ML Tutorial 1

Back to Basics

Wendy Carande - LASP/CU Boulder

SWx TREC

Deep Learning Laboratory

Tutorial Rules

Have fun!

Feel free to submit to the Kaggle leaderboard if you have a new/better solution.

Think like a scientist. Data science has that name for a reason. You should slow down and think about hypotheses, critically evaluate data, etc.

For a first pass, I recommend you follow along. I’ll give you freeform time to play around with different models, etc at the end.

I will post the “answers” to my github at the end if you want to reference them.

Setting the Scene

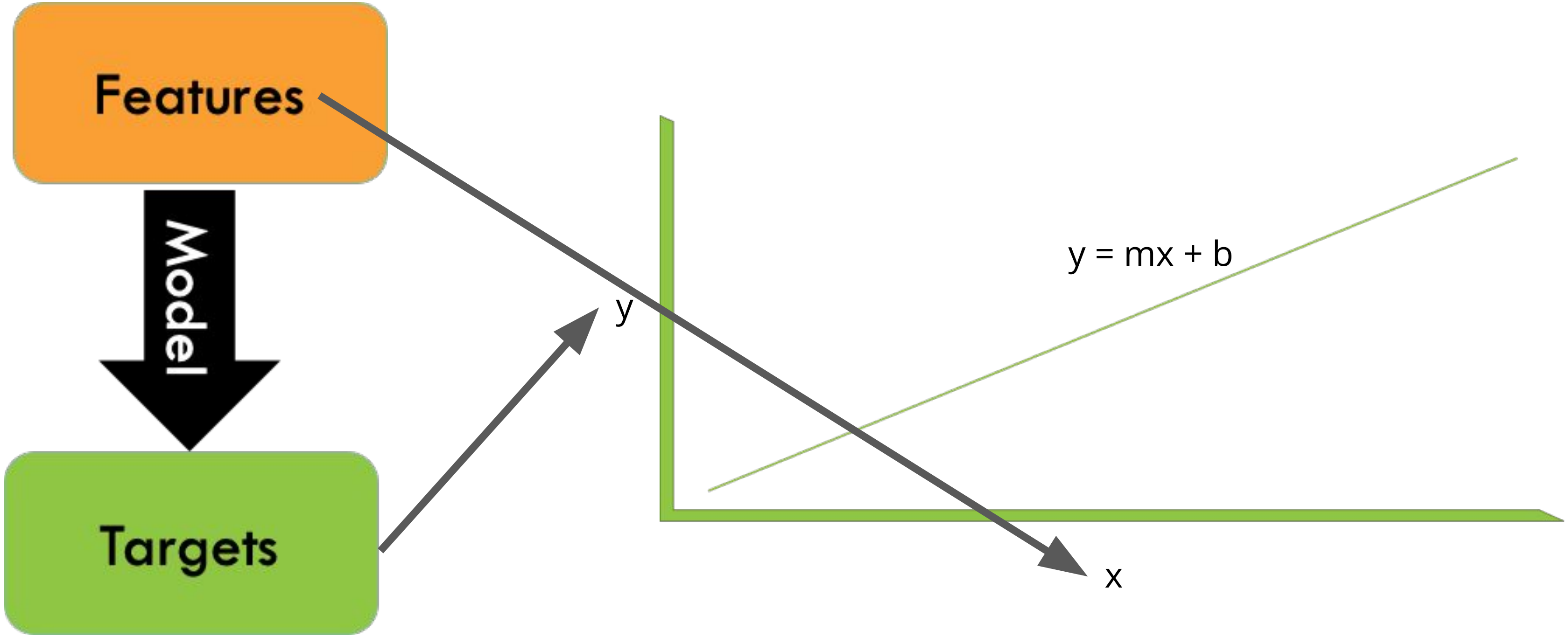
Year: 2912

Spaceship Titanic hits spacetime anomaly and mistakenly transports thousands of passengers to an alternate dimension!

Summary: The ship is intact, but some passengers are transported and still missing. If we can identify who the missing passengers are, we can transport them back to the ship. We have the transported/not transported data for some of the passengers (training data), but we are missing the transported/not transported data for a subset of passengers. In order to recover these passengers, we need to submit a list of passenger IDs and whether or not the passenger was transported (True/False) to headquarters. We have a variety of other information about the passengers.

Credit: https://www.kaggle.com/c/spaceship-titanic/overview

Features in, targets out



Features

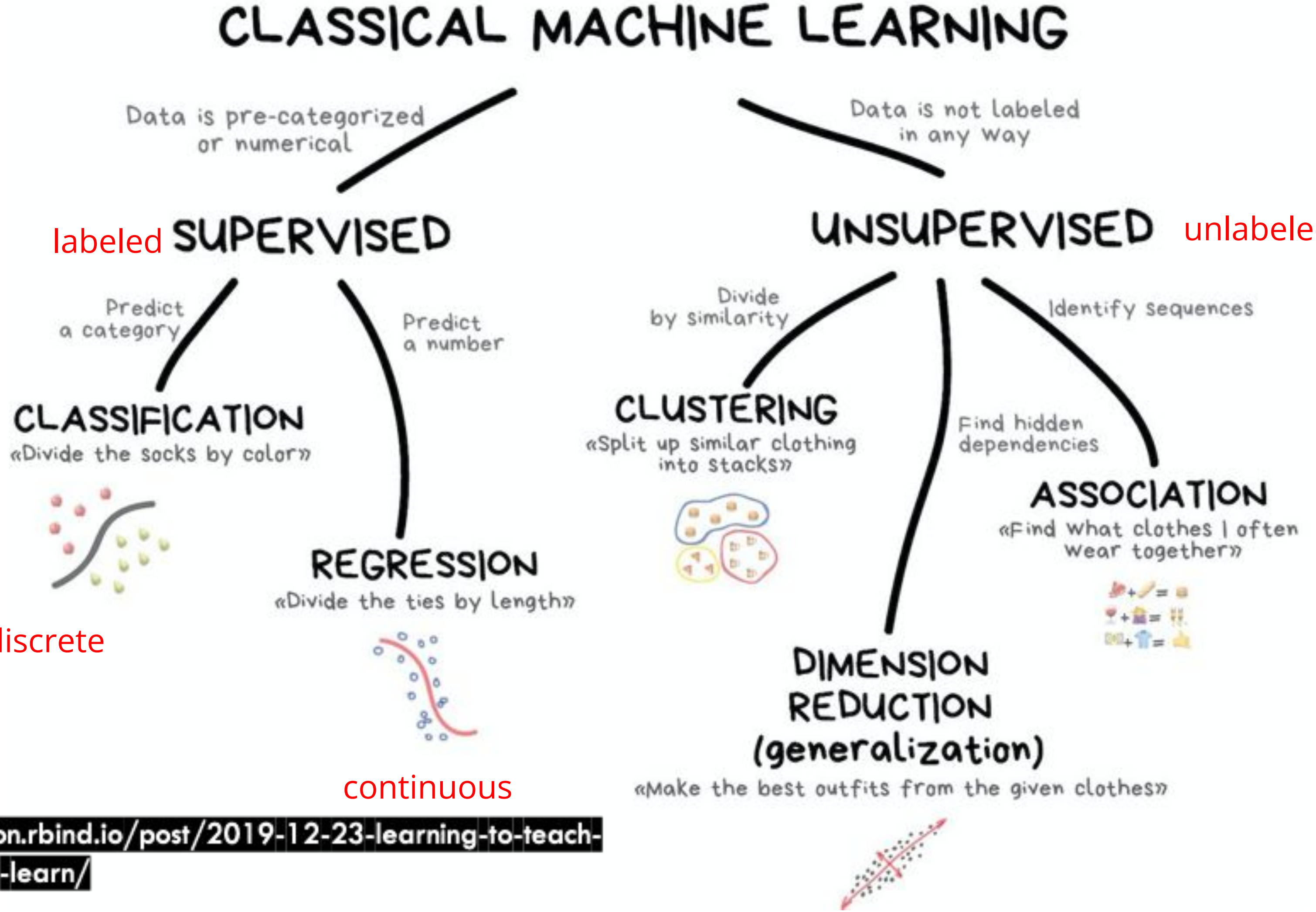
Speed

Legendary

Type

Defense

HP


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

labeled

unlabeled

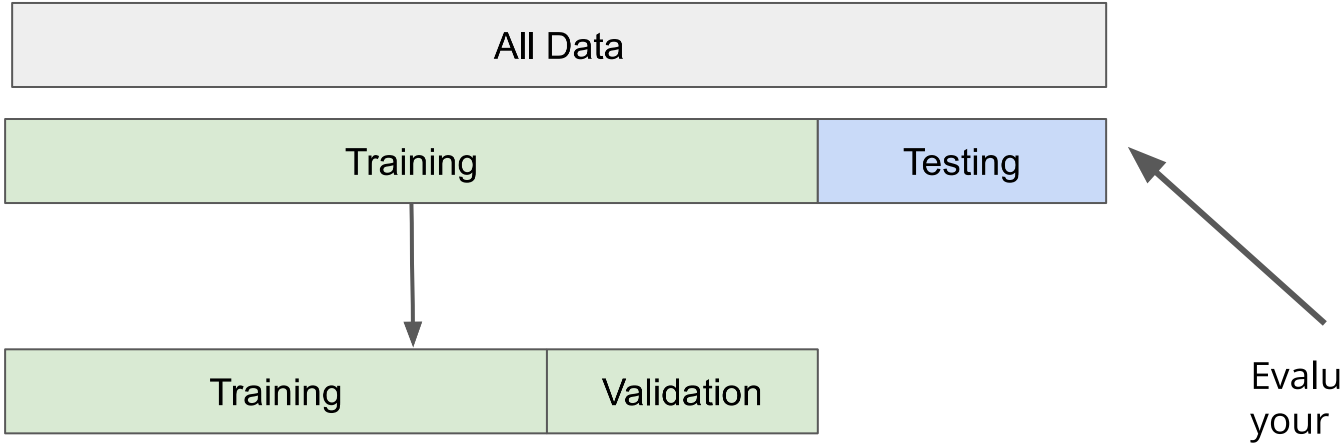
discrete

continuous

https://alison.rbind.io/post/2019-12-23-learning-to-teach-machines-to-learn/

Data Splitting & Evaluation

data split



All Data

Training

Testing

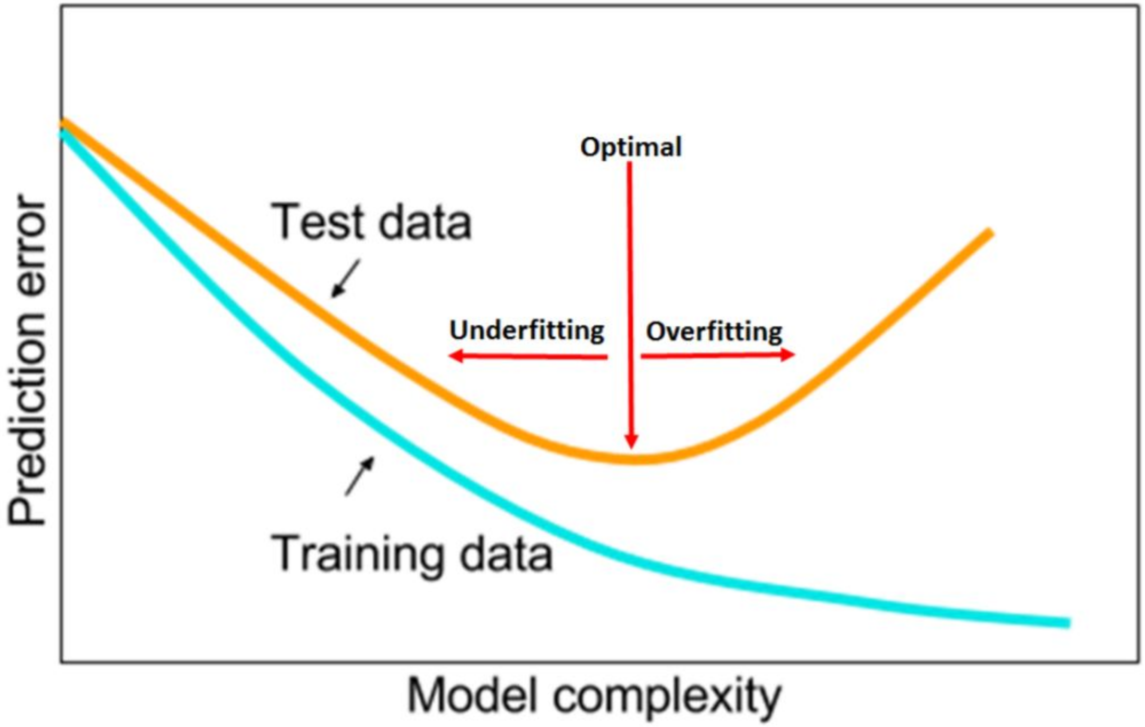
Training

Validation

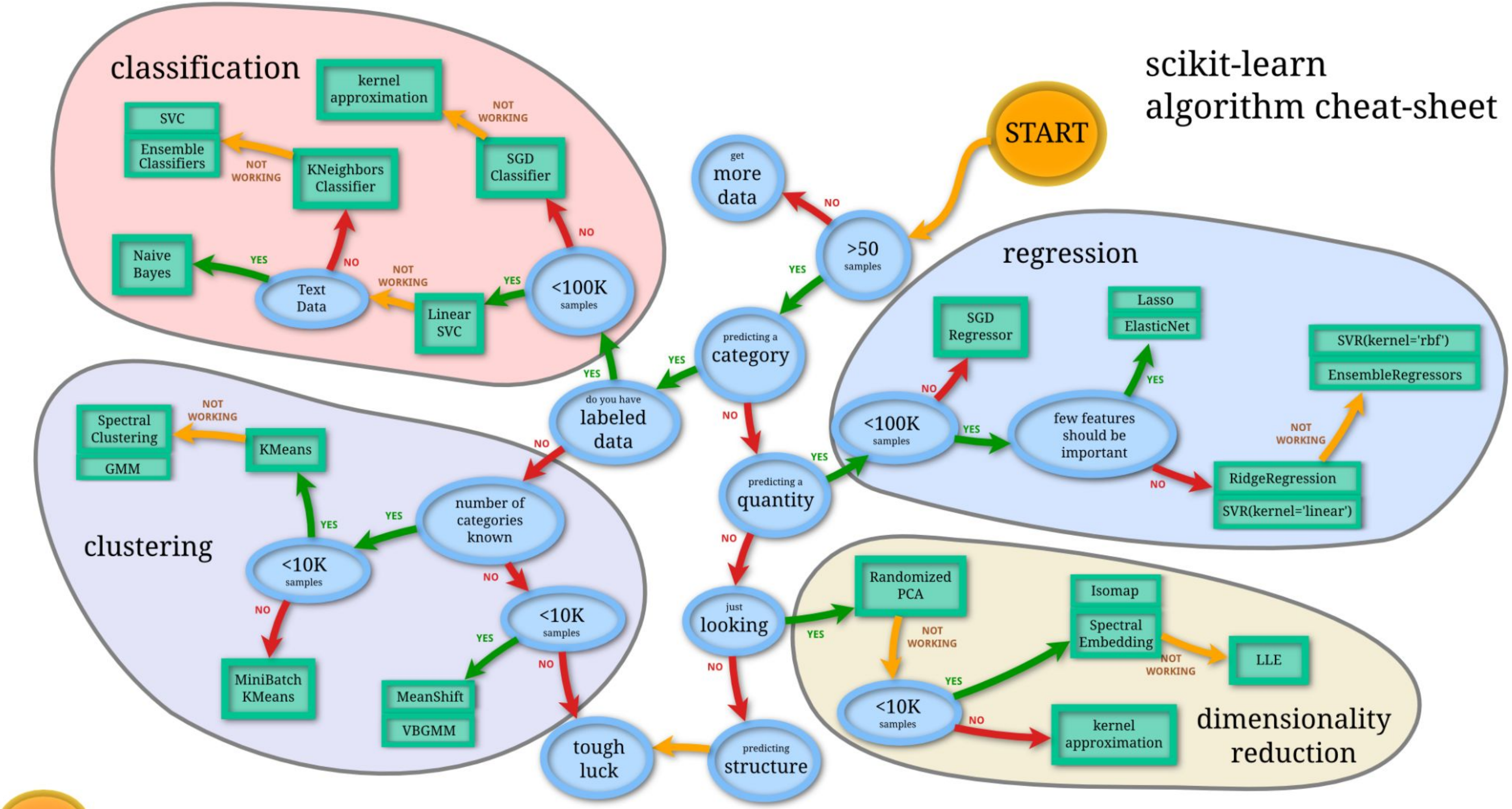
Validation is where you tune hyperparameters

Evaluate data your model has never seen

Model Tuning



https://towardsdatascience.com/hyper-parameter-tuning-techniques-in-deep-learning-4dad592c63c8



Back

scikit learn

https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

sklearn basics

Choose your model

Ex: clf = RandomForestClassifier(max_depth=2, random_state=0)

Fit

Ex: clf.fit(x_train, y_train)

Predict

Ex: y_pred = clf.predict(x_test)

Score

Ex: clf.score(x_test, y_test)