COMP 4432 Machine Learning

Lesson 1: Introduction

Yuanyuan Li
yuanyuan.li456@du.edu
09/14/2022

Agenda

- Syllabus
- My biography
- Introductions
- Introduction to machine learning

Syllabus

Course Overview & Objective

- This course explores machine learning techniques and theory. The course covers how to use popular machine learning libraries to develop, train, evaluate, and deploy predictive models on prepared data. Both design principles (machine learning types and tasks) and technical tools/languages will be covered.

- Students will understand and be able to apply machine learning techniques to train, evaluate, and tune models to increase performance, as well as prepare data for analysis via feature engineering. Students will write Python scripts utilizing the following packages: numpy, pandas, matplotlib, