameter of interest.

Some times we will have $x_i = (x_{i1}, \ldots, x_{id})^T \in \mathbb{R}^d$. In this case we can use the model

$$Y_i = \beta_0 + \sum_{j=1}^{d} \beta_j x_{ij} + \epsilon_i = \beta_0 + x_i^T \beta + \epsilon_i.$$

We will often simply write $y = x^T\beta + \epsilon$ where $x = (1, x_1, x_2, \ldots, x_d) \in \mathbb{R}^{d+1}$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_d) \in \mathbb{R}^{d+1}$.

When we refer to linear models we are assuming that the model is linear in β not necessarily in x. For example we have do polynomial regression when we have $x \in \mathbb{R}$ and use

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d.$$

In another example we may have Y = daily rainfall and X = day of the year $\in \{1, 2, \ldots, 365\}$. In that case we could use the model

$$Y = \beta_0 + \beta_1 \sin\left(\frac{2\pi x}{365}\right) + \beta_2 \cos\left(\frac{2\pi x}{365}\right).$$

This gives us cyclical data.