BREAK THROUGH TECH

Machine Learning Foundations

Lab 3

Week of June 9

BREAK
THROUGH
TECH

Week 3 Program Announcements

FAQs: Where can I find...?

- Academic support from peers and instructional staff on course content
  - → **Canvas Discussion Board**

- Community-building with peers and in-class lab discussions
  - → **Slack** (If you haven't joined Slack yet, ask me for a join link)

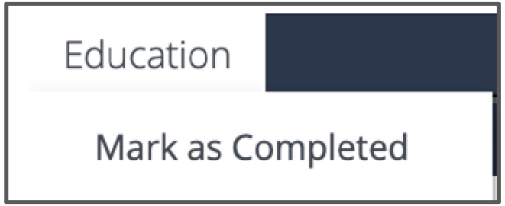
- Answers to programmatic and logistical questions from program staff
  - → **Support form** (https://bit.ly/4bDqjEV)

- Lab recordings and slide decks
  - → **"Live Recordings"** on the left side menu in Canvas
  - Please allow 24 hours after your lab ends for the upload to be available

FAQs: Grades and Assignments

- **Late work will not be accepted and assignment extensions are not permitted**
- All assignments are due by 11:59 pm Eastern Time on the deadline indicated in Canvas
- If you have questions regarding a graded assignment, contact your grader on Canvas

- For every coding assignment on Canvas, please remember to follow the instructions to **"Mark as Completed"** in order to submit your assignment successfully. Missing this step will jeopardize your work on the assignment – our auto-grader will not count your submission and you may receive a 0%.



Today's Agenda

(10 minutes)	Icebreaker
(25 minutes)	Week 3 Concept Overview + Q&A
(20 minutes)	Breakout Groups: Big Picture Questions
(10 minutes)	Class Discussion
(10 minutes)	Break
(35 minutes)	Breakout Groups: Lab Assignment Working
Session 1	
(10 minutes)	Working Session 1 Debrief
(35 minutes)	Breakout Groups: Lab Assignment Working
Session 2	
(10 minutes)	Working Session 2 Debrief
(15 minutes)	Concluding Remarks & Survey

BREAK
THROUGH
TECH

Pride Month

Icebreaker: Pride Month

Objectives:

- Acknowledge the contributions of marginalized folks committed to inclusion and tech equity
- Share something about yourself that you’re proud of
- Recognize your peers’ strengths and accomplishments

AI influencers & Tech Advances

Joy Buolamwini

Latanya Sweeney

Timnit Gebru

Explainable A.I.

Differential Privacy

Icebreaker: Pride Month

Students: What are you proud of?

BREAK
THROUGH
TECH

Week 3 Concept Overview + Q&A

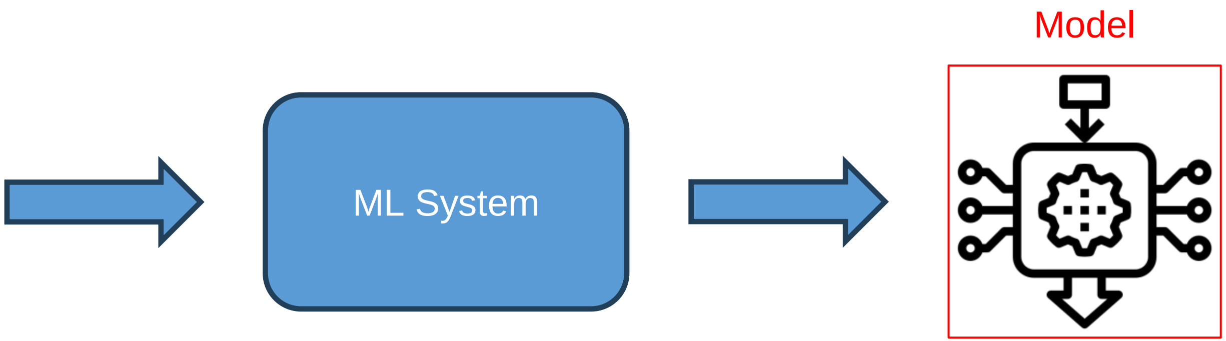
Week 3 Overview

This week you explored a number of topics. To refresh your memory, your goals were to:

- Define the core foundational elements of model training and evaluation
- Develop intuition for different classes of algorithms
- Analyze the mechanics of two popular supervised learning algorithms: decision trees and k-nearest neighbors
- Develop intuition on tradeoffs between different algorithmic choices

Supervised Machine Learning

X	Y	Label
1	4	+
3	4	+
1	2	+
5	4	+
3	1	-
3	2	-

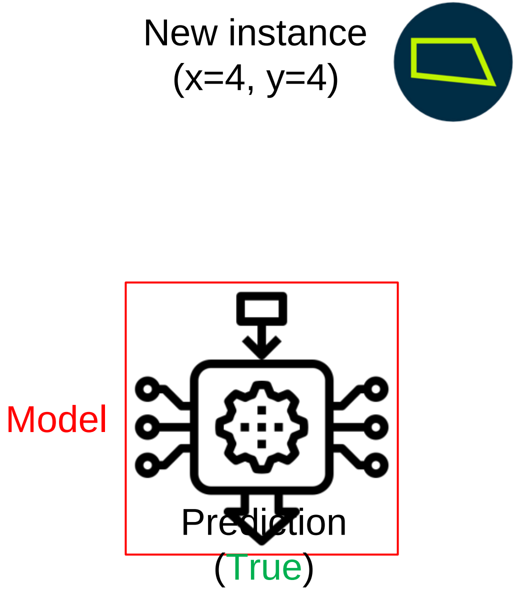


Goal: Create a model to "explain" training data and predict on new data

Learning is generalization

Goal: Create a model to "explain" training data and predict on new data

- Good model: Generalizes training data
- Want:
  - Low training error
  - Don't be "too good" on training data



Training Models With the Goal of Generalization: decision boundaries

1 squiggly line

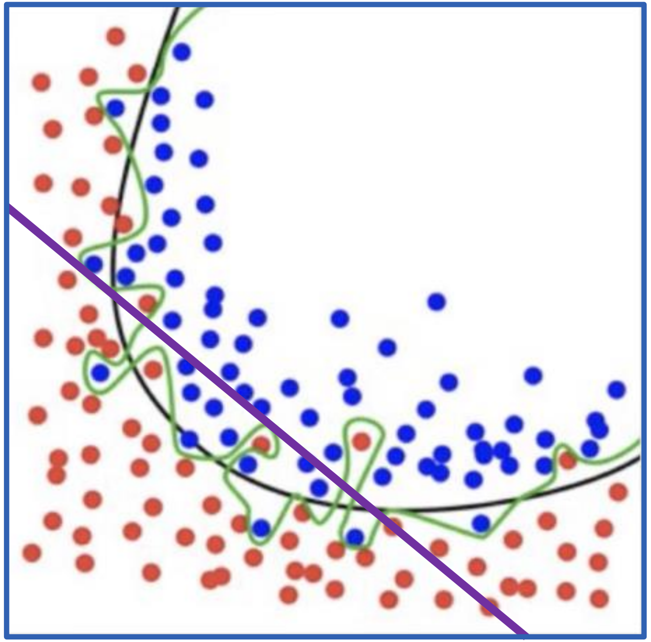
overfit

2 straight line

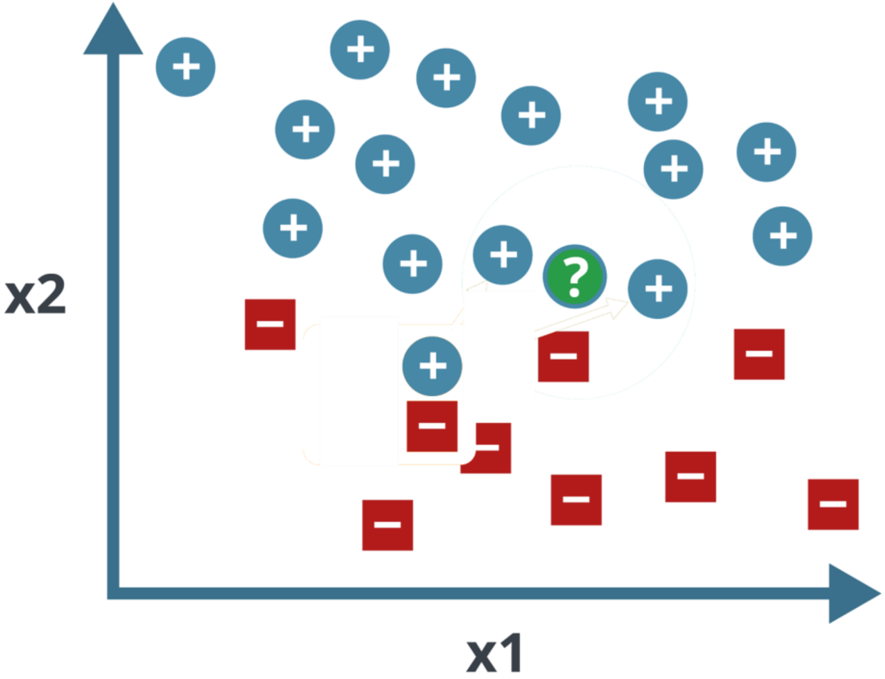
underfit

3 curve

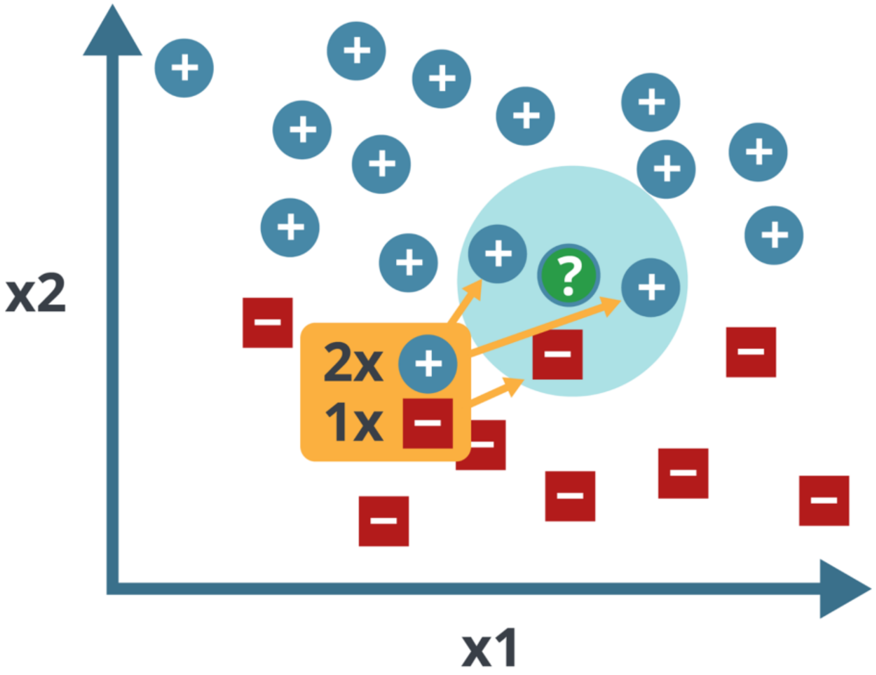
good



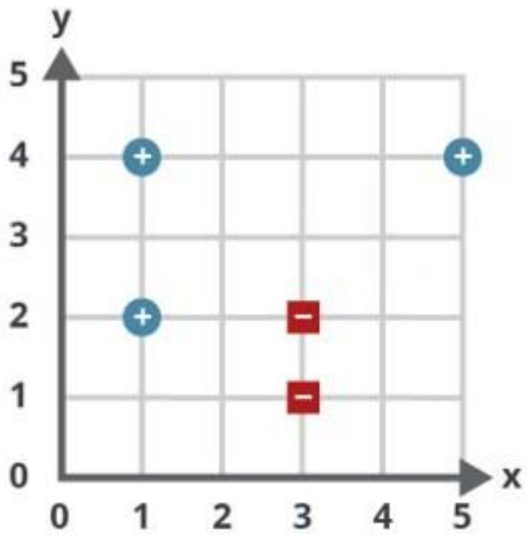
K Nearest Neighbors



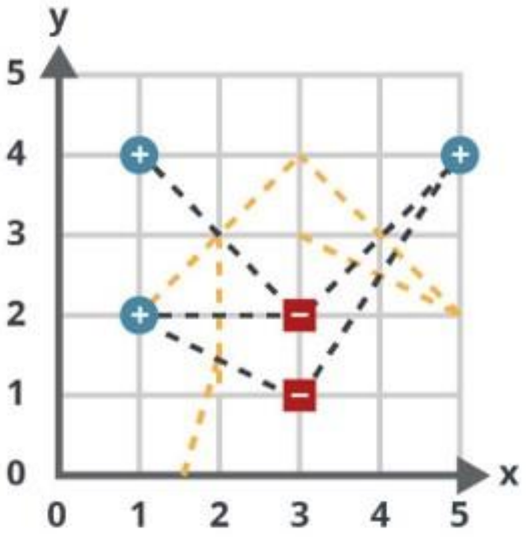
K = 3 Nearest Neighbors



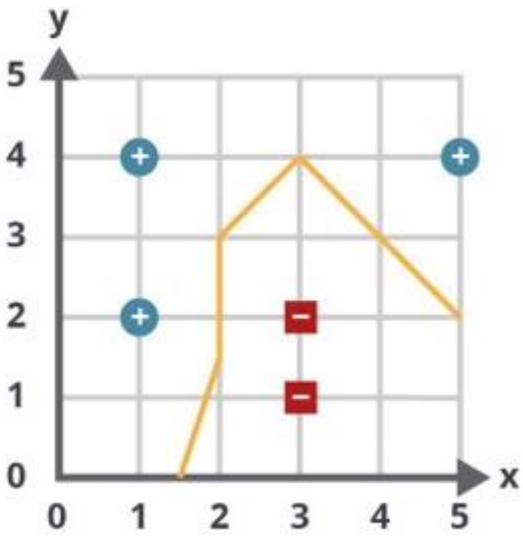
K-Nearest Neighbors (K = 1) Boundaries



K-Nearest Neighbors (K = 1) Boundaries



K-Nearest Neighbors (K = 1) Boundaries



Decision Trees: Understanding Entropy

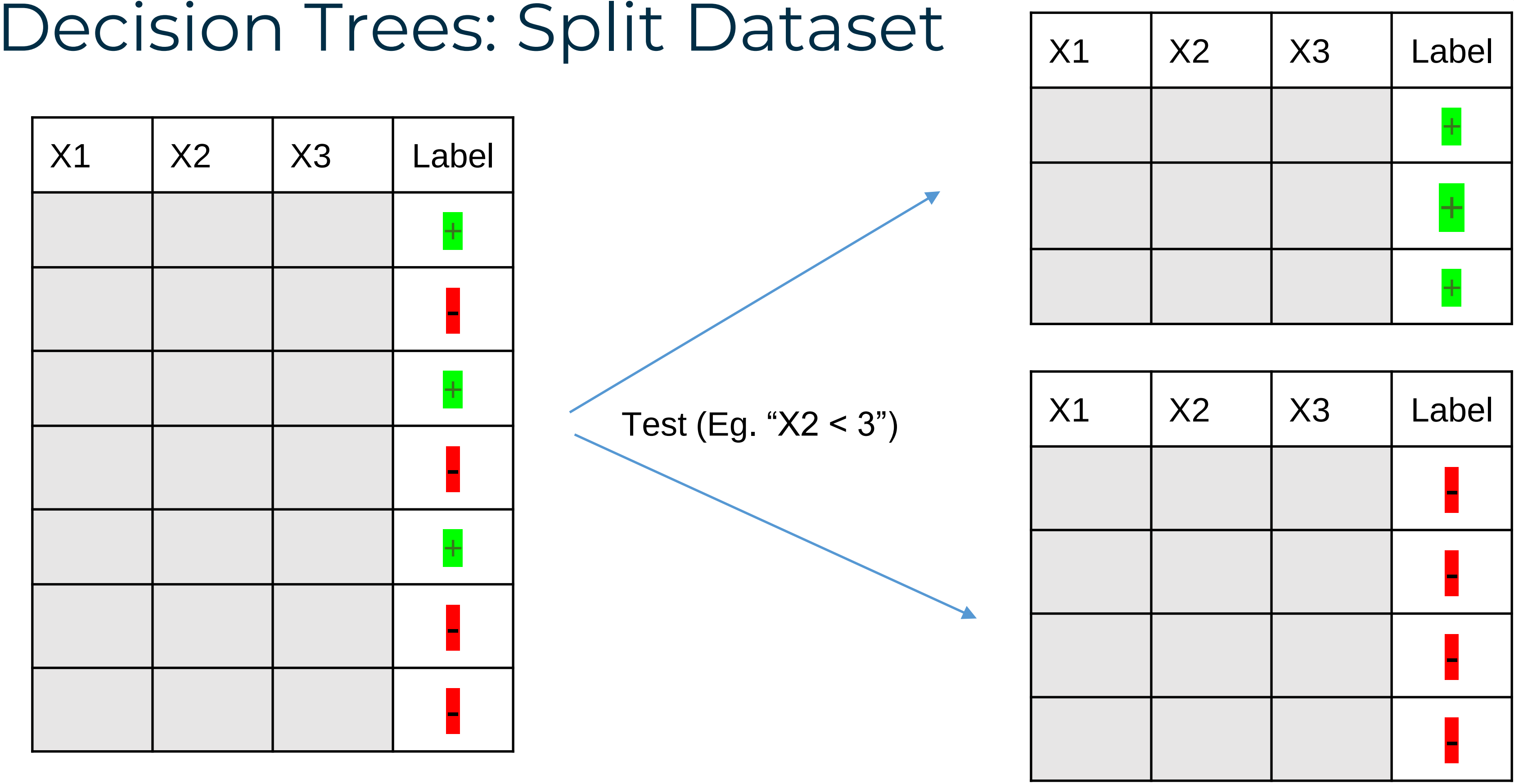
Predict the label

X1	X2	X3	Label
			+
			+
			+
			+
			+
			+
			-
			?

Entropy = Amount of uncertainty

X1	X2	X3	Label
			+
			-
			+
			-
			+
			-
			-
			?

Decision Trees: Split Dataset



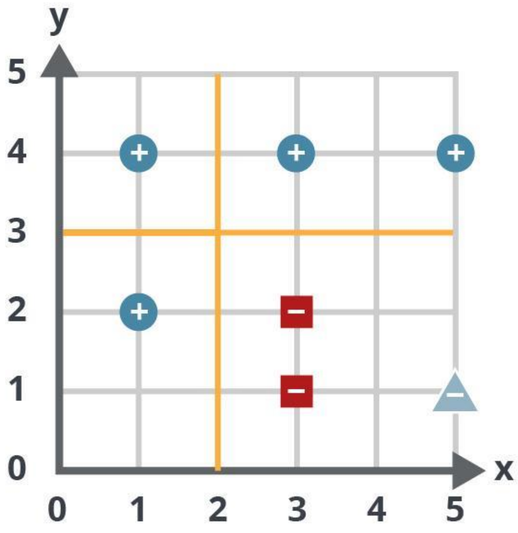
X1	X2	X3	Label
			+
			-
			+
			-
			+
			-
			-

Test (Eg. “X2 < 3”)

X1	X2	X3	Label
			+
			+
			+

X1	X2	X3	Label
			-
			-
			-
			-

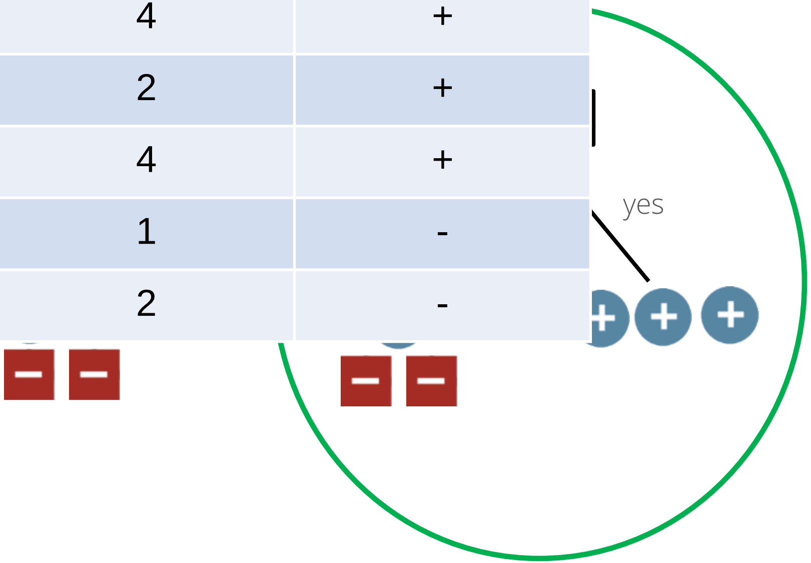
Decision Trees: Choosing a test



y > 3

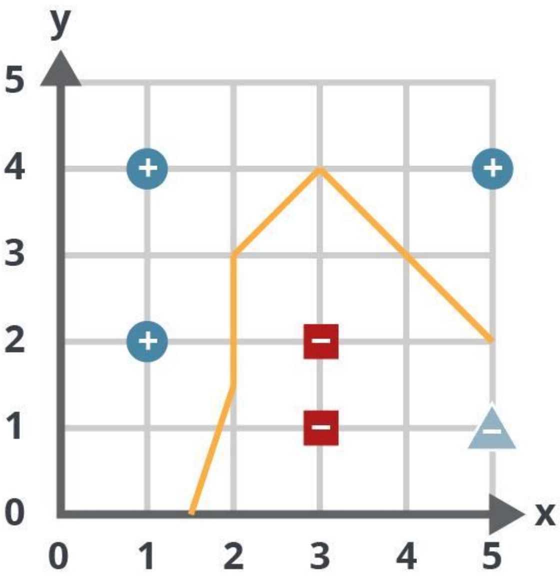
x > 2

X	Y	Label
1	4	+
3	4	+
1	2	+
5	4	+
3	1	-
3	2	-

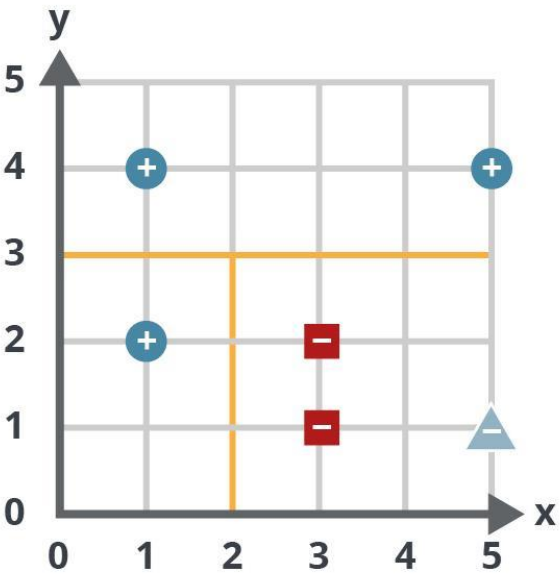


Higher Information Gain

KNN and Decision Trees



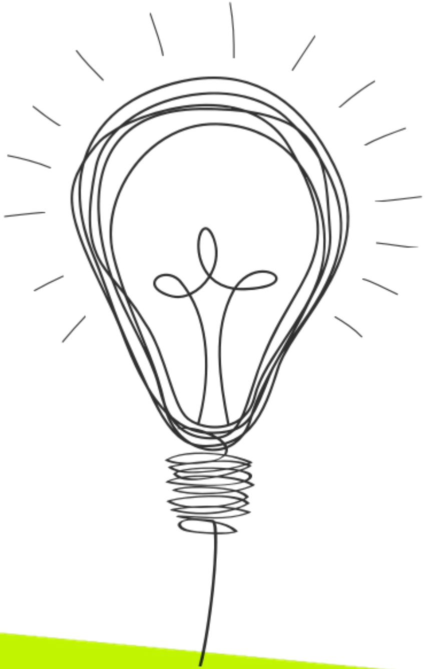
KNN, K=1 (1-NN)



Decision Tree: y > 3, x > 2

Questions & Answers

What questions do you have about the online content this week?



BREAK
THROUGH
TECH

Breakout Groups: Big Picture Questions

Big Picture Questions

You have 20 minutes to discuss the following questions within your breakout groups:

- How would you explain model complexity to a non machine learning person?
- What are the hyperparameters in KNN and decision trees? How do they impact the respective model's complexity?
- In less-technical terms, why do you think KNN and decision trees work? In other words, what is special about them that enables them to make accurate predictions on new data?
- How can you tell if a model is overfitting the data?
- How can you tell if a model is underfitting the data?

BREAK
THROUGH
TECH

Class Discussion

Break

BREAK
THROUGH
TECH

BREAK
THROUGH
TECH

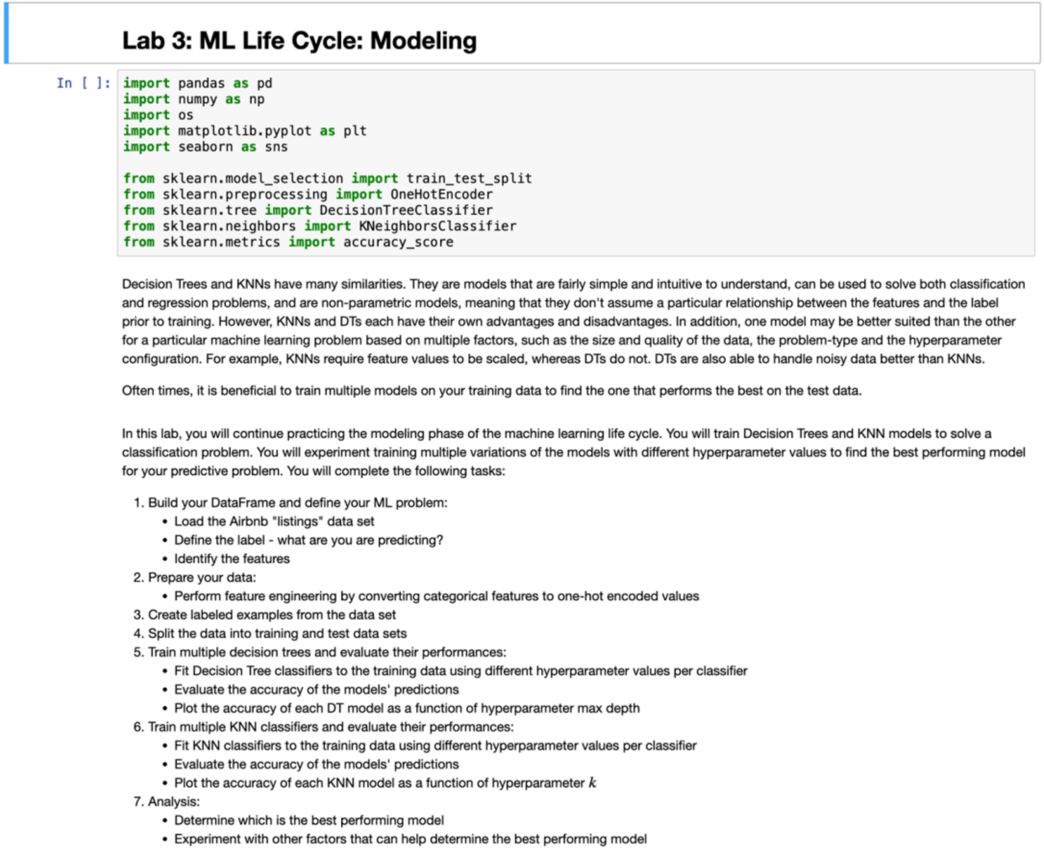
Breakout Groups: Lab Assignment

Lab 3

In this lab, you will:

- Define your ML problem and build your DataFrame
- Prepare your data:
  - Perform feature engineering by converting categorical features to one-hot encoded values
- Train multiple decision trees and evaluate their performances:
  - Train decision tree classifiers with various hyperparameter values.
  - Visualize and evaluate the accuracy of the models' predictions
- Train multiple KNN classifiers and evaluate their performances:
  - Train KNN classifiers with various hyperparameter values.
  - Visualize and evaluate the accuracy of the models' predictions
- Determine the best performing model for your predictive problem

Lab 3



Lab 3: ML Life Cycle: Modeling

```
In [ ]: import pandas as pd
        import numpy as np
        import os
        import matplotlib.pyplot as plt
        import seaborn as sns

        from sklearn.model_selection import train_test_split
        from sklearn.preprocessing import OneHotEncoder
        from sklearn.tree import DecisionTreeClassifier
        from sklearn.neighbors import KNeighborsClassifier
        from sklearn.metrics import accuracy_score
```

Decision Trees and KNNs have many similarities. They are models that are fairly simple and intuitive to understand, can be used to solve both classification and regression problems, and are non-parametric models, meaning that they don't assume a particular relationship between the features and the label prior to training. However, KNNs and DTs each have their own advantages and disadvantages. In addition, one model may be better suited than the other for a particular machine learning problem based on multiple factors, such as the size and quality of the data, the problem-type and the hyperparameter configuration. For example, KNNs require feature values to be scaled, whereas DTs do not. DTs are also able to handle noisy data better than KNNs.

Often times, it is beneficial to train multiple models on your training data to find the one that performs the best on the test data.

In this lab, you will continue practicing the modeling phase of the machine learning life cycle. You will train Decision Trees and KNN models to solve a classification problem. You will experiment training multiple variations of the models with different hyperparameter values to find the best performing model for your predictive problem. You will complete the following tasks:

1. Build your DataFrame and define your ML problem:
    - Load the Airbnb "listings" data set
    - Define the label - what are you are predicting?
    - Identify the features
2. Prepare your data:
    - Perform feature engineering by converting categorical features to one-hot encoded values
3. Create labeled examples from the data set
4. Split the data into training and test data sets
5. Train multiple decision trees and evaluate their performances:
    - Fit Decision Tree classifiers to the training data using different hyperparameter values per classifier
    - Evaluate the accuracy of the models' predictions
    - Plot the accuracy of each DT model as a function of hyperparameter max depth
6. Train multiple KNN classifiers and evaluate their performances:
    - Fit KNN classifiers to the training data using different hyperparameter values per classifier
    - Evaluate the accuracy of the models' predictions
    - Plot the accuracy of each KNN model as a function of hyperparameter k
7. Analysis:
    - Determine which is the best performing model
    - Experiment with other factors that can help determine the best performing model

BREAK
THROUGH
TECH

Working Session 1 Debrief

Lab Debrief

So far,

- What did you enjoy about this lab?
- What did you find difficult about this lab?

BREAK THROUGH TECH

Working Session 2 Debrief

Lab Debrief

- What did you enjoy about this lab?
- What did you find hard about this lab?
- What questions do you still have about this lab?
- How did you approach problem-solving during the exercise?
- What would you do differently if you were to repeat the exercise?

Concluding Remarks

Concluding Remarks

- Key takeaways
- Additional resources

Next week

In the following week, you will:

- Analyze the mechanics of logistic regression
- Understand the purpose of using gradient descent and loss functions
- Explore common hyperparameters for logistic regression
- Define the core math concepts required to solve common machine learning problems
- Use NumPy to perform vector and matrix operations
- Explore how linear regression works to solve real-world regression problems

And in the lab, you will:

- Write a Python class that will train a logistic regression model and make predictions
- Build your DataFrame and define your ML problem
- Create labeled examples from the data set
- Train a logistic regression model using your class
- Compare your Python implementation to scikit-learn's implementation.

BREAK
THROUGH
TECH

Content + Lab Feedback Survey

Weekly Survey + Early Program Survey

To complete your lab, please answer the following questions about BOTH your online modules and your lab experience. Your input will help pay it forward to the Break Through Tech student community by enabling us to continuously improve the learning experience that we provide to our community.

Thank you for your thoughtful feedback!

Weekly Content + Lab feedback: https://forms.gle/eUQQZgS6BPRpqgZ7A

Early-Program Feedback: https://forms.gle/dNRZnSmXzY3Y49ws5