Building first Scikit Learn Solution.

- By Janani Ravi (Pluralsight)

Machine Learning:

- An algorithm that can learn from data.
- The algorithm should be able to work with huge data.
- It finds patterns in data while learning and apply that on new data to predict the output for unknown data.

Types of ML

- Main categories of ML problem
  - ◊ Classification - Discrete value/output
  - ◊ Regression - continuous values
  - ◊ Clustering - based on similarities.
  - ◊ Dimensionality Reduction - reduce the features by identifying the significant attributes.

- Data is used to train the ML Model parameters.
- Model parameters pick the significant features from the data to predict the output.
- Model parameters refer to variables in the ML algorithm that can be trained using data.
- Two phases of building ML model:
  1. Training Phase:
     a. Feed in a large corpus of data classified correctly.
     b. Feedback – by using loss function, we need to reduce the loss function to improve the model.
     c. Learnable model parameters are trained in this phase.
  2. Prediction phase
     a. The trained model is used for prediction.
     b. To classify the new instance.

Traditional and Representational ML:

- Traditional Models: They have fundamental algo structure to solve problem.
  - Different models have different structures.
  - The algo is fed data which train the algo parameters called model parameters.
  - These ML based systems rely on expert to decide what features to pay attention to and how.
    - Regression models: Linear, Lasso, Ridge, SVR
    - Classification models: Naïve bayes, SVMs, Decision Trees, Random Forest
    - Dimensionality Reduction: Manifold learning, factor analysis
    - Clustering: K-means, DBSCAN, Spectral
- Representational Models: Also used to solve classification, regression, clustering, dimensionality reduction problems.
  - Used to solve complex use cases.
  - Learn significant features from underlying data, no expert decision required.
  - DL models such as NN
  - These ML based systems figure out by themselves what features to pay attention to and how.

What is NN?

DL: Algorithm that learn what features matter,

NN: The most common class of DL algorithms :- It is common class od DL algorithms : made up of neuron.

Neuron: simple building blocks that learn , mathematical functions that learn from data fed.

Traditional vs. Deep Learning Models

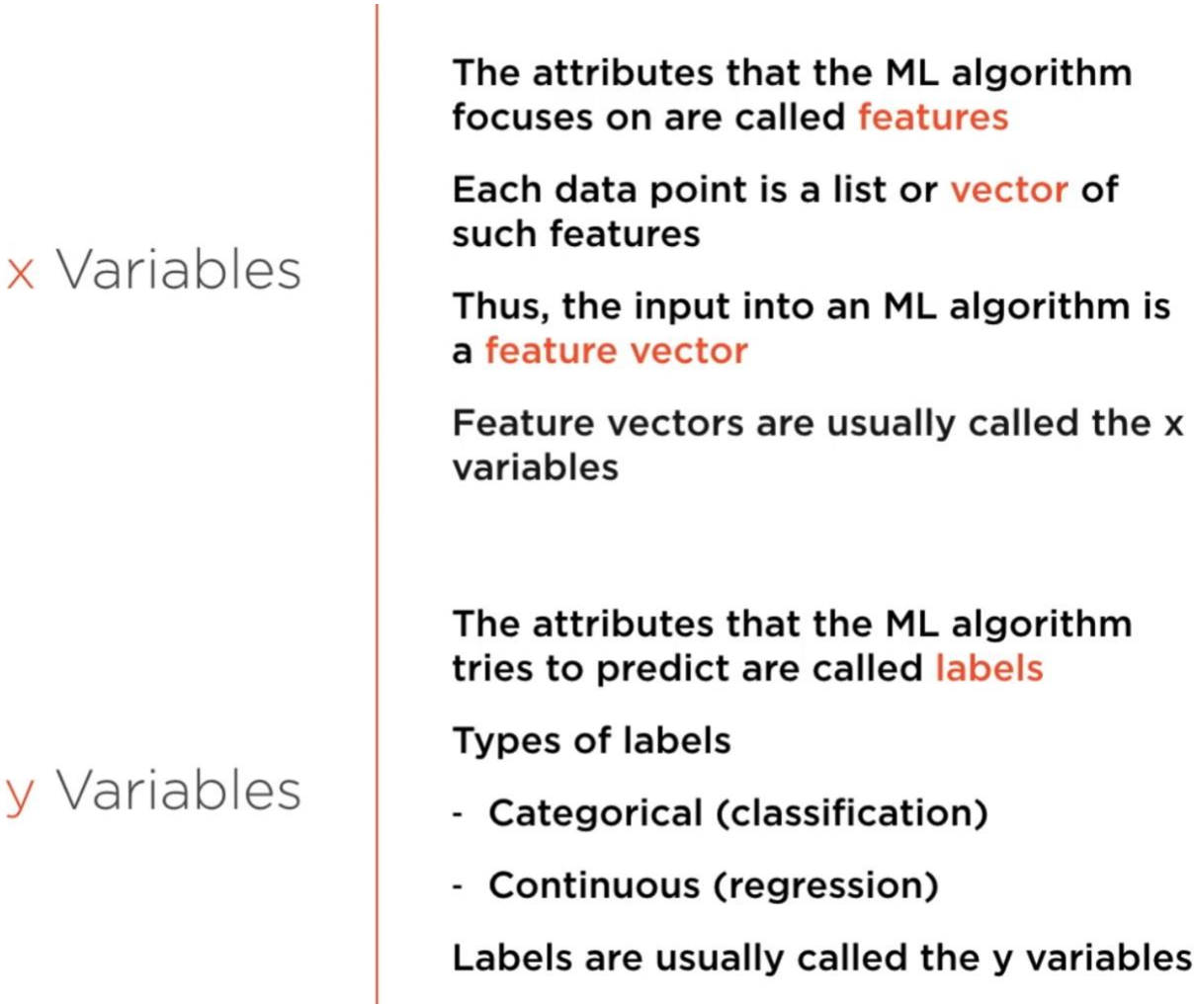
Traditional ML Models	Deep Learning ML Models
Features used in models explicitly chosen by domain experts	Features used in models implicitly chosen by model itself
Structured data such as numbers and probabilities	Unstructured data such as images and movies
Classification, regression, clustering, and dimensionality reduction	Classification, regression, clustering, and dimensionality reduction
Wide range of problem-specific solution techniques	Neural networks by far the most common solution technique
Each solution technique adopts characteristic approach	All solution techniques rely on neurons and interconnections between them
User has more insight into mechanics and internals of models	Black-box models that are hard to question or reverse-engineer
scikit-learn	TensorFlow, Keras, PyTorch

Sci-Kit Learn

- It is easy to use, very comprehensive and efficient python library for Traditional ML models.
- Easy to use: Estimator API for consistent interface.
  - Estimators of all kinds f models.
  - Create model object -> fit to training data -> predict for new data.
  - Create pipelines for complex operations.
- Comprehensive : All common families of models are supported. (reg, classification, cluster, DR)
  - Libraries available for Data Wrangling: that includes data preprocessing, cleaning, feature selection and extraction.
  - Hyperparameter tuning.
  - Libraries for model validation and evaluation (cross validation)
  - Data generation for artificial data generation: Swiss rolls, S-curve.
- Efficient: Highly optimized implementation
  - Built on SciPy
  - Popular libraries: NumPy, SciPy, Matplotlib, Sympy, Pandas and they all interact with Sci-Kit learn.

Link to scikit Learn: https://scikit-learn.org/stable/

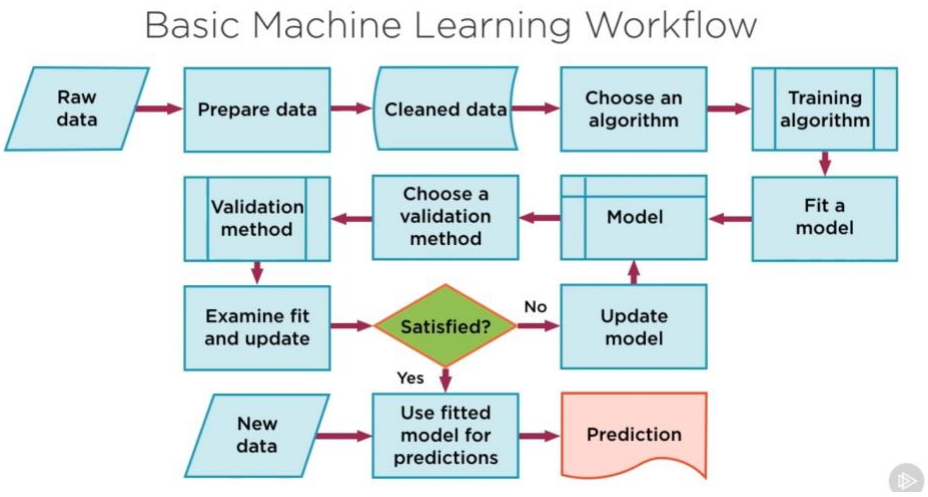
Supervised Learning: Label Associated with the training data. Y = f(x)
Regression and classification



