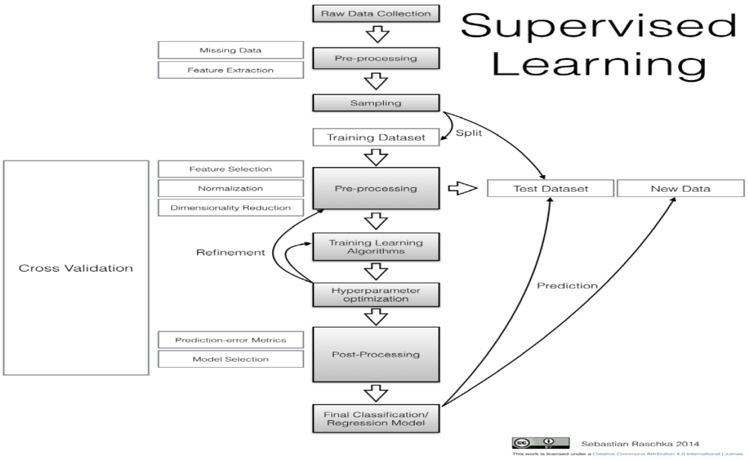
Machine Learning process

- Basic Process

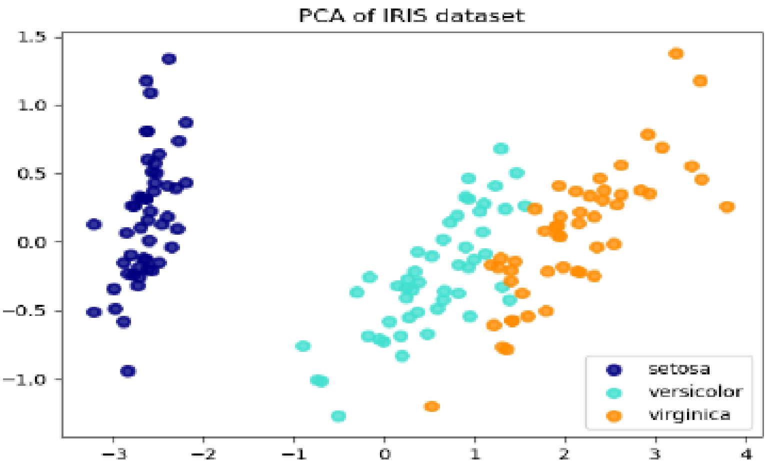
  - Data acquisition and sensing:
    - Measurements of physical variables.
    - Important issues: bandwidth, resolution , etc.
  - Pre-processing:
    - Removal of noise in data.
    - Isolation of patterns of interest from the background.
  - Feature extraction:
    - Finding a new representation in terms of features.
  - Classification/Clustering
    - Using features to learn models for different tasks.
  - Post-processing
    - Evaluation of confidence in decisions
- Preliminary preparation
  - Clarify the problem
    - What is input?
    - What is output?
  - Data selection
    - Representation of the data
    - The time range of the data
    - Data business scope
- Feature engineering
  - Exploratory Data Analysis (EDA)
    - Clean the data, describe the data (descriptive statistics, graphs), view the distribution of the data, compare

        - Compare the relationship between the data, cultivate the intuition of the data, and summarize the data.
    - Data overview
        - Check whether the data types, dimensions are consistent, and missing values
        - Find data outliers-box plot
    - Data distribution
        - For regression problems, the target variable should conform to the normal distribution as much as possible
            - Data visualization: histogram
            - Take the logarithm and fit the unbounded Johnson distribution
    - Data type view
        - Correlation analysis
        - Number of categories
- Data preprocessing
    - Outlier handling
        - Deal with man-made outliers, determine the outliers through business or technical means (such as the 3σ criterion), and then
        - (Regular expression matching) and other methods to filter abnormal information, and delete or replace values based on business conditions
    - Missing value processing
        - High missing rate => directly delete the corresponding feature variable, add bool type, missing situation, missing record
        - Is 1, non-missing is recorded as 0
        - The missing rate is low => some missing value filling methods can be used in combination with the business, such as pandas's fillna method
        - Method, training regression model to predict and fill in missing values;
        - No processing: some models such as random forest, xgboost, lightgbm can handle missing data
        - In addition, there is no need to deal with missing data.
    - Data discretization
        - Discretization is the segmentation of continuous data into segments of discretization. The original segmentation
        - There are methods such as equal bandwidth and equal frequency. Discretization can generally increase noise immunity and make features more professional
        - Service interpretability, reduce the time and space overhead of the algorithm
    - Data standardization(normalization)
        - The dimension of each feature variable of the data is very different, and you can use data standardization to eliminate different component dimensions

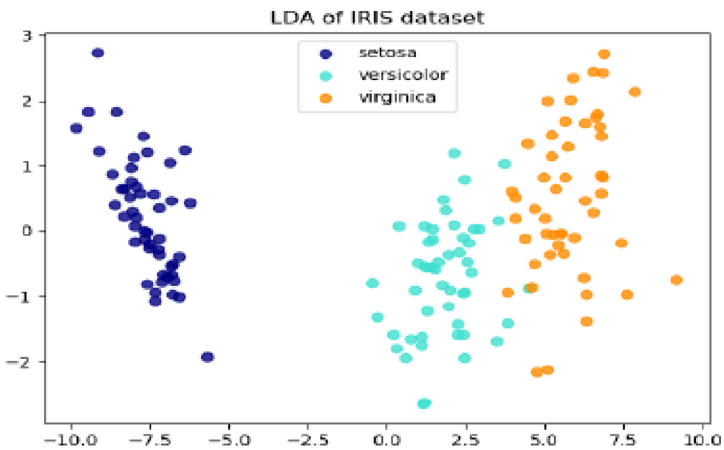
    - The impact of differences, accelerate the efficiency of model convergence
      - min-max: The value range can be scaled to (0, 1) without changing the data distribution. max is the most sample
      - Large value, min is the minimum value of the sample.
      - z-score: The value range can be scaled to near 0, and the processed data meets the standard normal score
      - cloth. Is the mean and σ is the standard deviation.
  - Data reduction
    - Dimensionality reduction -feature selection
    - Numerosity reduction –select/ sample records
- Feature extraction
  - Feature representation: For example, the picture is converted to a three-dimensional matrix (rgb) or a one-dimensional matrix (grayscale)
  - Feature derivation: Basic features have limited expression of sample information, and feature derivation can increase the number of features
  - Non-linear expression ability improves model effect. Feature derivation is to make a certain treatment of the meaning of existing basic features.
  - Management (aggregation/conversion, etc.), common methods of manual design, automated feature derivation
    - Group aggregation: Calculate the average, count, maximum value, etc. after the fields are aggregated => For example, through 12 months of work
    - Salary can be processed out: average monthly salary, maximum salary
    - Conversion: Add, subtract, multiply and divide between fields. For example, the 12-month salary can be processed: current month's work
    - The ratio and difference of capital income and expenditure, etc.
  - Feature selection: The goal of feature selection is to find the optimal feature subset, by filtering out salient features and abandoning redundancy
    - Curse of dimensionality
    - Retain only "useful" (discriminatory) information and avoid overfitting.
    - Reasons to reduce the number of features:
      - Computational complexity
      - Generalization properties
  - Additional features can reduce the risk of model overfitting and improve operating efficiency.
    - Filtering method: Calculate the lack of features, divergence, relevance, amount of information, stability, etc.
    - The indicators evaluate and select each feature, such as missing rate, single value rate, variance verification,

- Pearson correlation coefficient, chi2 chi-square test, IV value, information gain, PSI and other methods.
- Packing method: Iteratively train the model by selecting some features each time, and select according to the model prediction effect score
- Feature removal, such as sklearn's RFE recursive feature elimination.
- Embedding method: directly use some model training to the feature importance, and perform the feature at the same time as the model training
- choose. The weight coefficient of each feature is obtained through the model, and the characteristic is selected according to the weight coefficient from large to small.
- Levy. Commonly used are logistic regression based on L1 regular term, XGBOOST feature importance selection feature.

- Feature Dimensionality Reduction:
  - PCA: principal component analysis

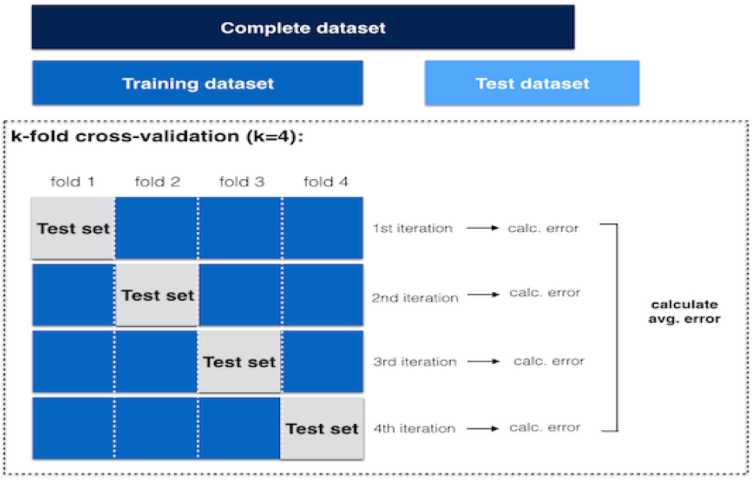


  - LDA(supervised):Linear Discriminant Analysis

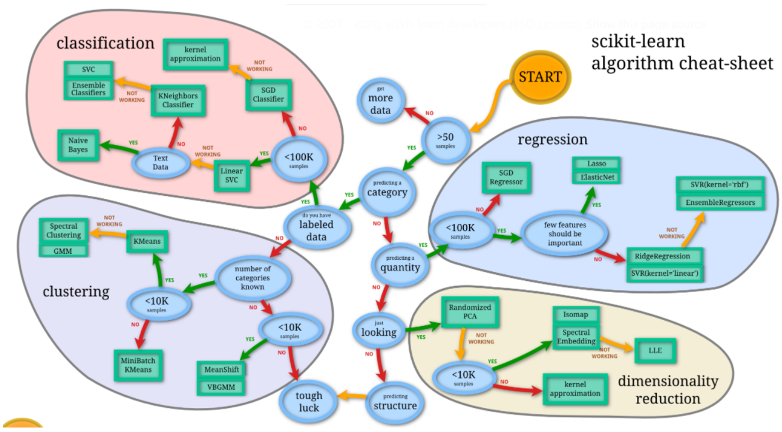


    - identify attributes that account for the most variance between classes.

- Model training
  - Data set division
    - HoldOut verification method, leave one method, k-fold cross verification



- Training set: used to run learning algorithms and train models.
- The valid set is used to adjust hyperparameters, select features, etc., to select suitable models
- The test set is only used to evaluate the performance of the selected model, but the learning algorithm or parameters will not be changed accordingly.
- Data enhancement (only for training set)
    - Data enhancement
    - Data oversampling/downsampling => improve data imbalance
- Select model



- Training process
- Training set
    - Use learning rate finder to select learning rate
    - Adam optimization
    - Cosine learning rate decay
    - Learning rate restart
    - If you do transfer learning, try a differentiable learning rate
    - Number of neurons in the hidden layer
    - minibatch size
    - Number of hidden layers

  - Improvement
    - Dropout
    - L2 regularization
    - Input feature normalization
    - Batch normalization
    - Data augmentation
    - Supplement data for the training set
    - Gradient disappears or explodes
    - He initialization
    - Use LSTM neurons
    - Gradient clipping
    - Adjust the neural network architecture
    - Visualize the training process
    - Training process log record
    - Hyperparameter optimization
      - Hyperparameter are parameters that are not directly learnt within estimators.
      - Methods used to find out Hyperparameters:
        - Manual Search
        - Grid Search
        - Random Search
        - Bayesian Optimization
        - Evolutionary Optimization
  - Learning from Imbalaned data
    - data augmentation
    - custom loss funciton
- Model evaluation
  - Evaluation index
    - Classification model
      - Commonly used evaluation standards include precision rate P, recall rate R, and the average F1-score of the two, etc.,
      - Calculate the corresponding number of matrix statistics:

混淆矩阵		预测类别	
		Positive	Negative
实际类别	Positive	TP	FN
	Negtive	FP	TN

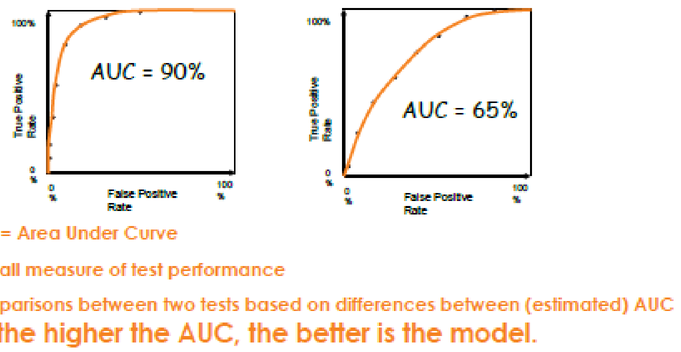
Actual class\Predicted class	Predicted C_1	Predicted $\neg C_1$
Actual C_1	True Positives (TP)	False Negatives (FN) Type-II Error
Actual$\neg$ C_1	False Positives (FP) Type-I Error	True Negatives (TN)

