Introduction to Statistical Learning

Malka Guillot* and Philine Widmer**

*ETH Zurich **University of St.Gallen

SICSS ETH Zurich, 14 June 2021

Readings we recommend

- James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.
- Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z. (2015). Prediction Policy Problems. American Economic Review, 105(5), 491-95.
- Mullainathan, S., Spiess, J. (2017). Machine Learning: an Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106.
- Prosperi, M., Guo, Y., Sperrin, M., Koopman, J. S., Min, J. S., He, X., ... Bian, J. (2020). Causal Inference and Counterfactual Prediction in Machine Learning for Actionable Healthcare. Nature Machine Intelligence, 2(7), 369-375. (Specifically on health but highlights causality/prediction concerns well!)

Setting in our class

- Many of you have an (political) economics(-ish) background
- **Poll:** how many of you have had classes in ...
  - Causal inference?
  - Prediction/machine learning?

Content of today[1]

- What is statistical learning?
- Statistics in social science: causality
- Statistics in machine learning: prediction
- Accuracy versus interpretability

[1] This material is partly based on Malka's class at ETH Zurich on "Big Data for Public Policy", teaching material by Professor Jason Anastasopoulos (https://anastasopoulos.io/), and James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

General setting in statistical learning

- Input variables $\mathcal{X}$
  - Also known as: features, independent variables, predictors
- Output variables $\mathcal{Y}$
  - Also known as: dependent variables, outcomes, etc.

Statistical learning theory

- $\mathcal{X} \to \mathcal{Y}$
- $\mathcal{X} \in \mathbb{R}^{n \times p}, \mathcal{Y} \in \mathbb{R}^{n}$

$\rightarrow$ Statistical learning: approaches for finding a function that accurately maps the inputs $\mathcal{X}$ to outputs $\mathcal{Y}$

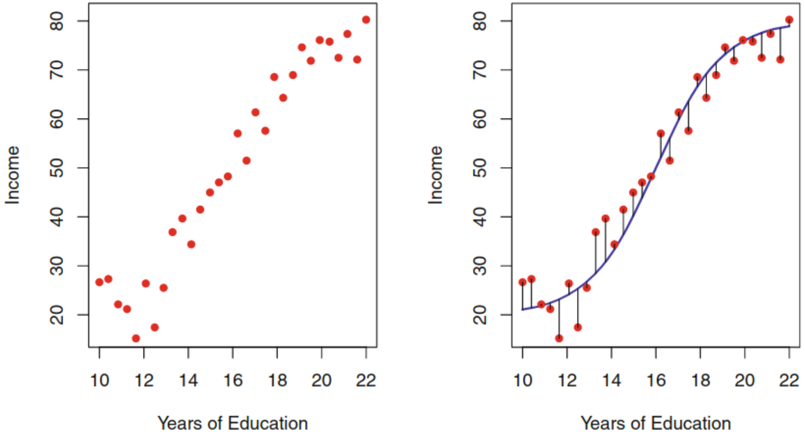
Statistical model

Finding $f(\bullet)$ such that $Y = f(X) + \epsilon$

- $f(X)$ is an unknown function of a matrix of predictors $X = (X_1, ..., X_p)$
- Y: a scalar outcome variable
- Error term ϵ with mean zero
- While X and Y are known, $f(\bullet)$ is unknown

$\rightarrow$ Goal of statistical learning: to utilize a set of approaches to estimate the "best" $f(\bullet)$ for the problem at hand

Example: income as a function of education



Source: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

Prediction

- Predict Y by $\hat{Y} = \hat{f}(X)$
- When do we care about "pure prediction"?
  - X readily available but Y is not
  - $\hat{f}$ can be a **black box**: the only concern is prediction accuracy

Inference

- Understanding the way that Y is affected as $X_1, ..., X_p$ change
- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?

$\rightarrow$ $\hat{f}$ cannot be a **black box** anymore

Approach in social science

- Objective: understanding the way that Y is affected as $X_1, ..., X_p$ change
- The goal not necessarily to make predictions for Y
- Often linear function to estimate Y: $f(X) = \sum_{i=1}^{p} \beta_i x_i$
- Assume $\epsilon \sim N(0, \sigma^2)$
- Parameters β are estimated by minimizing the sum of squared errors: $Y = \sum_{i=1}^{p} \beta_i x_i + \epsilon$

Approach in social science: causality (1/2)

- $Y = \beta_0 + \beta_1 T + \sum_{i=1}^{p-1} \beta_i x_i + \epsilon$
- Interested in the values of one or two parameters and whether they are **causal** or not
- Framework to interpret statistical causality: counterfactuals

Approach in social science: causality (2/2)

- Causal inference requires that $T \perp \epsilon$ or $T|X \perp \epsilon$
  - $\rightarrow$ Can be achieved through randomization of T
- This implies that we are not really all that interested in choosing an optimal $f(\bullet)$
- (We want to estimate unbiased coefficients)

Approach in machine learning: prediction

- $\hat{Y} = \hat{f}(X)$
- Objectives:
  - Find the "best" $f(\bullet)$ and the "best" set of X $\rightarrow$ "best" means giving the most accurate predictions $\hat{Y}$
  - Accuracy: find the function that minimizes the difference between *predicted* and *observed* values
  - That is: we seek to minimize the prediction error

Reducible and irreducible error (1/2)

- Estimated function: $\hat{f}(X) = \hat{Y}$
- True function: $f(X) + \epsilon = \hat{Y}$
- Reducible error: $\hat{f}$ is used to estimate f
  - But it's not perfect
  - $\rightarrow$ Accuracy can (maybe) be improved by adding more features and/or data
- Irreducible error: ϵ = all other features that can be used to predict f
  - $\rightarrow$ Unobserved $\rightarrow$ irreducible

Reducible and irreducible error (2/2)

$$E(Y - \hat{Y})^2 = E[f(X) + \epsilon - \hat{f}(X)]^2$$

$$= \underbrace{[f(X) - \hat{f}(X)]^2}_{Reducible} + \overbrace{Var(\epsilon)}^{Irreducible}$$

$\rightarrow$ **Objective**: estimating f with a minimal reducible error

How do we estimate f?

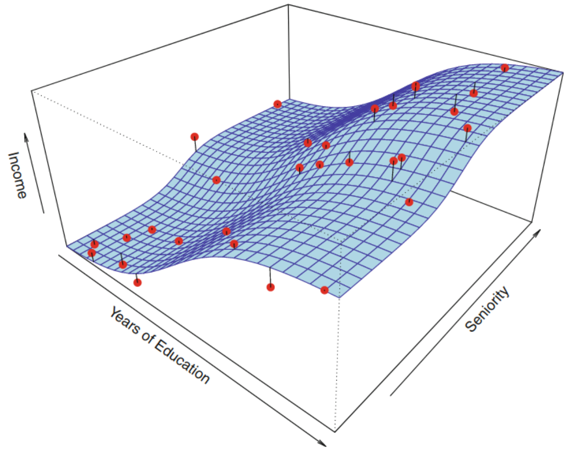
- We use observations to “teach” our ML algorithm to predict outcomes
- Training data: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Thereby, $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$
- We use observations to “teach” our ML algorithm to predict outcomes
- Two types of SL methods: **parametric vs. non-parametric**

Parametric methods

Model-based approaches, 2 steps:

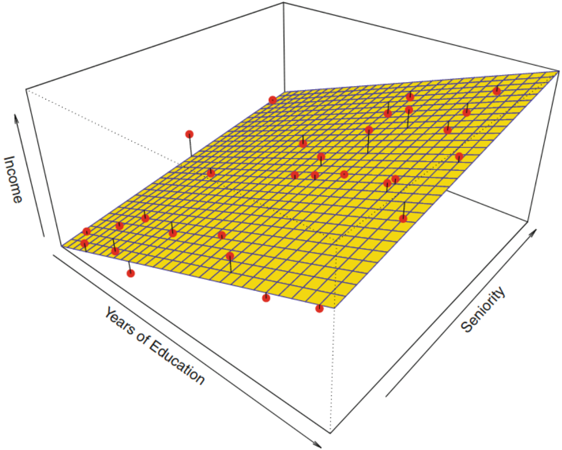
1. Specify a **parametric (functional) form** for $f(X)$
   - For example: linear $f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$
   - Parametric means that the function depends on a finite number of parameters, here $p + 1$
2. Training: estimate the parameters (e.g., by OLS) and predict Y
   - $\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$

Imagine the true relationship looks like this:



Source: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

A linear model could approximate the function like this:



Source: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

Parametric methods: issues

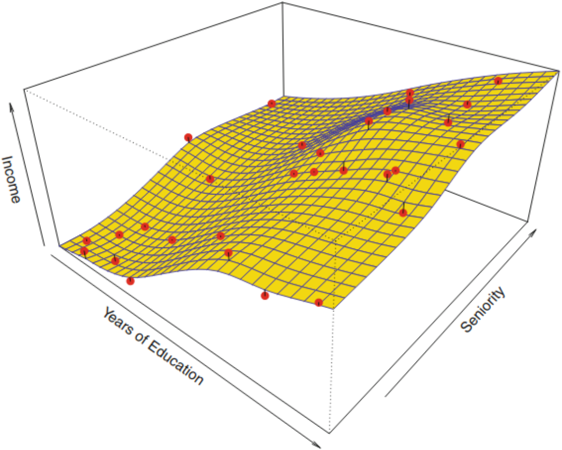
Mis-specification of $f(X)$

- Rigid models (e.g. strictly linear) may not fit the data well
- More flexible models require more parameter estimations
  - $\rightarrow$ Potentially overfitting

