CS4641B Machine Learning

Lecture 02: Machine learning workflow and the term project

Rodrigo Borela ▸ rborelav@gatech.edu

Recap: what is machine learning?

Study of algorithms that

- improve their **performance** P
- at some **task** T
- with **experience** E

well-defined learning task: <P, T, E>

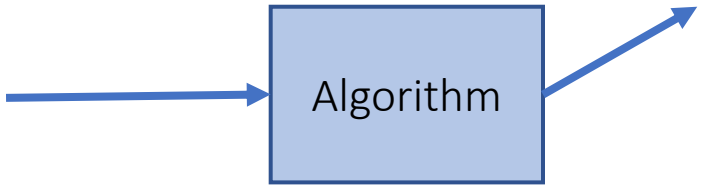
— Tom Mitchell

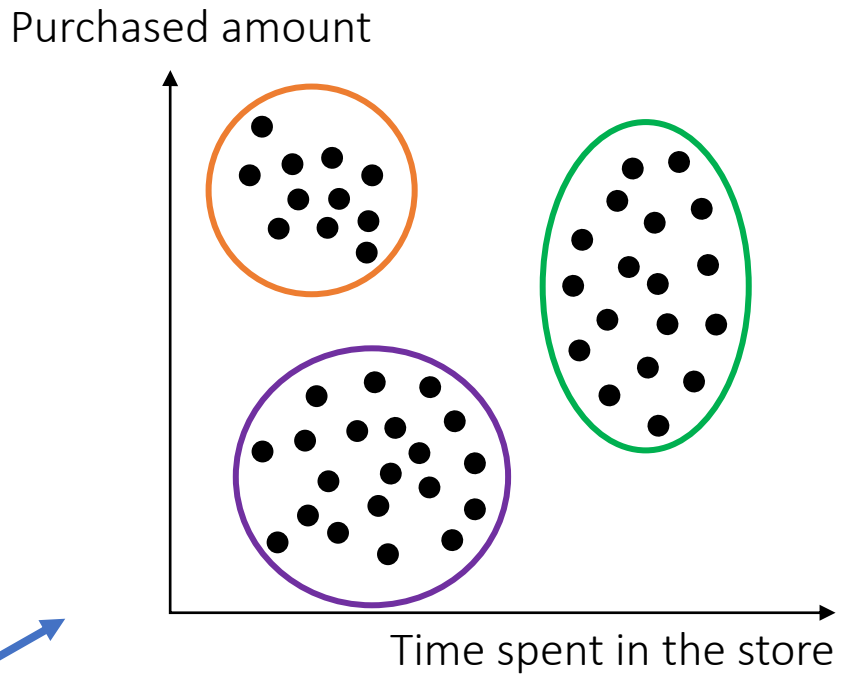
Recap: Unsupervised learning

Example: clustering for market research

What is the task?
Identify different customer groups in a department store
How do we measure performance?
Similarity between the customers in a group
What is the experience?
Dataset containing information about the amount spent and how long they were in the store.

$$\begin{matrix} \$ & \text{T [min]} \\ \begin{bmatrix} 380 & 30 \\ 122 & 17 \\ 48 & 84 \\ \vdots & \vdots \\ 87 & 5 \end{bmatrix}_{N\times 2} \end{matrix}$$





