MY474: Applied Machine Learning for Social Science

Lecture 1: Getting Started, Brief Introduction to Machine Learning

Blake Miller

18 January 2023

Agenda

1. Motivations
2. Logistics
3. What is machine learning?
4. A real-world example
5. Which model do we choose?
6. Quiz review

Motivations

Learning from Data

- **Fact:** The amount of data and information collected and stored is constantly increasing, due to advances in data collection, computerization of many aspects of life and breakthroughs in storage technology.
- **Consequence:** Statistical problems have increased both in size and complexity.
- **The data analyst's job:** make sense of all these data! Identify patterns and trends, uncover interesting relationships among the variables and/or the observations, predict future behavior.
- Machine learning can help us process and understand these data.

Learning from Data

- Technology helps
  - Faster computers $\rightarrow$ more flexible and thus more powerful techniques $\rightarrow$ fewer modeling assumptions
  - New capabilities from graphical processing units (GPUs)
- But not always: Faster computers do not solve all problems

Logistics

About Me

- **Assistant Professor in Computational Social Science at the Methodology Department, LSE**
- Previously at **Dartmouth College**
- PhD in Political Science and Scientific Computing, **University of Michigan**
- My research:
  - Chinese politics, and authoritarian politics
  - Information control (censorship, propaganda, etc.), information and mobilization of violence
  - Applied machine learning for text data
- **Contact:**
  - `b.a.miller@lse.ac.uk`
  - `www.blakeapm.com`
  - Office hours: book through Student Hub

Books

What is machine learning?

Machine learning is many things

- A method of mining data for **patterns** or unobserved **latent variables**
- A method of **approximating unknown functions** using available data
- A predictive framework for forecasting future events
- A framework for **augmenting** and **automating** human decisions or tasks

Unsupervised Learning

- **Objective:** Find patterns and structure among inputs $\mathbf{x}$ in the data $\mathcal{D}$.
- Types of unsupervised learning:
  - **Clustering:** Given data $\mathbf{X}$, identify clusters of objects that balance within-cluster similarity and distinctness between clusters.
  - **Dimension Reduction:** Given data $\mathbf{X}$, identify manifolds or underlying factors that represent the data in fewer dimensions
- Training data does not contain any outputs $\mathbf{y}$. Structure is

A real world example

A real world example

Which of these Wikipedia Talk Page comments are insults?

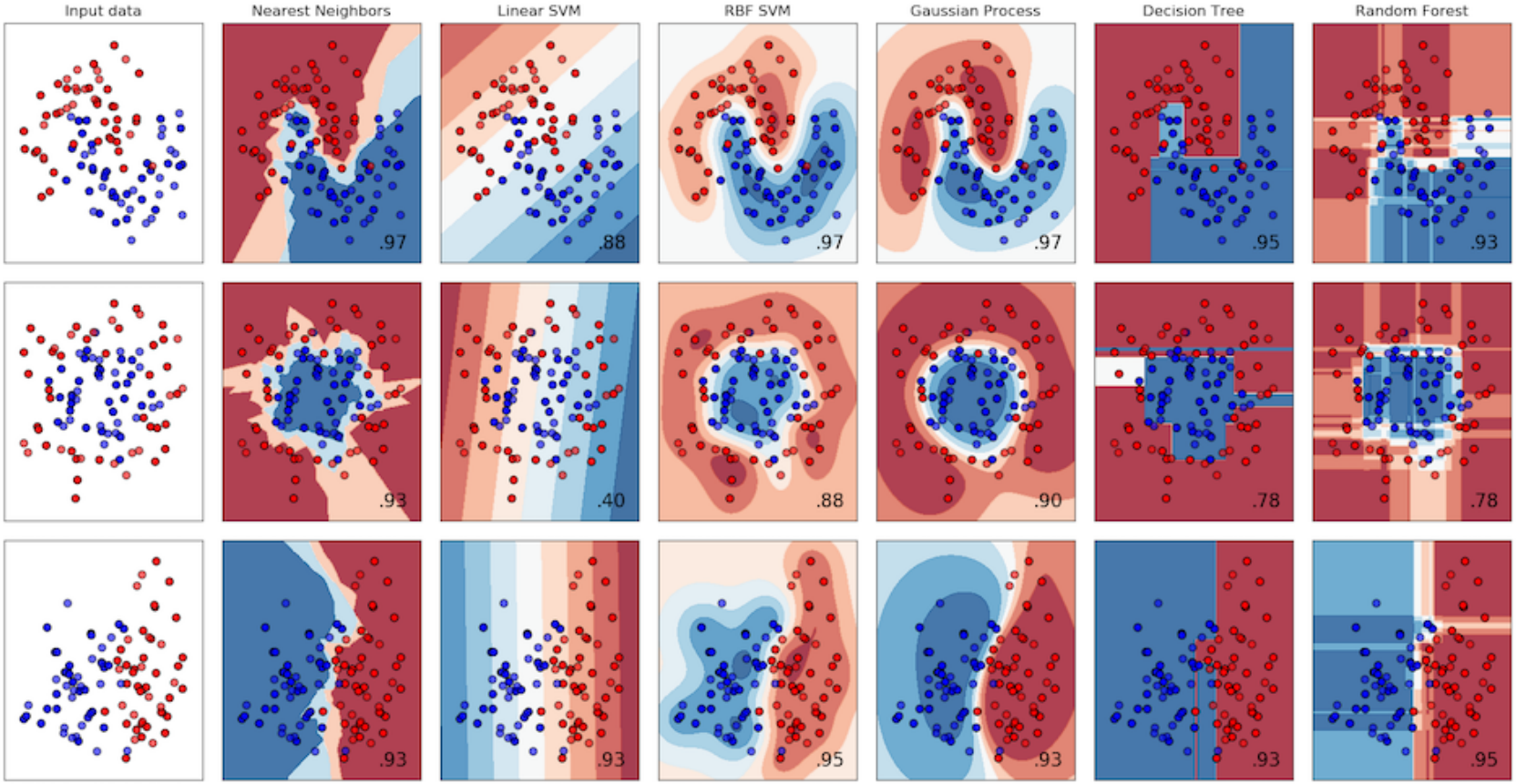
1. "I HATE your freakin guts!!"
2. "YOU ARE AN IDIOT!!!!"
3. "Before we can accept this change, we need a citation for this second claim."
4. "I'm afraid I reverted your change. You erased all of the existing discussion there; we want to keep that."
5. "STOP YOUR DAMN NONSENSE!!!!! YOU HAVE DONE NOTHING TO IMPROVE THESE ARTICLES!!! ASSHOLE!!"

A real world example

Which of these Wikipedia Talk Page comments are insults?

Which model do we choose?

Which model do we choose?



Which model works best really depends on the task and the data ("no free lunch"). (source: sklearn.org)

Types of Models

- Machine learning models come in many forms, each differing in

Some quick crowdsourcing

Click to access the idea board

Quiz Study Guide

Concepts to Know

- Curse of dimensionality
- "No free lunch"
- KNN, perceptron
- Interpretability
- Parametric vs. non-parametric
- Unknown target function
- Hypothesis set, learning model, learning algorithm
- Unsupervised vs. supervised learning

Review Questions

1. What is the difference between a learning model and a learning algorithm?
2. Why is KNN considered a non-parametric model? Why is the perceptron model considered a parametric model?
3. In what scenarios might the perceptron model out-perform the KNN model?
4. What are some examples of supervised learning problems?
5. What are some examples of unsupervised learning problems?