Non-parametric methods

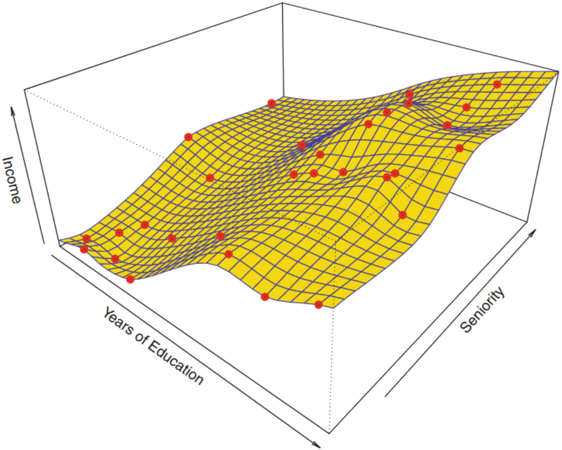
- **No assumptions** about the functional form of f
- Estimates a function only **based on the data itself**
- Disadvantage: large number of observations is required to obtain an accurate estimate of f

Remember our example – a smooth non-linear estimate could look like this:



Source: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

Remember our example – a rough, overfitted non-linear estimate could look like this:



Source: James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning.

Recap: parametric vs. non-parametric approaches

Quiz: which statements apply to parametric models?

- Only estimate a set of parameters
- Give insights on the data when nothing is known
- Offer better predictions with little data
- Rely on assumptions about functional form

Accuracy and interpretability trade-offs

- More accurate models often require estimating more parameters and/or being more flexible
- Models that are better at prediction generally are less interpretable
- For inference, we do care about interpretability

Machine learning: supervised vs. unsupervised learning

- **Supervised learning:** estimating functions with known observation and outcome data
  - We observe data on Y and X and want to learn the mapping $\hat{Y} = \hat{f}(X)$
  - Classification for discrete Y, regression for continuous Y
- **Unsupervised learning:** estimating functions without the aid of outcome data
  - We only observe X and want to learn something about its structure
  - E.g., clustering (partition data into homogeneous groups based on X) or PCA for dimensionality reduction

Examples of social science studies using machine learning for prediction (1/2)

- Glaeser, Kominers, Luca, and Naik (2016) use images from Google Street View to measure block-level income in New York City and Boston
- Jean et al. (2016) train a neural net to predict local economic outcomes from satellite data in Africa
- Chandler, Levitt, and List (2011) predict shootings among high-risk youth so that mentoring interventions can be appropriately targeted
- Kleinberg, Lakkaraju, Leskovec, Ludwig, and Mullainathan (2018) predict the crime probability of defendants released from investigative custody to improve judge decisions

Examples of social science studies using machine learning for prediction (2/2)

- Kang, Kuznetsova, Luca, and Choi (2013) use restaurant reviews on Yelp.com to predict the outcome of hygiene inspections
- Huber and Imhof (2018) use machine learning to detect bid-rigging cartels in Switzerland
- Kogan, Levin, Routledge, Sagi, and Smith (2009) predict volatility of firms from market-risk disclosure texts (annual 10-K forms)

The machine learning workflow

- Look at the big picture
- Get the data
- Discover and visualize the data to gain insights
- Prepare the data for Machine Learning algorithms
- Select a model and train it
- Fine-tune your model
- Present your solution
- Launch, monitor, and maintain your system

From Aurelien Geron, Hands-on Machine Learning with Scikit-Learn TensorFlow, Chapter 2 (cf. our GitHub)

Conclusion: econometrics vs. machine learning (1/2)[2]

- Common objective: to build a predictive model, for a variable of interest, using explanatory variables (or features)
- Different cultures:
  - Econometrics: probabilistic models designed to describe economic phenomena
  - ML: algorithms capable of learning from their mistakes

[2] Charpentier A., Flachaire, E. Ly, A. (2018). Econometrics and Machine Learning. Economics and Statistics, 505-506, 147–169.

What does all of this have to do with SICSS?

- ML inherent to many computational approaches in social science (e.g., computational linguistics)
- As a social scientist in general, you will likely encounter ML in some ways
- Interdisciplinary focus of SICSS: bringing together the best of many worlds

There is a quickly growing literature on econometrics + machine learning

- More on this later this week
- Next week: guest lecture by Professor Michael Knaus on "**Double Machine Learning based Program Evaluation**"

  22 June 2021, 17-18h, IFW building, ETH Zurich

Thank you for your attention ☺

Questions?