Machine Learning Introduction

CDIPS Data Science Workshop, Day 5

Jackie Brosamer

Overview

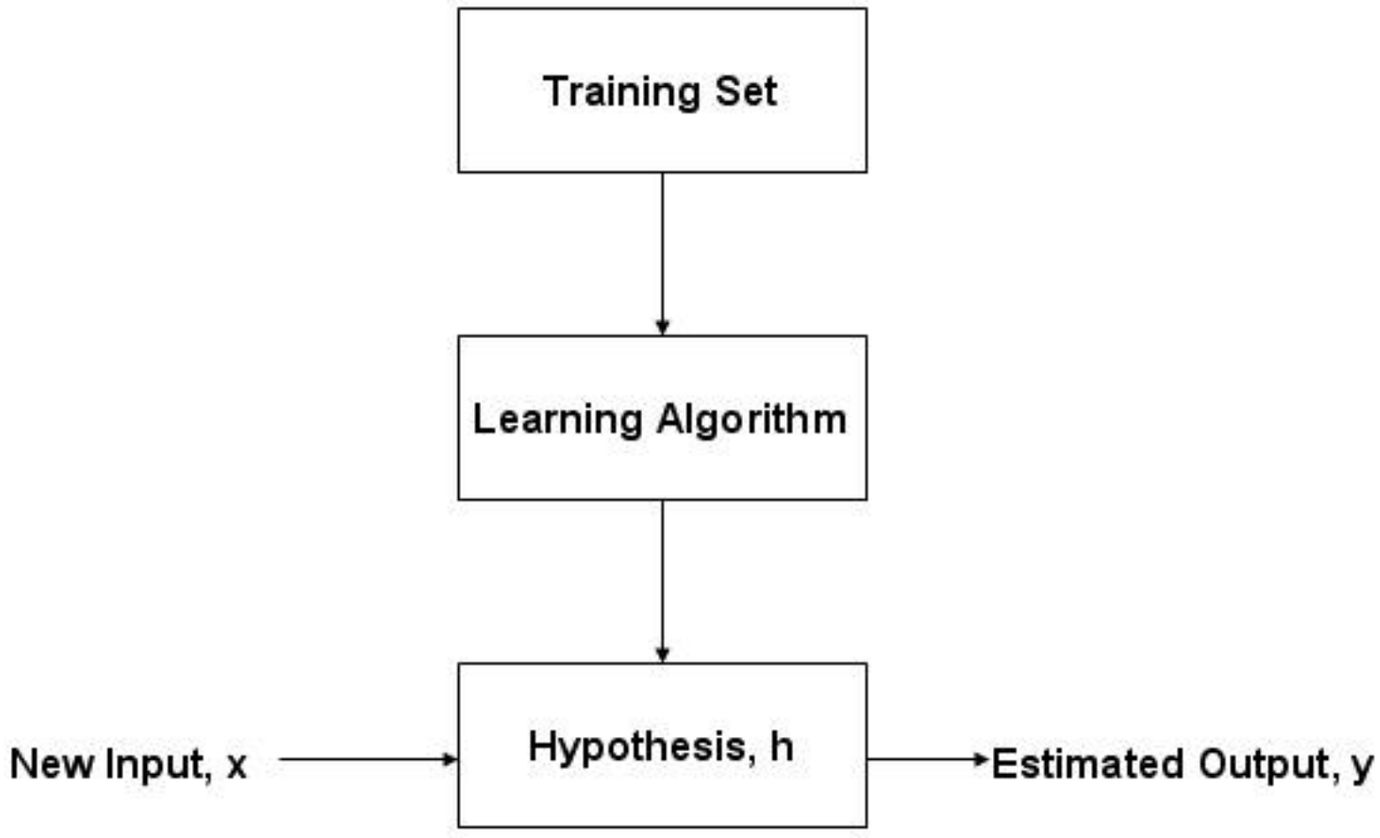
- What is machine learning?
  - Concepts
  - Types of algorithms
- Strategies and concepts
  - Preprocessing
  - Feature engineering
  - Visualization
- Workflow for improving your algorithm
- Resources

Machine learning algorithms

Machine Learning

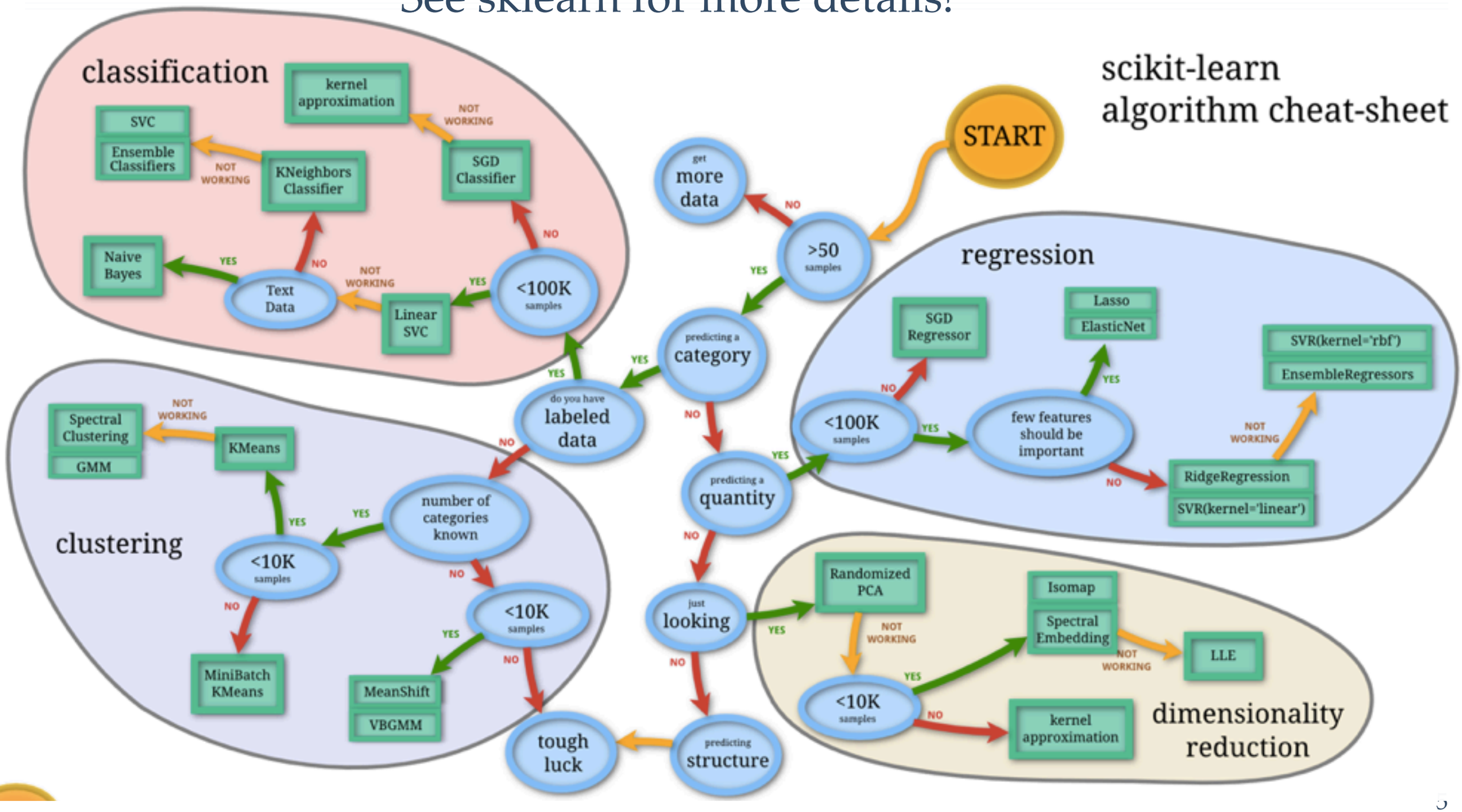
- Algorithm to generate a predictive **hypothesis** by learning from a **training set** of examples

- Used large datasets or complex problems: credit scoring, drug design, physics, pretty much everything on the internet..



Algorithm Taxonomy

See sklearn for more details!



Components of a classifier

Learning=representation+evaluation+optimization

- Representation
  - What type of inputs (*feature values)* and outputs (*classes)* are in our problem?
- Evaluation (*objective function* or *scoring function*)
  - How to distinguish a good model from a bad one?
- Optimization
  - Look for the highest scoring algorithm that satisfies our error constraints

Concepts for working with machine learning algorithms

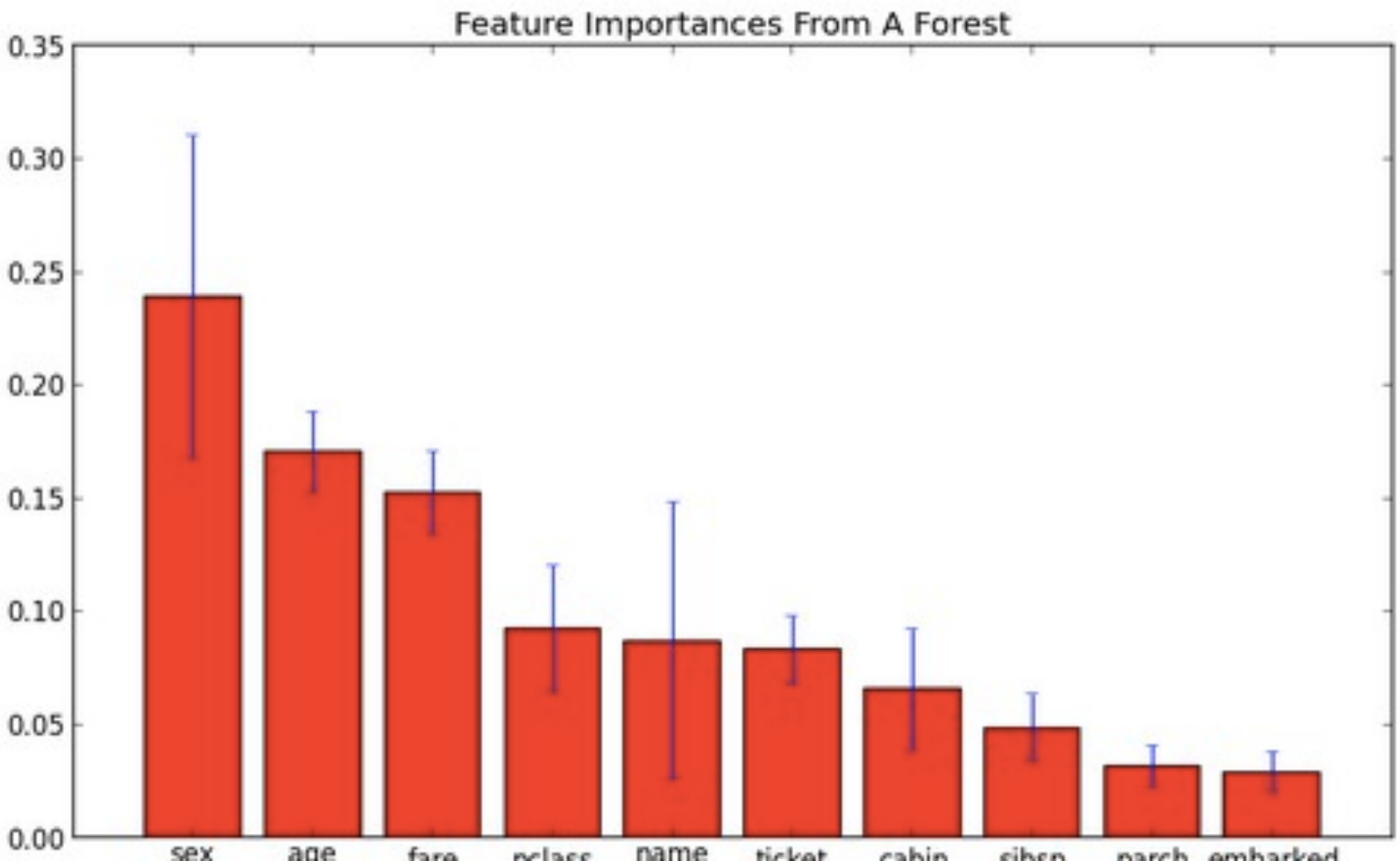
Preprocessing data

- Make features easier for your algorithm to process
- Encoding categorical features
  - example: categorical feature “sex” with values [“male”, “female”,”other”] could be represented as [0,1,2]
- Sparse arrays: efficient way to store mostly empty (or zero) arrays

See more at: http://scikit-learn.org/stable/modules/preprocessing.html

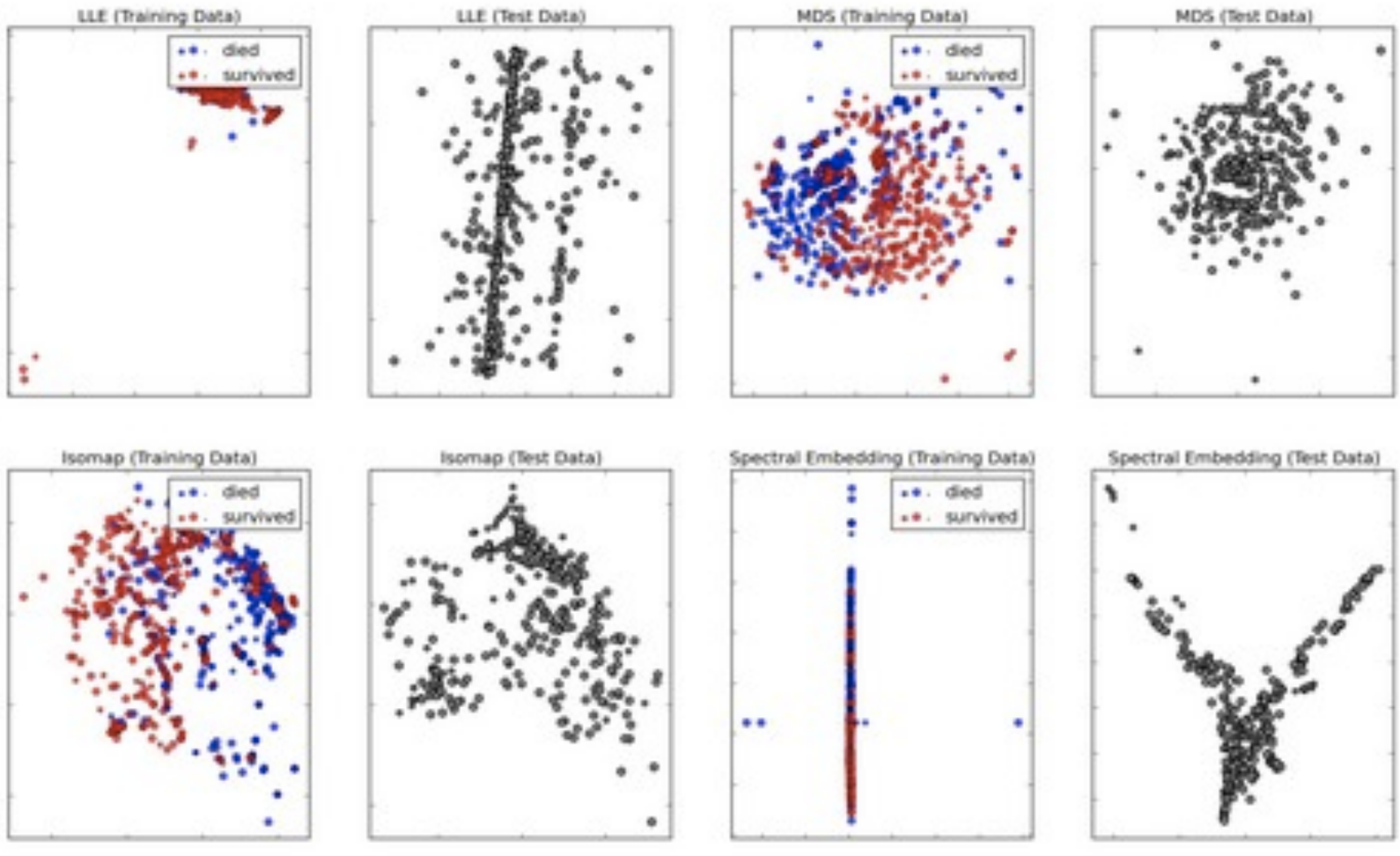
Feature Engineering

Use common sense to constrain your data before training. Are features redundant (pclass and fare)?



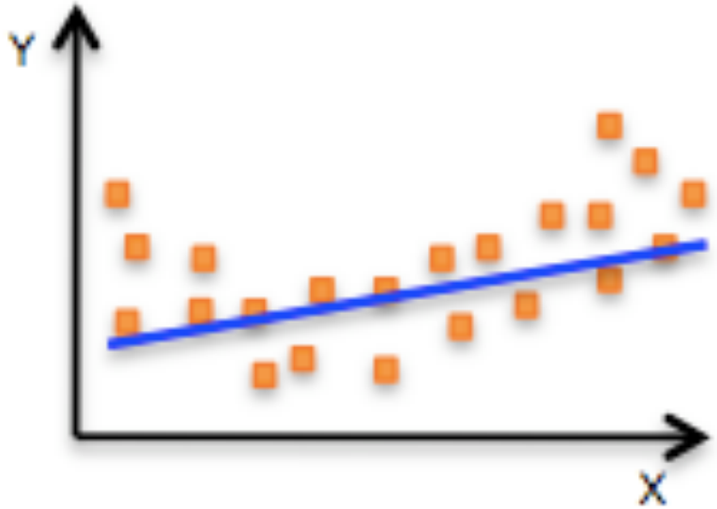
Visualization

Good to understand your results or intermediate steps. When you get stuck, try a picture!