Accuracy = (TP + TN)/All
Sensitivity = True Positive Rate = Recall= TP/(TP+FN)
Specificity = True Negative Rate = TN/(FP+TN)
Precision = TP/(TP+FP)
F1 score = 2*Precision* Recall/(Precision + Recall)

- The precision refers to the number of positive samples (TP) that are correctly classified by the classifier, which accounts for all the positive predictions of the classifier.
- The ratio of the number of samples (TP+FP); recall rate refers to the number of positive samples that are correctly classified by the classifier
- (TP) the proportion of all positive samples (TP+FN). F1-score is the precision rate P, recall rate
- Harmonic average of R:

$$\text{F1\_score} = \frac{2PR}{P + R}$$

- ROC(Receiver Operating Characteristic)Curve



- Use cross entropy as loss function

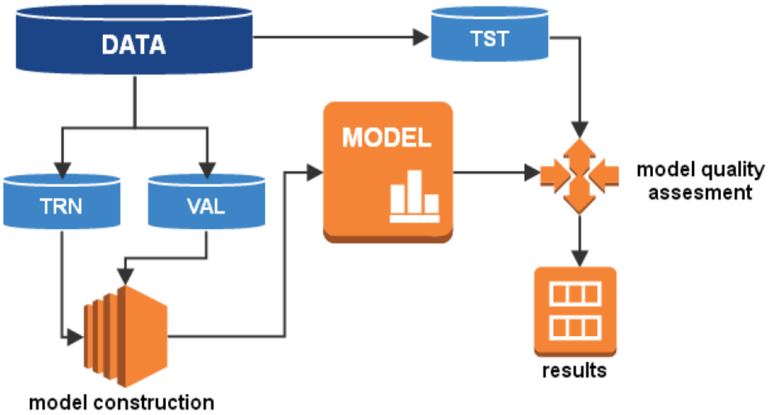
- Evaluate the regression model
  - Commonly used evaluation indicators include MSE mean square error and so on. The feedback is the fit between the predicted value and the actual value condition

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y^{(i)} - f(x^{(i)}; w))^2$$

- Evaluate the clustering model

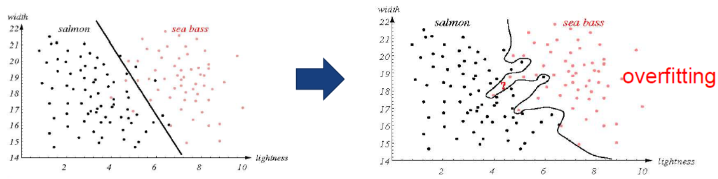
        - The clustering results are compared with the results of a certain "reference model", called "external indicators" (external indicators).Such as Rand index, FM index, etc.
        - Directly inspect the clustering results without using any reference model, which is called "internal indicators" (internal indicators).Such as compactness, separation, etc.

- Model evaluation and optimization

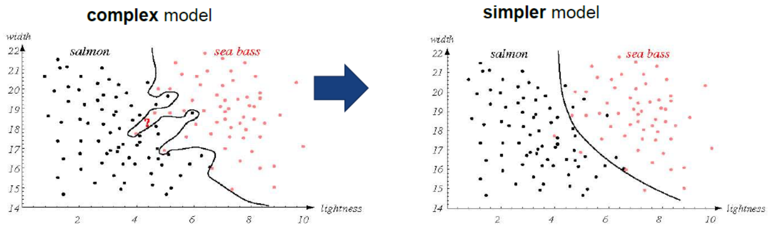
    - The error in the training data is called training error, and the error in the test data is called
    - It is test error or generalization error.
    - Overfitting

      Complex models are tuned to the particular training samples, rather than on the characteristics of the true model.

        - Reasons: too little training data, remember the noise of the sample, and the model complexity is too high
        - Solution: data cleaning, increase the number, regularization, reduce model complexity, early
        - stopping/dropout, integrated learning, pruning
    - Underfitting
        - Reason: The model is too simple and lacks the characteristics of strong predictive ability
        - Solution: Choose a model with stronger model capacity, and add effective features to feature engineering
    - Generalization
        - Generalization is defined as the ability of a classifier to produce correct results on novel patterns.

- How can we improve generalization performance ?

  - More training examples (i.e., better model estimates).
  - Simpler models usually yield better performance.

- Model deployment
  - Explain model predictions
  - Engineering is result-oriented. The effect of online model operation directly determines the success or failure of the model, not just its accuracy.
  - Accuracy, error, etc., as well as its operating speed (time complexity), resource consumption (space complexity)
  - Comprehensive consideration of stability) and stability.