

GEORGE MCINTIRE

INTRODUCTION TO MACHINE LEARNING

TABLE OF CONTENTS

▶ I. PRESENTATION

- ▶ What is Machine Learning?
- ▶ Supervised Learning
 - ▶ Examples, classification vs regression.
- ▶ Unsupervised Learning

▶ II. FINAL PROJECT

- ▶ Objective and presentation requirements



WHAT IS MACHINE LEARNING

- ▶ “A field of study that gives computers the ability to learn without being explicitly programmed” (1959)
 - Arthur Samuel, AI pioneer, coined the term “Machine Learning”
- ▶ “The automation of activities that we associate with human thinking, activities such as decision-making, problem solving, learning...” (1978)
 - Richard Bellman, applied mathematician

WHAT IS MACHINE LEARNING (CONT.)

- ▶ Examples in the form of data are passed through algorithms that look for patterns in that data in order to make predictions and decisions on future data.
- ▶ The computer observes that data of a certain category exhibits certain characteristics and data of another category exhibits a whole set of different characteristics. Allows the computer to properly classify data without that labeling.

FACE



FACE



NOT A FACE



?

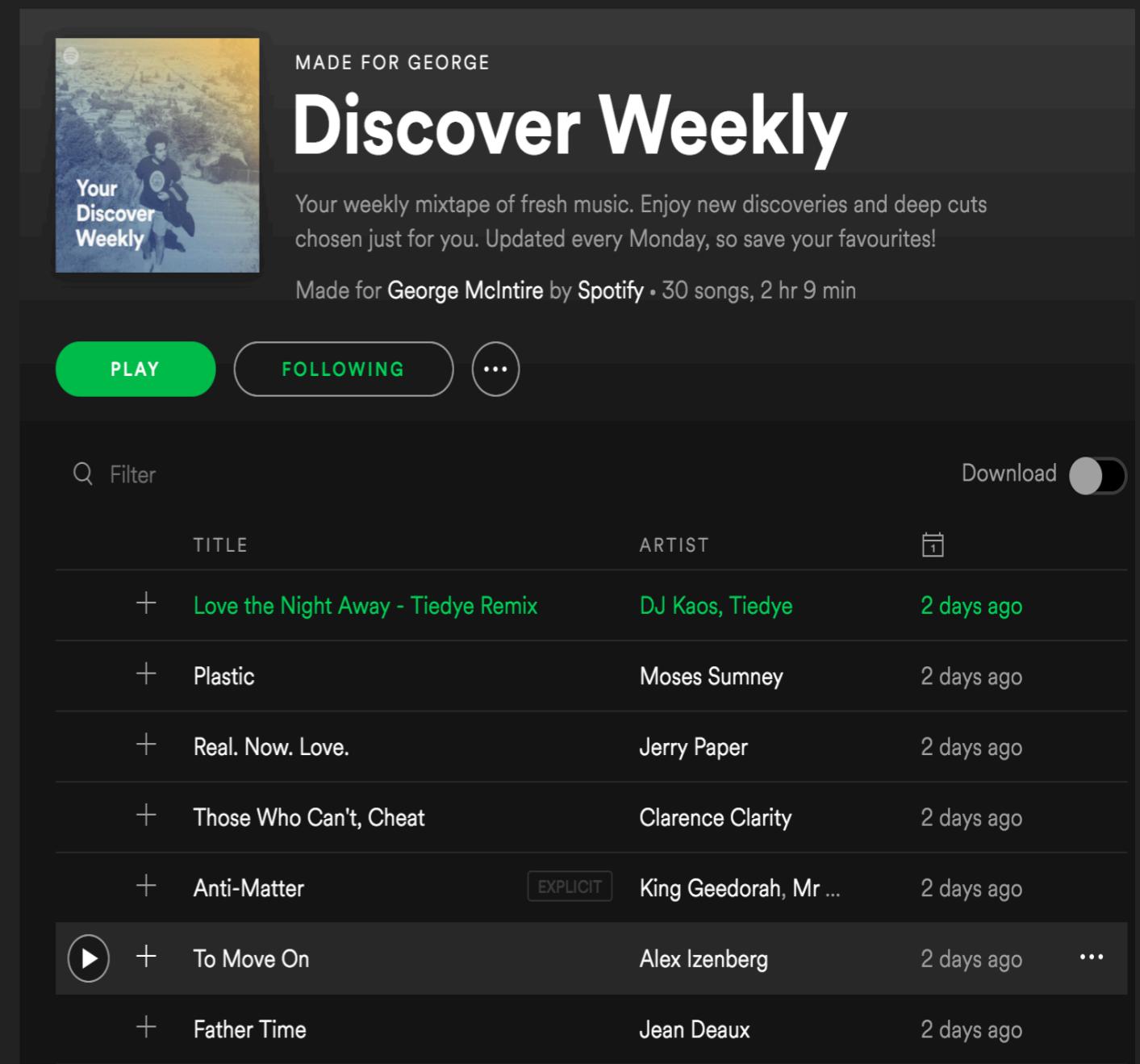


?



EXAMPLES

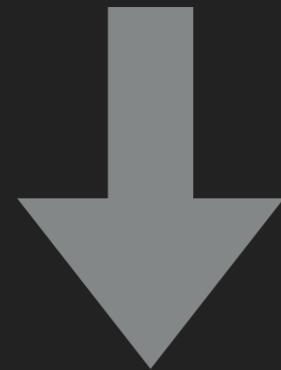
- ▶ Netflix and Spotify recommendations.
- ▶ Credit card fraud detections
- ▶ Loan approvals
- ▶ Zillow's Zestimate
- ▶ Email spam prevention/fake news
- ▶ Personal assistant machines: Siri, Alexa, etc...
- ▶ Crime pattern detections



TYPES OF MACHINE LEARNING

SUPERVISED

UNSUPERVISED



MAKING PREDICTIONS

FINDING STRUCTURES

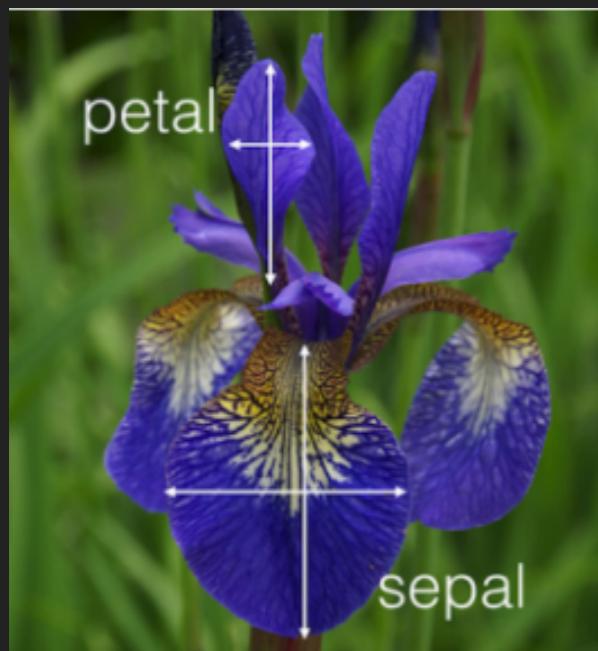
SUPERVISED LEARNING

SUPERVISED LEARNING

- ▶ Main goal is making predictions/classifications. Uses the past to predict the future.
- ▶ Data is composed of observations/events/instances.
- ▶ Predictors aka “X” aka the independent variables aka the features aka the input aka the attributes.
- ▶ Response variable aka “Y” aka the outcome aka the label aka the target aka the dependent variable.

SUPERVISED LEARNING DATASET

OBSERVATIONS



Fisher's Iris Data

Sepal length ↕	Sepal width ↕	Petal length ↕	Petal width ↕	Species ↕
5.1	3.5	1.4	0.2	<i>I. setosa</i>
4.9	3.0	1.4	0.2	<i>I. setosa</i>
4.7	3.2	1.3	0.2	<i>I. setosa</i>
4.6	3.1	1.5	0.2	<i>I. setosa</i>
5.0	3.6	1.4	0.2	<i>I. setosa</i>
5.4	3.9	1.7	0.4	<i>I. setosa</i>
4.6	3.4	1.4	0.3	<i>I. setosa</i>
5.0	3.4	1.5	0.2	<i>I. setosa</i>

PREDICTORS

RESPONSE

TYPES OF SUPERVISED LEARNING

CLASSIFICATION

- ▶ Outcome variable is a category:
 - ▶ good/bad
 - ▶ 1/0
 - ▶ sports/tech/politics/style
- ▶ Types of algorithms:
 - ▶ Logistic Regression, Naive Bayes, K-Nearest Neighbors, Decision Tress

REGRESSION

- ▶ Outcome variable is a continuous:
 - ▶ 3.6, 9.7, 2.3, 8.9, 11.1, 18.3, 23.6, 4.2, 6.9
- ▶ Types of algorithms:
 - ▶ Linear regression, Ridge regression, Lasso Regression

CLASSIFICATION EXAMPLE: LOAN DEFAULTS

- ▶ **Problem:** Lenders lose money when loanees fail to pay back loans.
- ▶ **Goal:** Develop a system that can efficiently identify high risk loans so lenders know which applications to reject
- ▶ **Data:** Records of previous loans marked as successful or failure that includes relevant information such as income, credit score, loan term, loan amount, etc...



REGRESSION EXAMPLE: HOUSING PRICES

- ▶ **Problem:** A home owner wants to sell her home but can't decide on an asking price.
- ▶ **Goal:** Accurately appraise the true value of the property
- ▶ **Data:** Home sale records labelled with their sale prices.
Dataset features # bedrooms/bathrooms, sq ft, location, year sold, etc...



UNSUPERVISED LEARNING

UNSUPERVISED LEARNING

- ▶ Absence of outcome variable/labels. Only features.
- ▶ Objective is looser and more exploratory.
 - ▶ Find groups of observations that exhibit similar characteristics.
 - ▶ Find combinations of features that explain the variation in the data.
- ▶ Useful as a preprocessing/exploratory data analysis step but too difficult to evaluate how well you're doing.

TYPES OF UNSUPERVISED LEARNING

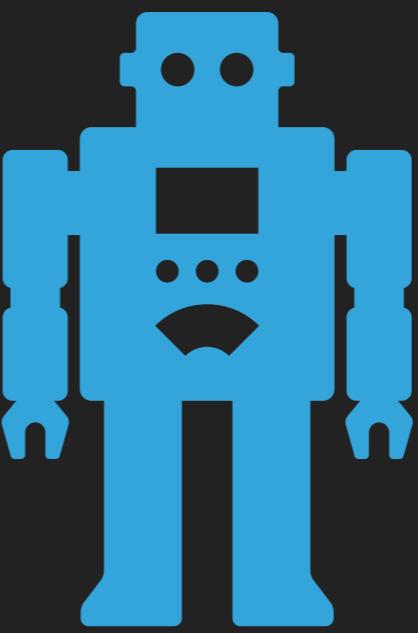
CLUSTERING

- ▶ Make labels from unlabelled data.
- ▶ Examples: detect segment of users, derive micro-positions in sports
- ▶ Algorithms: KMeans, Hierarchical, DBScan

DIMENSIONALITY REDUCTION

- ▶ Deals with two many variables.
Compresses data
- ▶ Great visualizing data with $>=4$ dimensions
- ▶ Algorithms: PCA, Truncated SVD, NMF

?? QUESTIONS ??



FINAL PROJECT

FINAL PROJECT OBJECTIVE

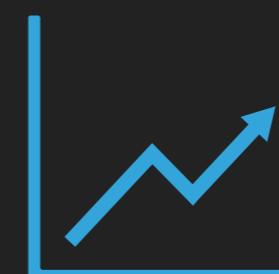
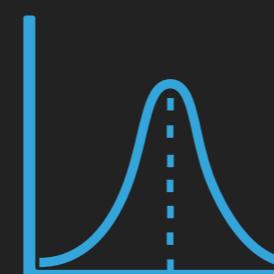
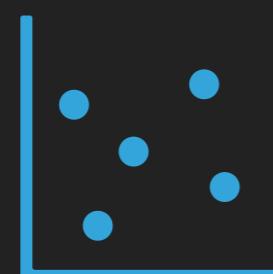
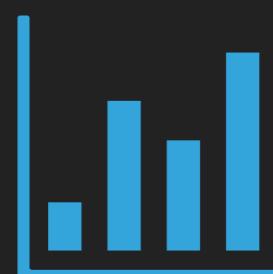
- ▶ Conduct a machine learning project in which you build a machine learning model trained on a dataset of your choosing.
- ▶ Use the models and tools we learned in this class to create the best possible model.
- ▶ Choose classification or regression.
- ▶ Unsupervised/clustering is allowed, but I prefer you choose the supervised route.
- ▶ If you want to use text data, then please let me know ASAP.

FINAL PROJECT PRESENTATION.

- ▶ 1. Introduce issue/problem and the dataset.
 - ▶ Give us context. Demonstrate your domain expertise.
- ▶ 2. State objective
 - ▶ “I am trying to use X to predict y”, “I am trying to analyze the relationship between A and B”, “I want to build a loan approval model”

FINAL PROJECT PRESENTATION 2

- ▶ 3. Data acquisition/wrangling/munging/cleaning
 - ▶ Tell us where the data came from. Which steps did you take to transform this dataset in to a workable machine-learning dataset.
- ▶ 4. Exploratory data analysis.
 - ▶ Use visualizations to give the audience a better understanding of the data and the context. Show us what you found interesting about the data. Charts, charts, charts!



FINAL PROJECT PRESENTATION 3

- ▶ 5. Modeling
 - ▶ Which algorithms did you test? Which algorithms and configurations did you choose as your model? What steps did you take to improve your model?
 - ▶ What sort of feature engineering steps did you take to improve the model? Which features did you keep, drop, and transform? What were the best features?

FINAL PROJECTION PRESENTATION 4.

- ▶ 4. Model evaluation
 - ▶ Show how well your best model did. Use accuracy, sensitivity, recall, precision, and roc auc metrics. Highly recommend you include a roc curve. Scores should be cross validated or tested against a validation set.
- ▶ 5. Conclusion/further steps
 - ▶ What did you learn? How much of your objective did you achieve? What else could be done to solve this problem?

EXAMPLES AND DATASETS

- ▶ Project examples from past GA students:
 - ▶ <https://github.com/ga-students/DS-SF-40/blob/master/project-examples.md>
- ▶ Datasets galore:
 - ▶ https://github.com/ga-students/DS-SF-40/blob/master/public_data.md
 - ▶ <https://www.reddit.com/r/datasets>
 - ▶ <https://www.kaggle.com/datasets>
 - ▶ driven-by-data.net
 - ▶ <https://data.world/>
 - ▶ <https://github.com/caesar0301/awesome-public-datasets>
- ▶ Can use work data, ask for permission early and deal with privacy related issues.

FINAL PROJECT PROPOSAL

- ▶ What are two or three potential questions you would like to try and answer? What do you want to find out in this project? If possible state your objective.
- ▶ What is the data? Where did it come from? Why did you choose it?
- ▶ Proposal is due week of December 5. After proposal is submitted, schedule a 1-1 feedback session with me.
- ▶ More details here: [https://github.com/ga-students/DS-SF-40/
blob/master/project.md](https://github.com/ga-students/DS-SF-40/blob/master/project.md)

FINAL PROJECT TIPS

- ▶ Pick a machine learning dataset. Not all datasets can be used for machine learning.
- ▶ Don't pick a very dirty dataset or any other dataset that requires tremendous time and effort to assemble.
- ▶ **ITS NEVER TOO EARLY TO START WORK ON YOUR FINAL PROJECT**
- ▶ Time management is especially crucial. Budget your time and take advantage of the holiday breaks.
- ▶ Use me and your fellow classmates for your help.

?? QUESTIONS ??