Recap: Supervised learning

Example: predicting house prices within a neighborhood

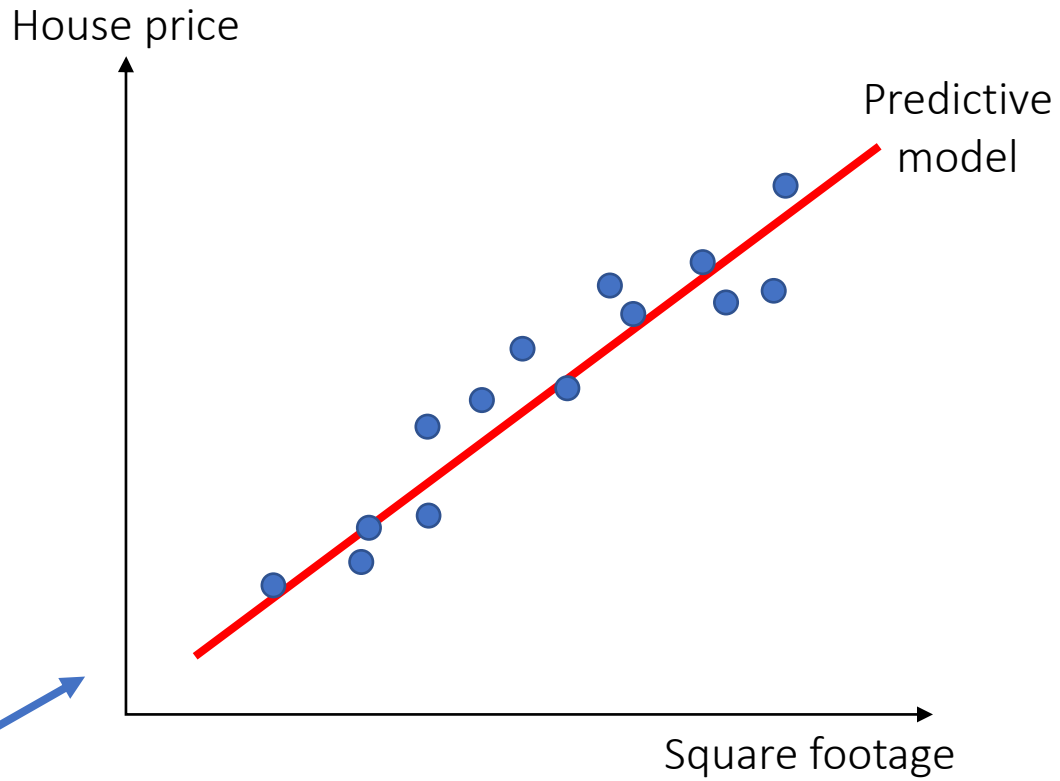
What is the task?
Predict the house prices within a neighborhood based on their square footage

How do we measure performance?
A measure of the difference between predicted and actual prices

What is the experience?
Dataset containing house square footage and prices

Area [sqft] $\begin{bmatrix} 1530 \\ 2010 \\ \vdots \\ 970 \end{bmatrix}_{N\times 1}$

$ $\begin{bmatrix} 258{,}000 \\ 295{,}000 \\ \vdots \\ 197{,}000 \end{bmatrix}_{N\times 1}$



Unsupervised and supervised learning

Example: predicting house prices in the city of Atlanta

What is the task?

Predict house prices anywhere in the city of Atlanta

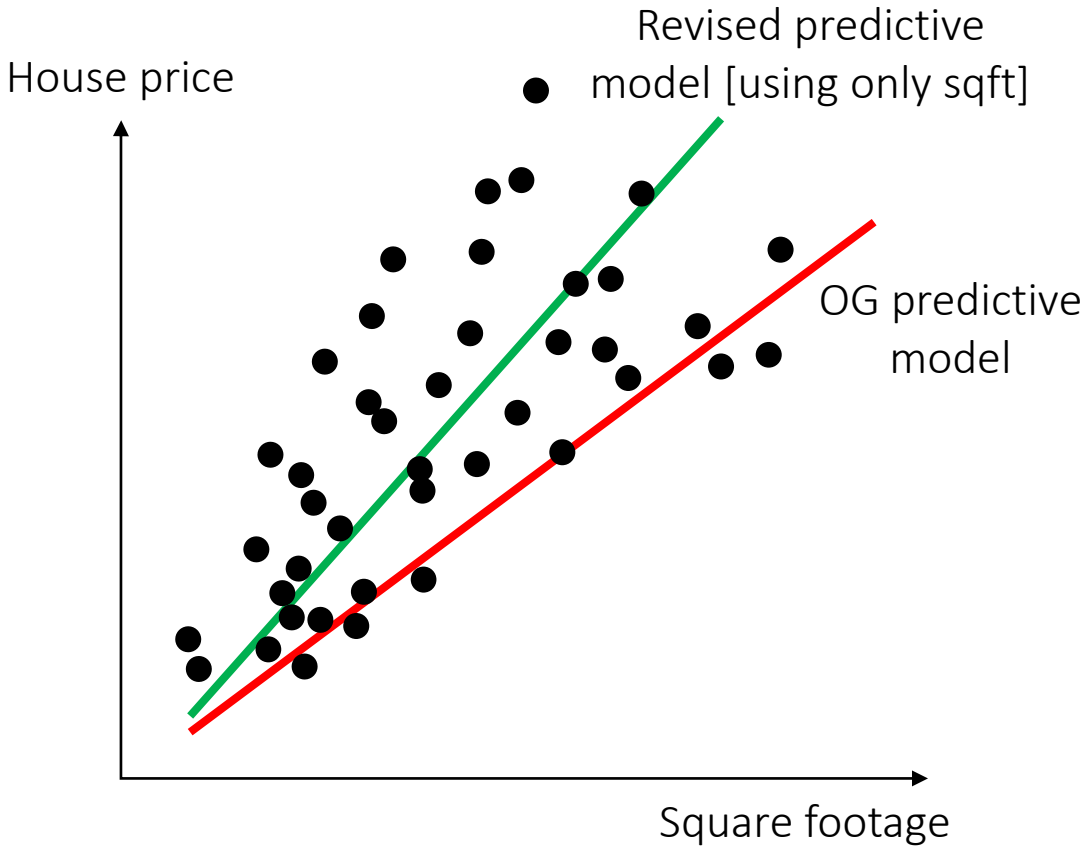
How do we measure performance?

A measure of the difference between predicted and actual prices

What is the experience?

Dataset containing information about location, square footage and house prices

Lat [°N]	Long [°W]	Area [sqft]	$
33.67	84.44	1830	258,000
33.83	84.41	1310	395,000
⋮	⋮	⋮	⋮
33.78	84.36	1100	297,000
$N \times 3$			$N \times 1$

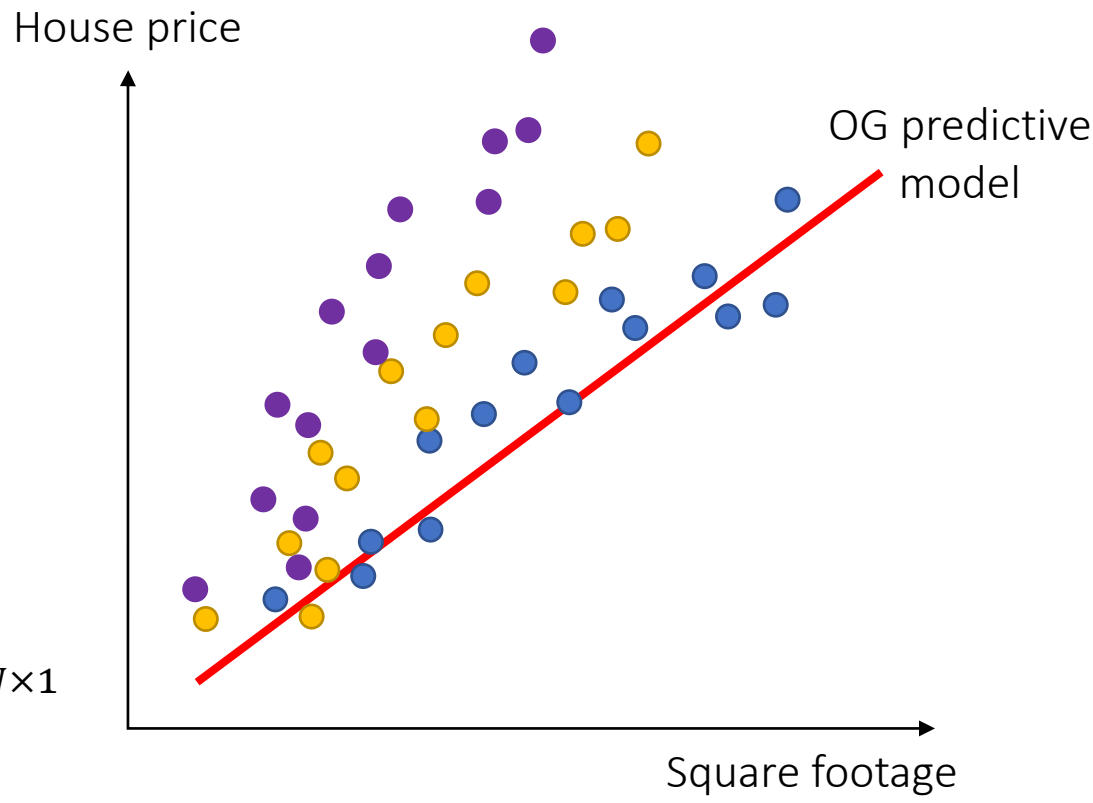


Unsupervised and supervised learning

Example: predicting house prices in the city of Atlanta

- Unsupervised learning allows us to explore and understanding our data
- With this knowledge we can select appropriate models and develop tools to make predictions about a dataset

