Machine Learning Methodology

Systematic Process | Workflow | Pipeline | Steps

Nury Yuleny Arosquipa Yanque

Department of Computer Science
Institute of Mathematics and Statistics (IME)
University of São Paulo (USP)

Schedule

Terminology

Machine Learning Workflow

Present results

Conclusions

Terminology

What is machine learning?

[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.
Arthur Samuel, 1959

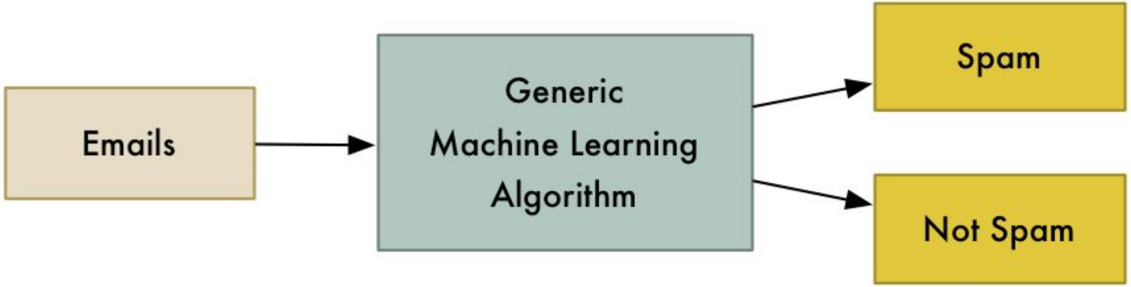
A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.
Tom Mitchell, 1997

Example

T := Classifying emails as spam or not spam.

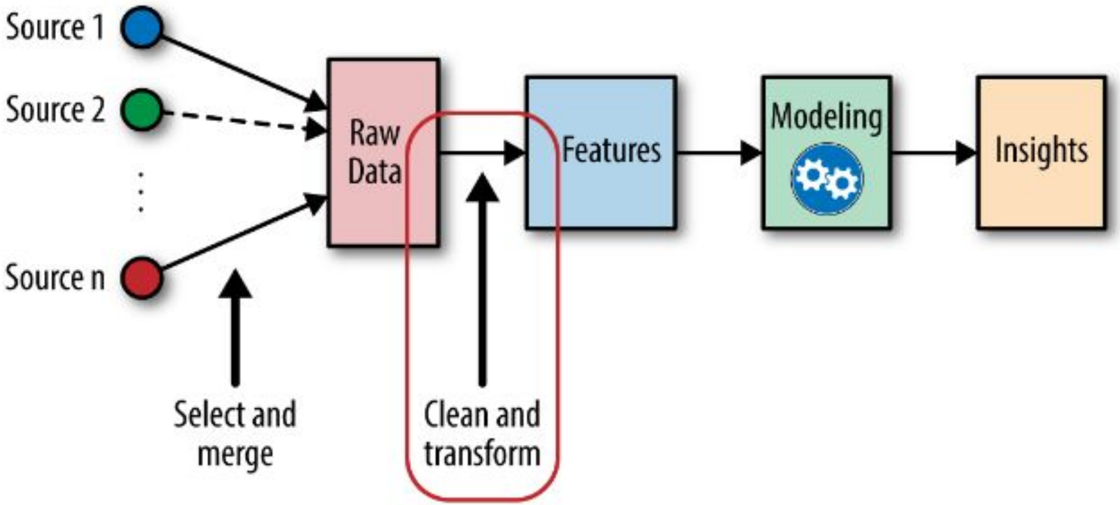
E := Watching you label emails as spam or not spam.

P := The number (or fraction) of emails correctly classified as spam/not spam.



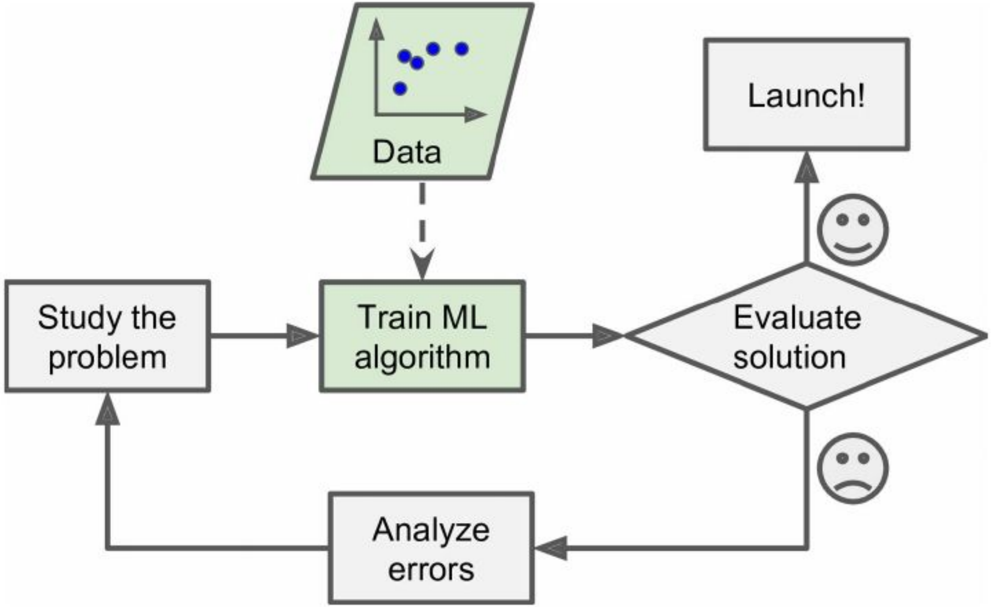
Key concepts

- Data
- Tasks
- Models
- Features
- Model Evaluation



Machine Learning Workflow

1. Define problem
2. Data pre-processing
3. Model design
4. Improve results
5. Using the Model



1. Define problem

Step 1: **What** is the problem?

Step 2: **Why** does the problem need to be solved?

Step 3: **How** would I solve the problem?

Problem Definition Framework

Step 1: What is the problem?

Describe the problem, list assumptions and similar problems.

Step 2: Why does the problem need to be solve?

Motivation for solving the problem, benefits a solution provides and how the solution will be used.

Step 3: How would I solve the problem?

Describe how the problem would be solved.

2. Data pre-processing

Machine learning algorithms learn from data.

It is critical that you feed them the right data for the problem you want to solve.

Even if you have good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included.

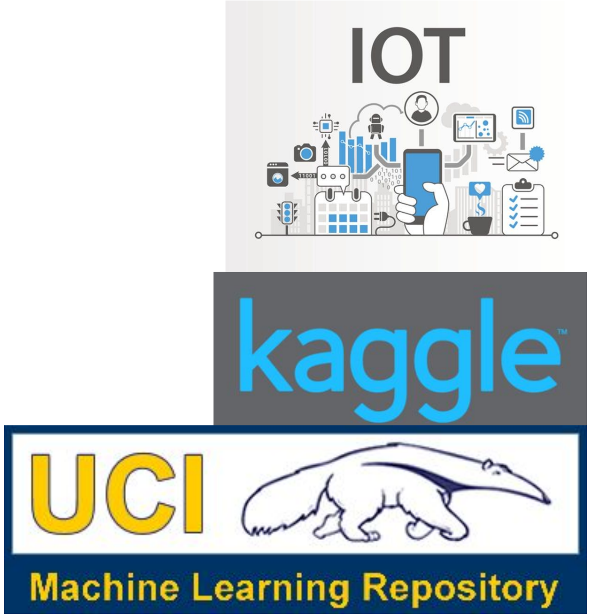
Gathering data

Popular open data repositories:

- UC Irvine Machine Learning Repository
- Kaggle datasets
- Amazon’s AWS datasets

Meta portals:

- http://dataportals.org/
- http://opendatamonitor.eu/
- http://quandl.com/



Gathering data

Other pages listing many popular open data repositories:

- Wikipedia’s list of Machine Learning datasets
- Quora question: “Where-can-I-find-datasets-for-machine-learning”
- Datasets subreddit

Types of data

- **Numeric** e.g. age
- **Categorical** e.g. gender, nationality
- **Ordinal** e.g. low/medium/high

Pre-processing

Most of the real-world data is messy:

1. Missing data: missing values and/or attributes (salary = "").

2. Noisy data: data with errors and/or outliers (salary = -150).

3. Inconsistent data: have discrepancies in codes and names (1→A, 2→B, 3→C).

Data Preparation Process

Step 1: Data Selection

What data is available, what data is missing and what data can be removed.

Step 2: Data Preprocessing

Organize your selected data by formatting, cleaning and sampling.

Step 3: Data Transformation

Transform preprocessed data ready for machine learning by engineering features.

EDA Methods

Descriptive statistics

Central tendency of a distribution

Mean	$\frac{\text{Sum of all values}}{\text{Total number of values}}$
Median	Middle value(when data are arranged in order
Mode	Most common value

Measure of Variation

Variance	**how far a set of numbers are spread out from mean**
Interquartile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

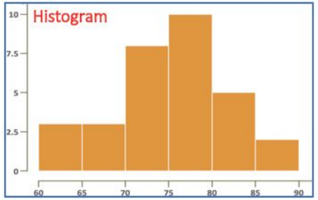
Skewness & Kurtosis

Skewness	**Measure of symmetry**
Kurtosis	**Kurtosis is a measure of "peakedness" relative to a Gaussian shape**

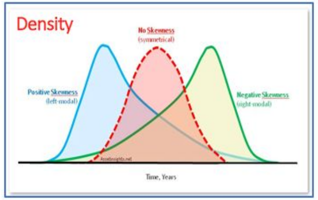
Visualizations

1-dimension

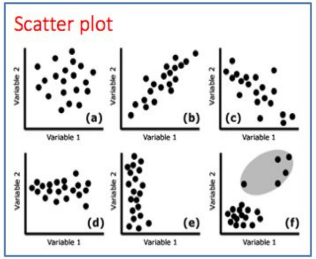
Few data points



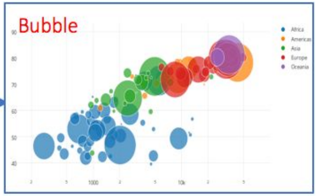
Many data points



2-dimension



3-dimension



3. Model Design

Standard methodology:

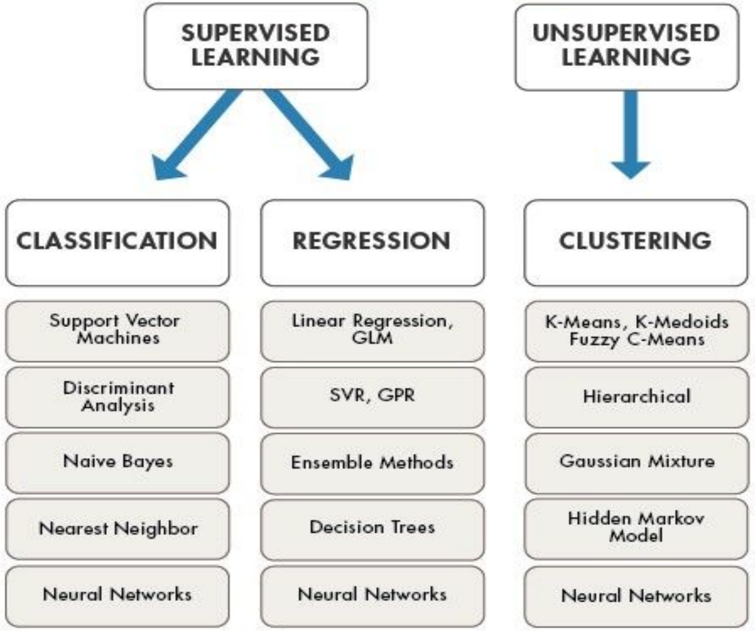
a. Collect large set of examples with correct classifications.

b. Divide collection into two disjoint sets: training and test.

c. Apply learning algorithm to training set giving hypothesis *H*.

d. Measure performance of *H* with respect to test set.

Researching the model



SUPERVISED LEARNING

UNSUPERVISED LEARNING

CLASSIFICATION

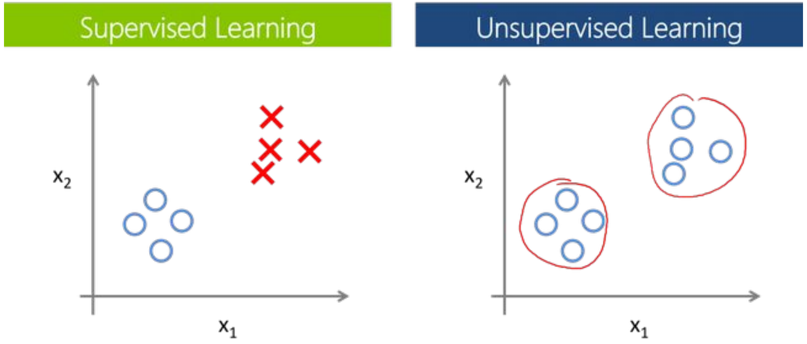
- Support Vector Machines
- Discriminant Analysis
- Naive Bayes
- Nearest Neighbor
- Neural Networks

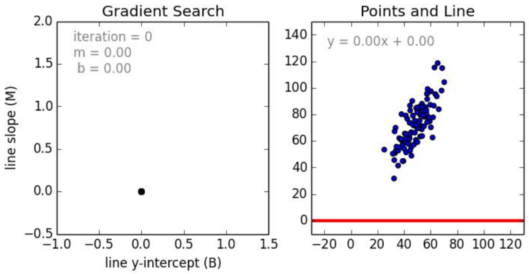
REGRESSION

- Linear Regression, GLM
- SVR, GPR
- Ensemble Methods
- Decision Trees
- Neural Networks

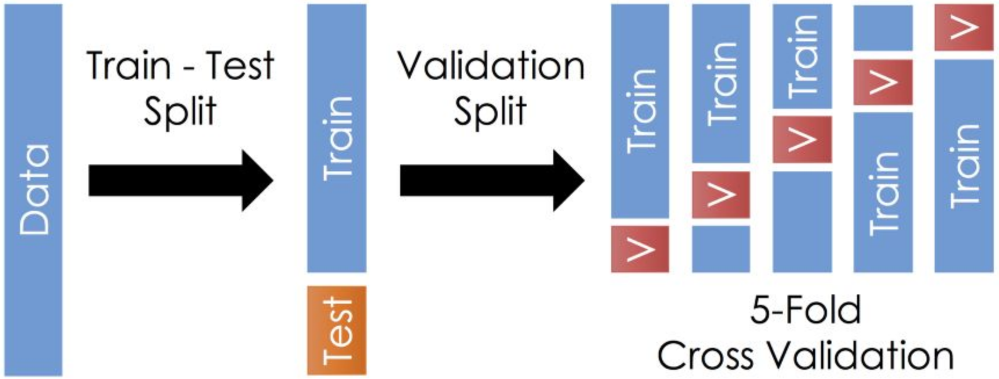
CLUSTERING

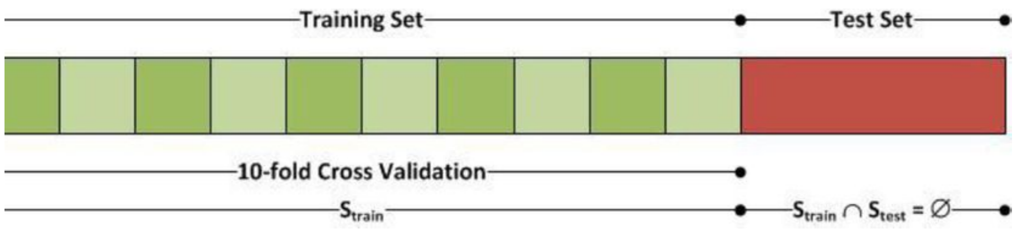
- K-Means, K-Medoids Fuzzy C-Means
- Hierarchical
- Gaussian Mixture
- Hidden Markov Model
- Neural Networks





Training and testing the model

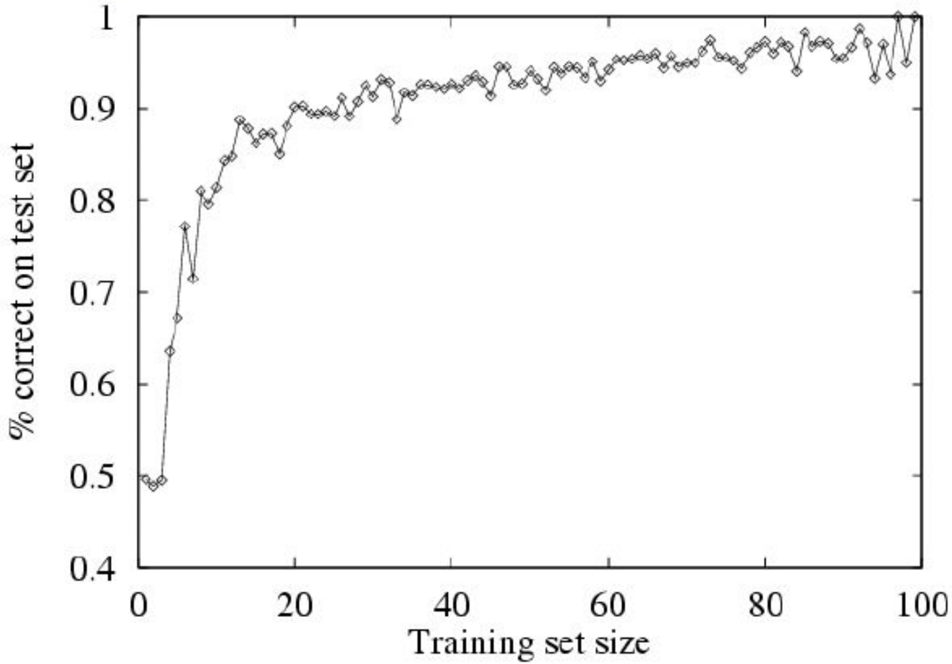




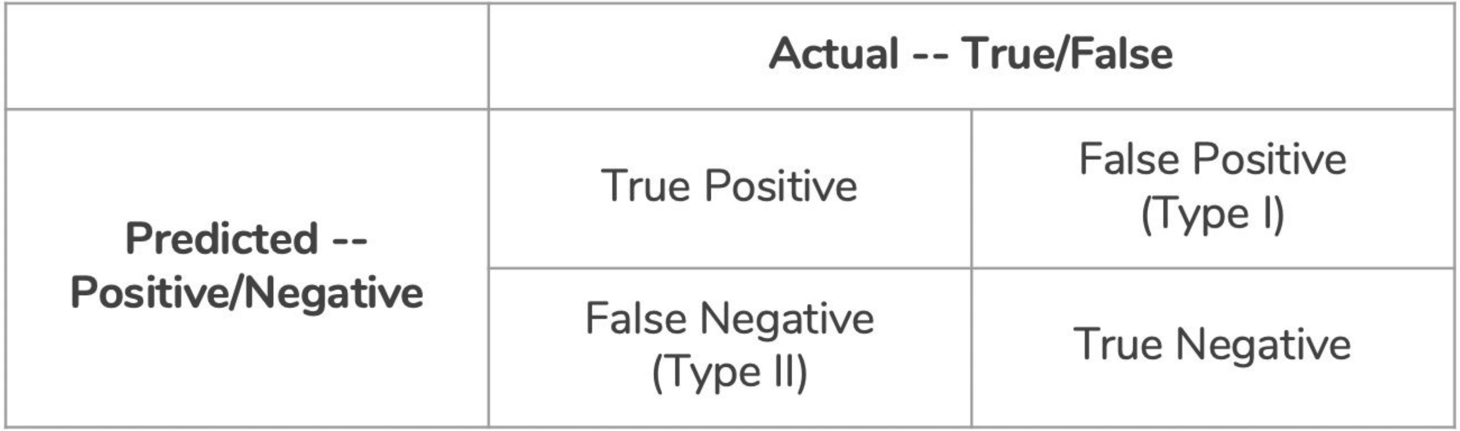
Performance Measure

- The way you want to evaluate a solution to the problem.
- It is the measurement you will make of the predictions made by a trained model on the test dataset.
- Performance measures are typically specialized to the class of problem you are working with.

Learning curve

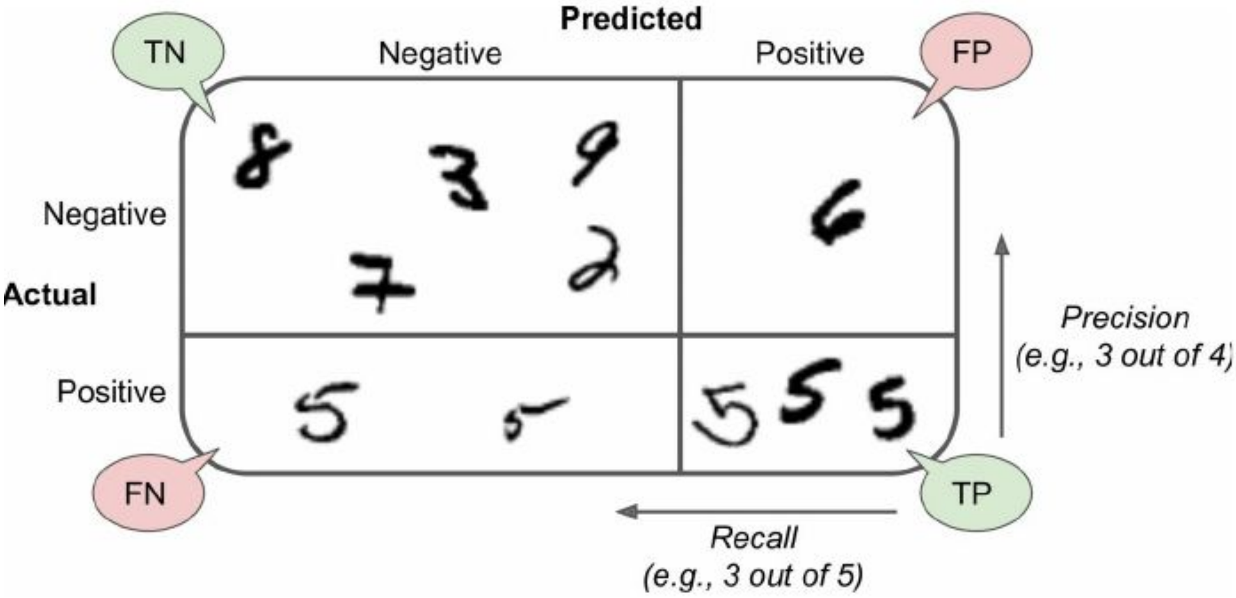


Confusion Matrix



	Actual -- True/False	
Predicted -- Positive/Negative	True Positive	False Positive (Type I)
	False Negative (Type II)	True Negative

Illustrated confusion matrix



Metrics

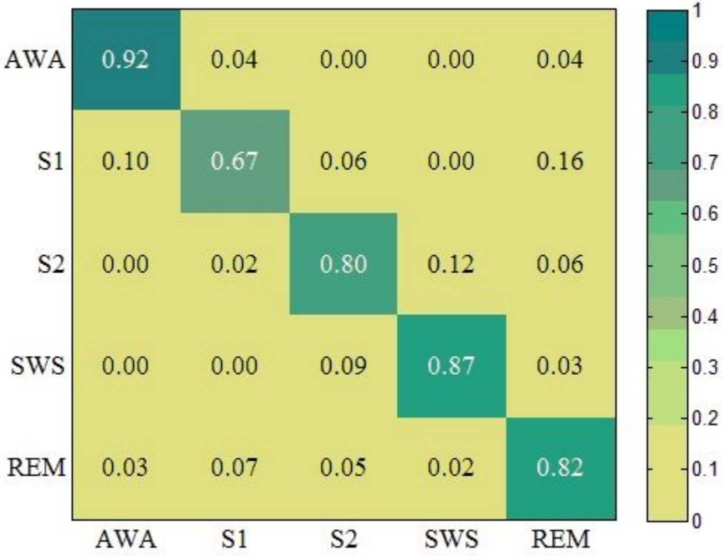
$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Multiclass confusion matrix



	AWA	S1	S2	SWS	REM
AWA	0.92	0.04	0.00	0.00	0.04
S1	0.10	0.67	0.06	0.00	0.16
S2	0.00	0.02	0.80	0.12	0.06
SWS	0.00	0.00	0.09	0.87	0.03
REM	0.03	0.07	0.05	0.02	0.82

4. Improve Results

Algorithm Tuning: where discovering the best model is treated like a search problem.

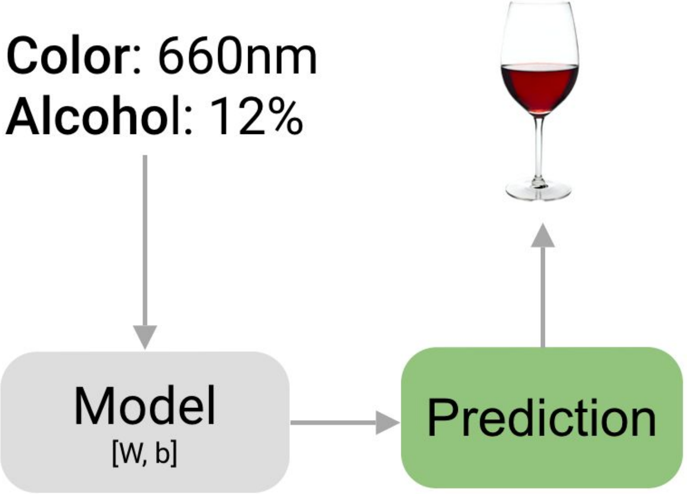
Ensemble Methods: where the predictions made by multiple models are combined.

Extreme Feature Engineering: where the attribute decomposition and aggregation seen in data preparation is pushed to the limits.

5. Using the model

Machine learning is using data to answer questions.

So **Prediction**, is the step where we get to answer some questions.



Present results

Present results

Context: why

Problem: question

Solution: answer

Findings: bulleted lists of discoveries

Limitations: where the model does not work

Conclusions: why + question + answer

Conclusions

Conclusions

- The steps may not be linear! As you clean your data, you may uncover a better question to ask. As you tune your model, you may realize you need more data, and go back to the collection step.
- The important part is to stay curious, and to keep iterating until you find a model that works the best!

References

https://www.csee.umbc.edu/courses/pub/www/courses/graduate/671/fall12/notes/14/14b.pptx.pdf

https://towardsdatascience.com/workflow-of-a-machine-learning-project-ec1dba419b94

https://towardsdatascience.com/the-7-steps-of-machine-learning-2877d7e5548e

https://machinelearningmastery.com/process-for-working-through-machine-learning-problems/

Book: Hands-On Machine Learning with Scikit-Learn and TensorFlow

Book: Feature Engineering for Machine Learning

Thanks!

Questions?