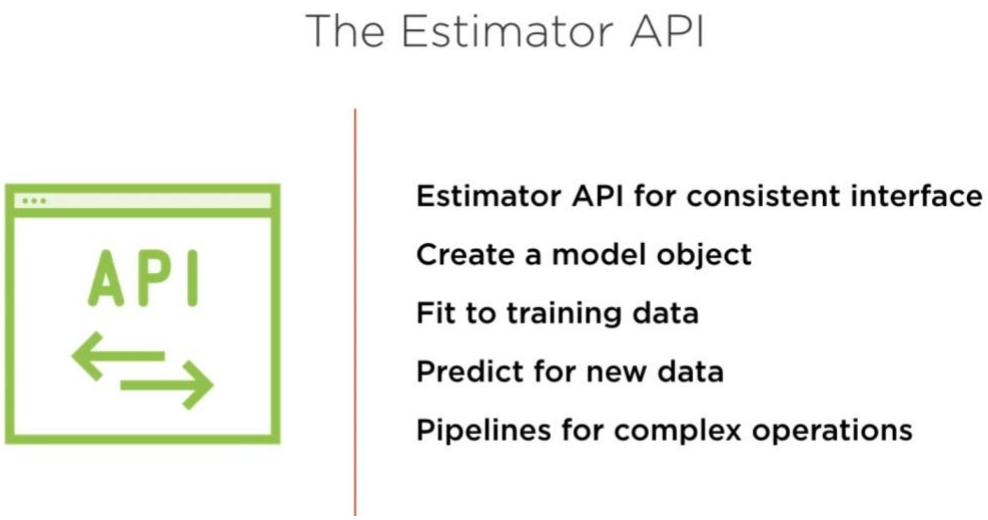
Unsupervised Learning: The data is not labeled. Model learns from data.

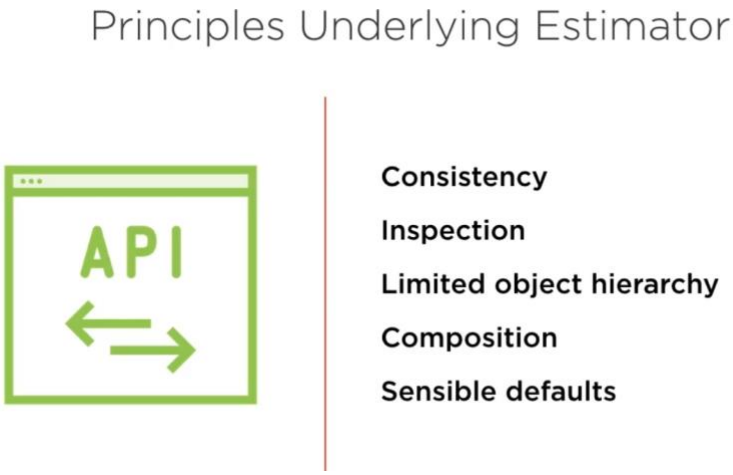
- They self-discover the patterns and Structure in the data.
  Clustering and dimensionality reduction(PCA).

1. Scikit-Learn in the typical ML workflow.



2. Estimators and pipelines



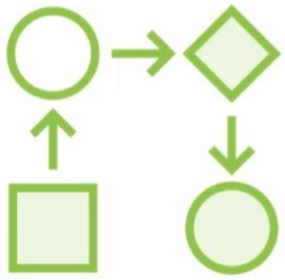


- Raw data – Pandas, NumPy.
- Prepare data – standardization, normalization, scaling.
- Cleaned data – missing values, outliers.
- Choose an algorithm – suite of algorithms for regression, classification, clustering, DR.
- Fit a model – find best model parameters.
- Model – fit () or fit_transform().
- Choose a validation method –
- Validation method – cross validation, k-fold.
- Examine and update model - Metrics of evaluation.
- Predict – invoke predict().

scikit-learn Pipeline

Estimator object that sequentially applies several transforms. Pipeline can be evaluated and tuned as a whole.

Pipelines

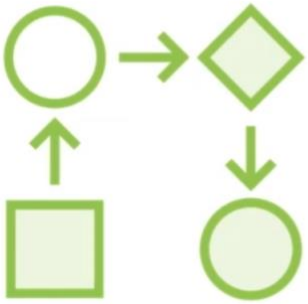


Pipeline objects are estimators too

Apply transforms sequentially

Return fitted estimator

Final output only implements fit



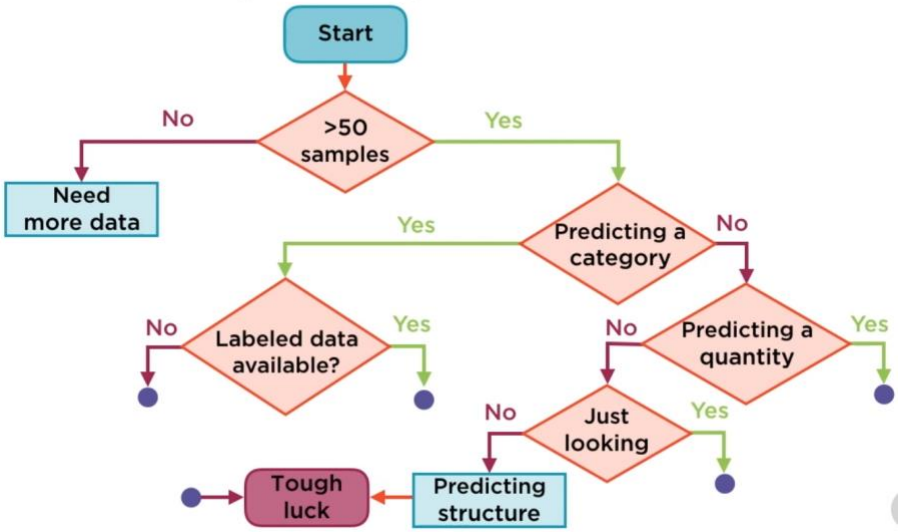
Intermediate transforms can be cached

Easily tune pipeline as a whole

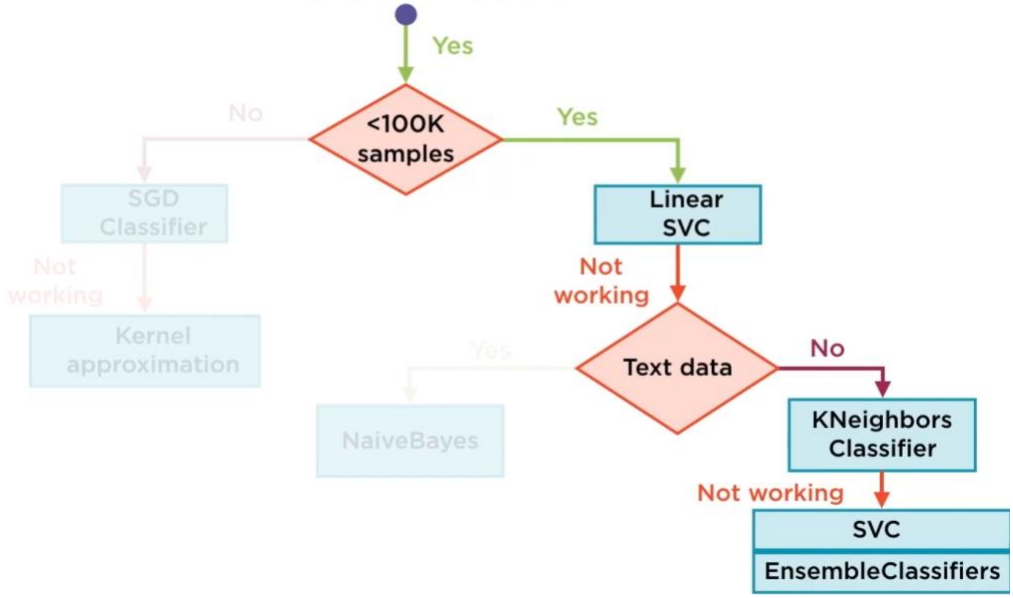
Cross-validation

Switch in or switch out individual steps

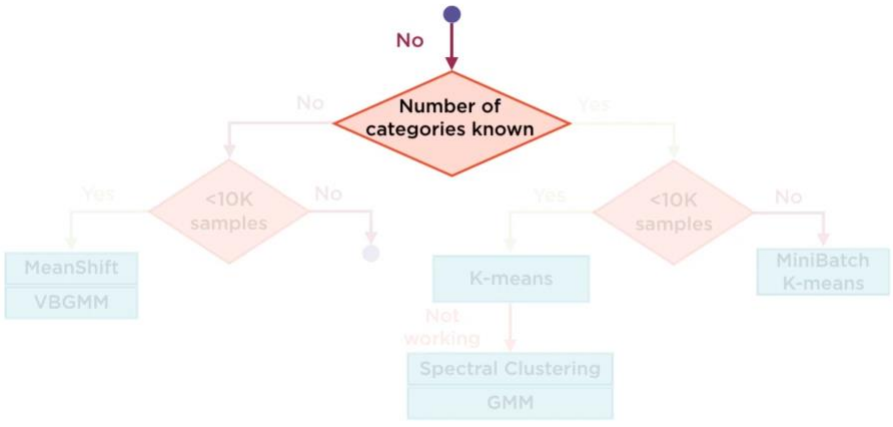
Choosing the Right Estimator

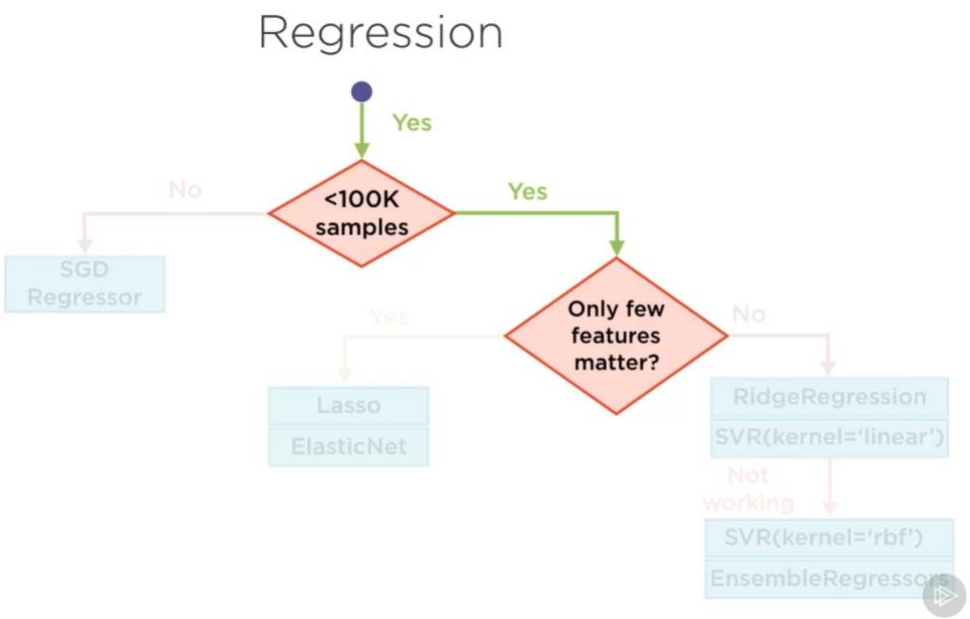


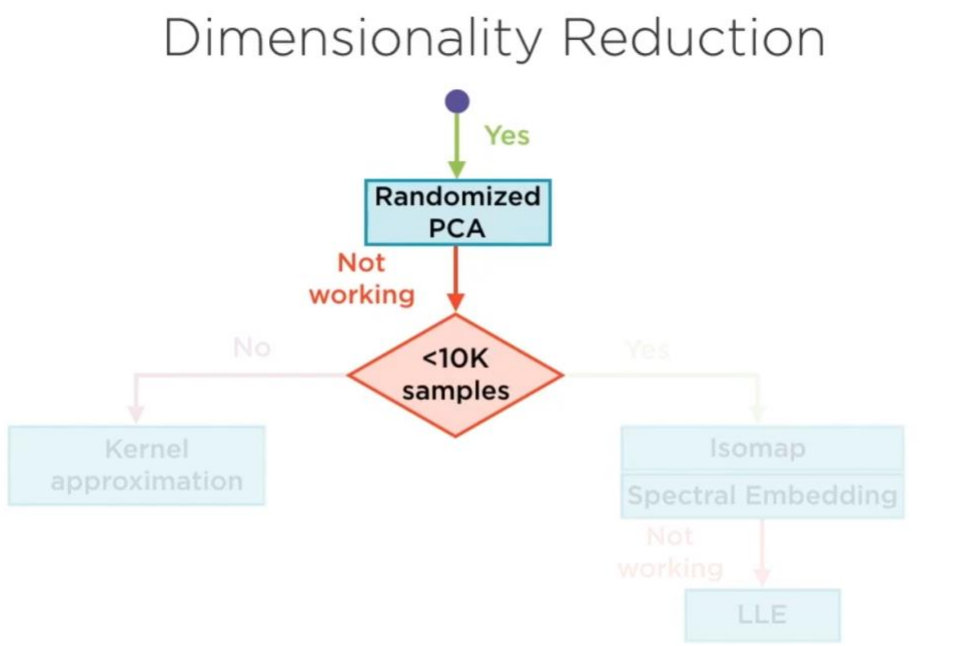
Classification



Clustering







3. Model evaluation and transformation
4. Loading, cleaning, transforming and visualizing datasets.

Follow the .ipynb files for examples of using scikit learn.