$$
\begin{array}{r} \\ \text{East Point} \\ \text{Buckhead} \\ \\ \text{Virginia Highlands} \end{array}
\begin{array}{c}
\begin{array}{ccc} \text{Lat [°N]} & \text{Long [°W]} & \text{Area [sqft]} \end{array} \\
\begin{bmatrix} 33.67 & 84.44 & 1830 \\ 33.83 & 84.41 & 1310 \\ \vdots & \vdots & \vdots \\ 33.78 & 84.36 & 1100 \end{bmatrix}_{N\times 3}
\end{array}
\quad
\begin{array}{c}
\$ \\
\begin{bmatrix} 258{,}000 \\ 395{,}000 \\ \vdots \\ 297{,}000 \end{bmatrix}_{N\times 1}
\end{array}
$$



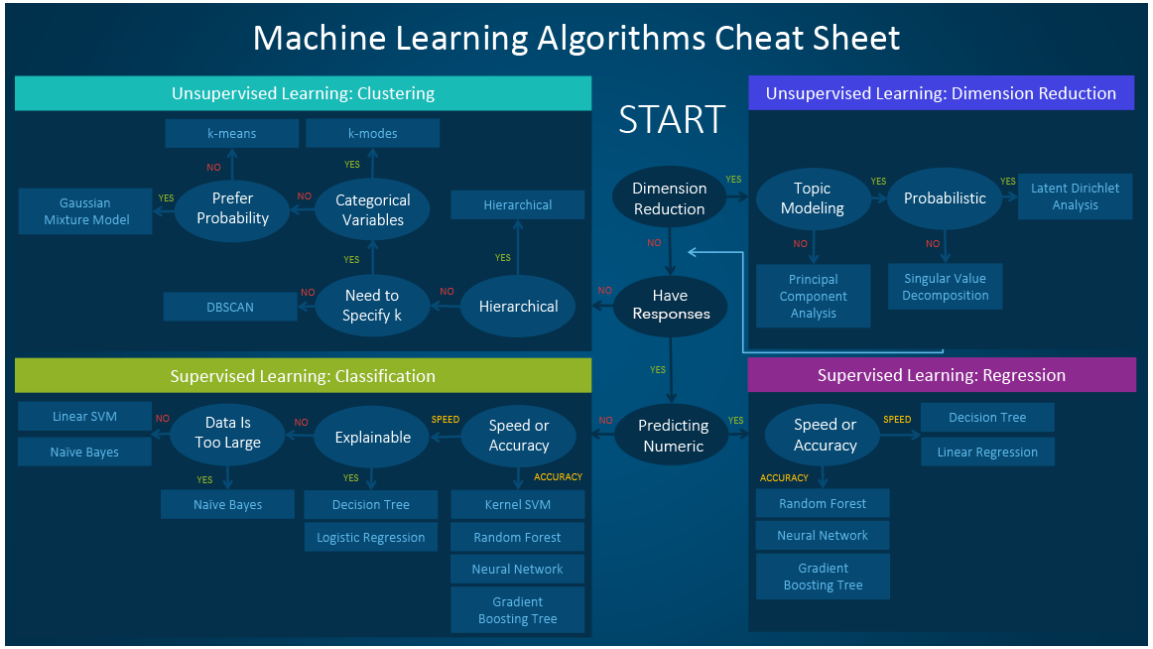
- Cluster data geographically
- Improve predictive model

Machine learning workflow process

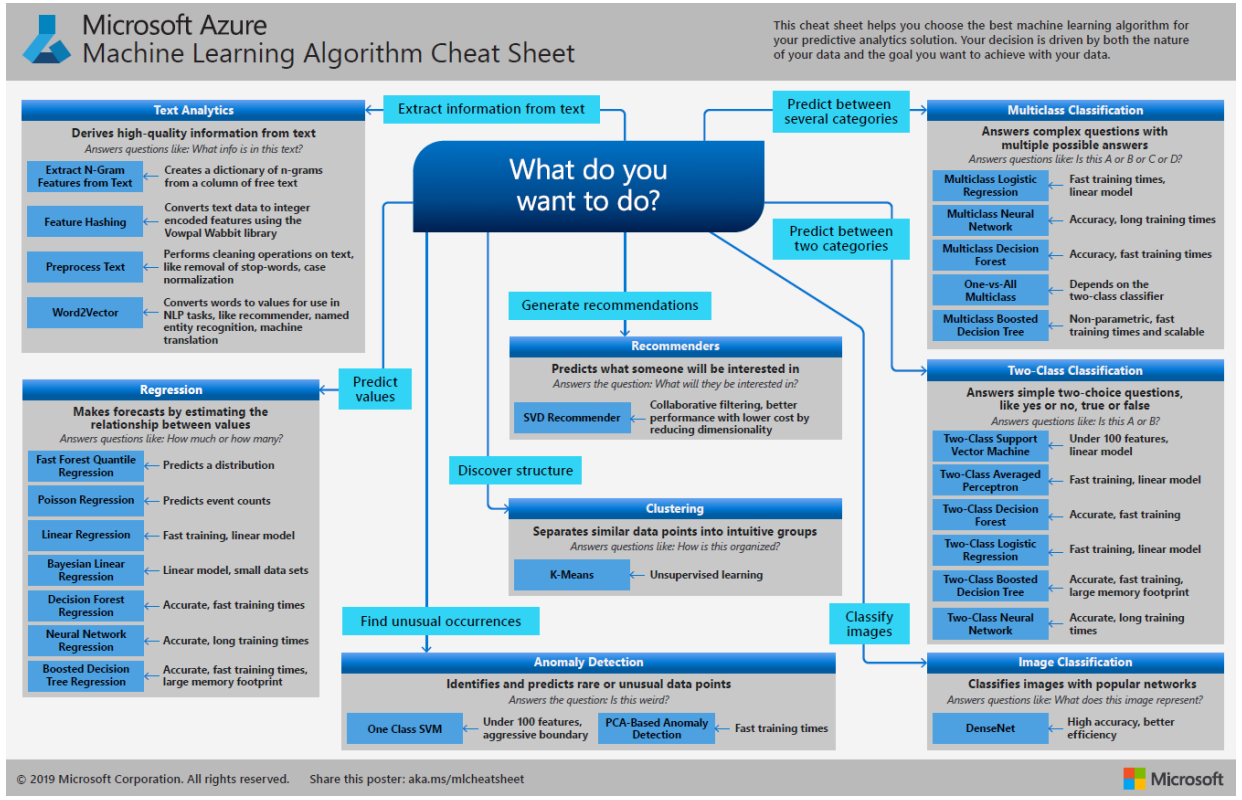
- Defining a problem
- Is this a machine learning task?
- What kind of machine learning task is it?
  - Clustering, distribution estimation, classification, regression, other?
- Do I have the data to support it?
- Do I have the resources to do it?
- What kind of data am I working with?
  - Spatial (map, trajectory), visual (images), text (documents, tweets, customer reviews), behavioral data (smoking habits), time-series data (stock prices)
- How am I evaluating success?

What algorithm should I use?

(Not a comprehensive list)



Infographic by SAS Data Science Blog. Click on the figure to access the source.



Infographic by Microsoft Azure. Click on the figure to access the source.

Logistics and teamwork conduct

- Create groups of 4 students each **(due Sep 8th)**
- Only the group leader needs to submit the deliverables using a group submission format on Gradescope (a template will be available)
- Work on your projects consistently and do not postpone it to the last days before each deadline
- Working as a team can often be a challenging experience, here are some things to keep in mind:
  - Define clear means of communication (I strongly recommend you create a channel for your group on MS Teams or Slack)
  - Listen to your teammates and be upfront about your availability
  - Play to your individual strengths when tackling an activity
  - Resolving conflicts within the team is part of the job
  - The people you work with are part of your professional network
  - Peer reviews will be used to assess your participation in the project

Important aspects to consider

- Complex tasks demand large datasets in order to generalize to different problems. The course webpage has an extensive list of databases from which you can obtain datasets to work on your project
- The training phase of techniques involving large datasets and/or deep learning architectures tend to be computationally intensive and require GPU access in order to be performed in a reasonable amount of time. Make sure you have the appropriate resources when dealing with such techniques. Here are some options of free GPU resources:
  - Colab
  - Kaggle
  - AWS Educate

Project examples

Previous editions of the course

- Fall 2019
- Spring 2020
- Summer 2020

Seminars

Starting on Aug 27th until Sep 24th two of our TAs will present a project to inspire you and motivate you in picking a task you would like to apply machine learning to. The presentations will be pre-recorded and accompanied by a Piazza thread. You are expected to watch at least three of these seminars and post a question or comment on the corresponding Piazza thread. Your contributions to the Piazza thread will compose a portion of your participation grade.

Proposal (10%)

Deliverables

- Single-slide presentation of your project proposal **(Sep 28th)**
- Three-minute pre-recorded presentation with your project proposal pitch **(Sep 28th)**
- GitHub Page. You can download an HTML template from GitHub and format it according to the structures indicated in the following. Your page should not exceed 600 words total. If you do not have a GitHub Student account yet, please apply for one ASAP here. **(due Oct 2nd)**

GitHub page structure

- Summary figure (one infographic prepared by your team that summarizes your project goal)
- Introduction/Background (Proposed)
- Methods (Expected)
- Results (what results are you trying to achieve?)
- Discussion (best outcome, what it would mean, what is next.....)
- References (list containing at least three references, preferably peer reviewed)

Heilmeier catechism

- What are you trying to do? Articulate your objectives using absolutely no jargon. *(Introduction)*
- How is it done today, and what are the limits of current practice? *(Introduction)*
- What is new in your approach and why do you think it will be successful? *(Methods)*
- Who cares? If you are successful, what difference will it make? *(Discussion)*
- What are the risks? *(Methods)*
- How much will it cost? *(Methods)*
- How long will it take? *(Methods)*
- What are the mid-term and final "exams" to check for success? *(Results)*

Source: Defense Advanced Research Projects Agency (DARPA)

Touch-point 1: Project proposal (Sep 30th)

- Watch the project proposal presentations of fellow teams that belong to your cluster before the touch-point (approximately 30 mins total)
- Prepare questions and suggestions for your fellow teams
- On the date of the touchpoint, we will meet (remotely or in-person) during class time, and we will go over the proposal made by each team, giving you an opportunity to share your thoughts and learn from each other
- A form will be sent out two weeks in advance so you can register for the in-person classroom version of the touch-points
- You should incorporate the feedback you received into your project proposal before submitting the final version on Oct 2nd

Midterm report (10%)

Deliverables

- Single-slide presentation outlining progress highlights and current challenges **(Oct 30th)**
- Three-minute pre-recorded presentation with your progress and current challenges **(Oct 30th)**
- GitHub Page. Expanded version of your previous page **(due Nov 6th)**

GitHub page structure

- Expanded and revised version of the proposal including the results you have achieved so far relating to the unsupervised learning portion of the assignment

Touch-point 2: Unsupervised learning (Nov 2nd)

- Same format as for touch-point 1

Final report (15%)

Deliverables

- Single-slide presentation outlining progress highlights and current challenges **(Nov 20th)**
- Three-minute pre-recorded presentation with your progress and current challenges **(Nov 20th)**
- GitHub Page. Expanded version of your previous page **(due Dec 7th)**
- Final seven-minute long pre-recorded presentation **(due Dec 7th)**

GitHub page structure

- Expanded and revised version of the midterm report including both unsupervised and supervised learning results for your project

Touch-point 3: Supervised learning (Nov 23rd)

- Same format as previous touch-points

Questions?