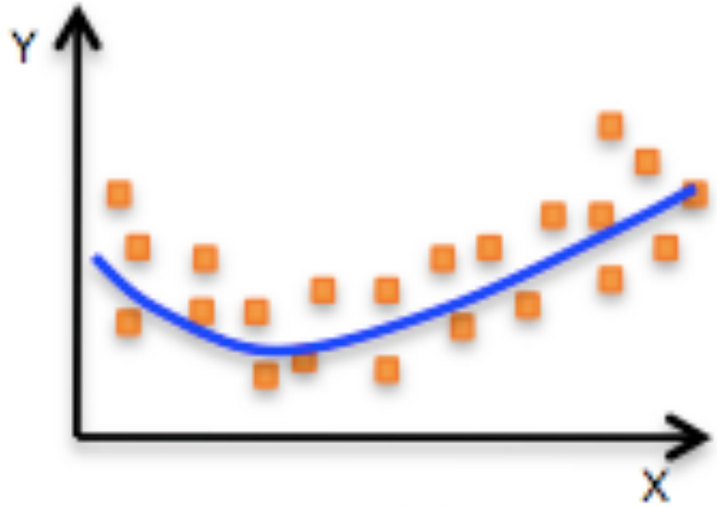


Overfitting

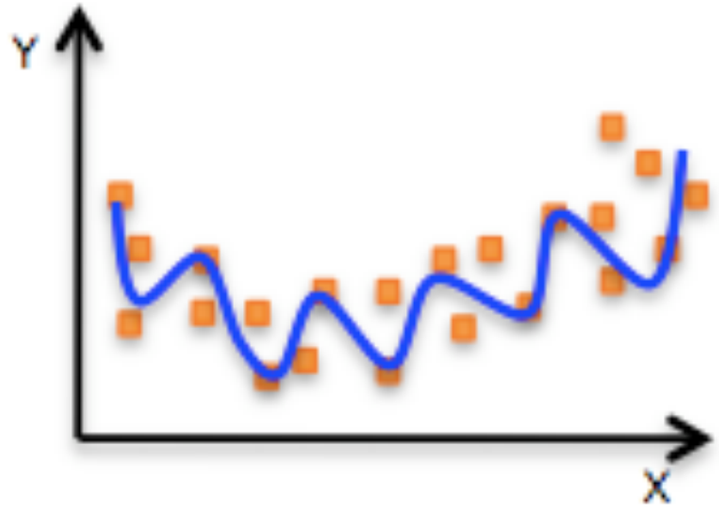
Be on guard against one of the biggest pitfalls of machine learning. Remember Occam's razor! Usually best to start simple.



Underfitting



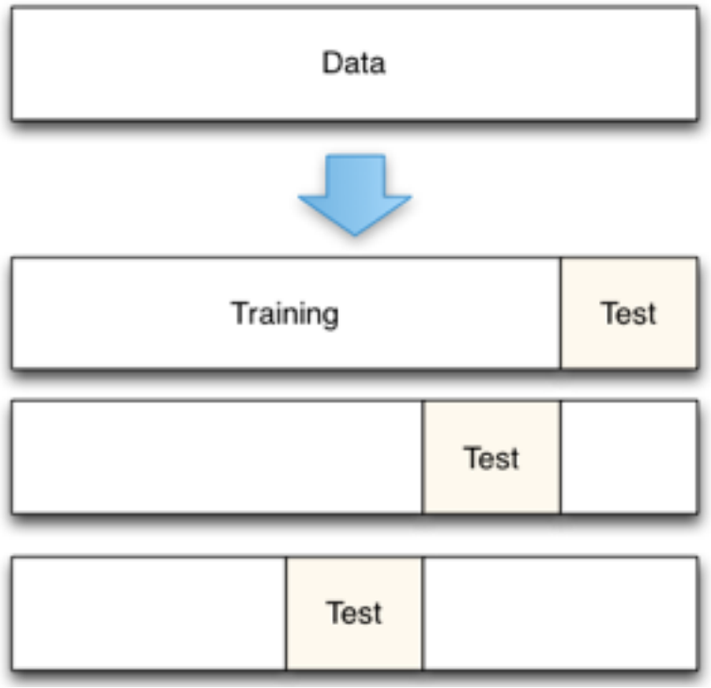
Just right!



overfitting

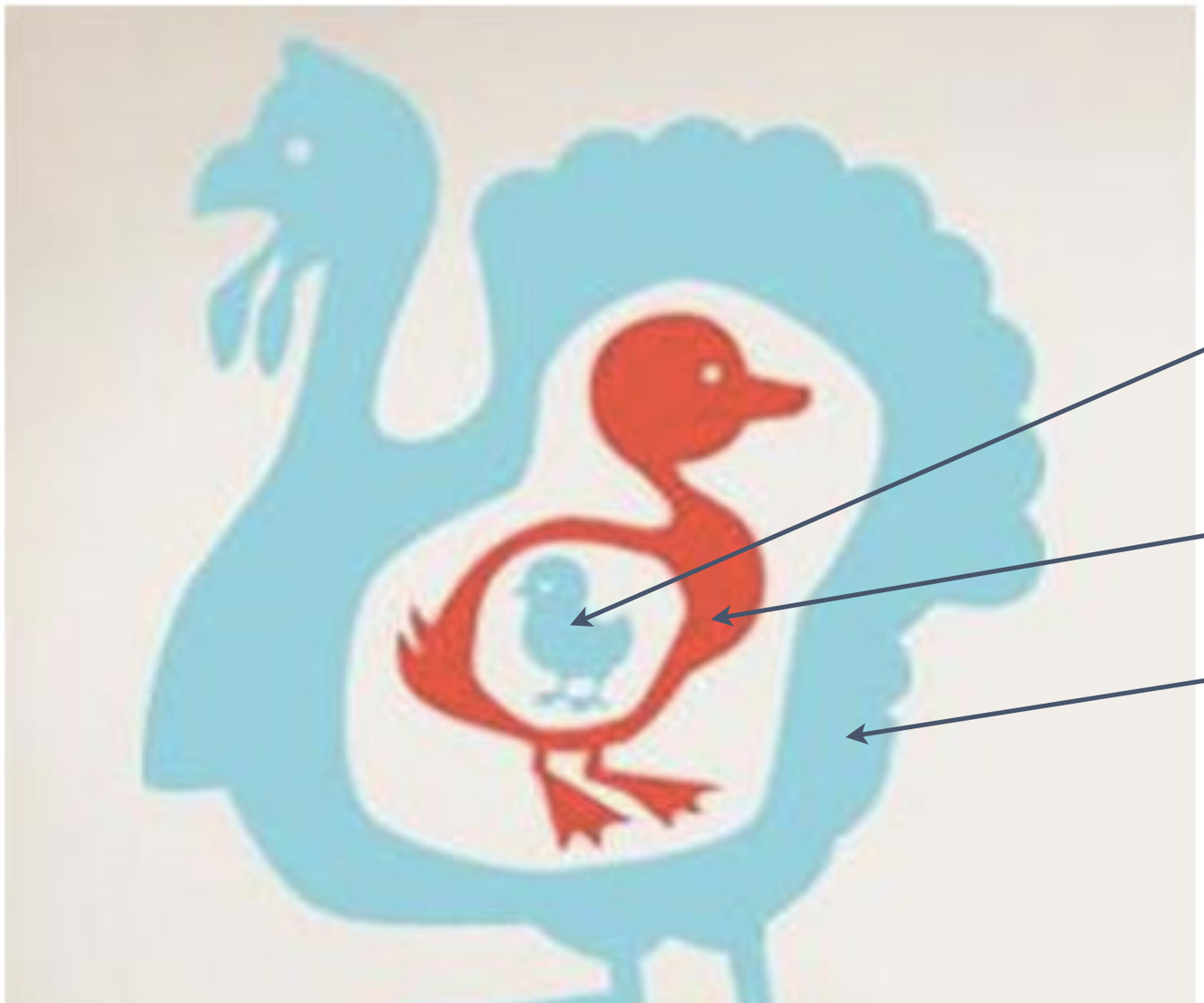
Cross-Validation

Technique for assessing how the results of a model will generalize to an independent data set by splitting the data into subsets and running through algorithm. Good guard against overfitting!



Combining algorithms

Take a combination of several algorithms to get a better result. Many already do this (e.g. random forest)



turducken > turkey
turducken > duck
turducken > chicken

So what now?

Suggestions for getting started

1. Start with a general purpose algorithm
    - Random forest, k-nearest neighbors: compare to linear regression
2. Look at output and input
    - Use visualization
    - Draw some hypotheses with your
3. Tweak algorithm, drawing on assumptions from 2.
    - Revisit your assumptions about how you cleaned and filled the data.
    - Be creative with additional feature engineering, so that your chosen model has more columns to train from.
    - Use the sklearn documentation to experiment with different parameters for your algorithm.
    - Consider a different model approach. For example, a logistic regression model is often used to predict binary outcomes like 0/1.
    - New algorithm?
4. Back to 2. and repeat! Keep experimenting.
5. *Don't forget to utilize your data science mentor!*

Resources

- **Sklearn documentation and tutorials: http://scikit-learn.org**
- **Kaggle forum: https://www.kaggle.com/forums**
- Preparing data: http://machinelearningmastery.com/how-to-prepare-data-for-machine-learning/
- Aggregation of data science tutorials: http://datacommunitydc.org/blog/2013/07/python-for-data-analysis-the-landscape-of-tutorials/
- Machine learning "folk wisdom": http://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf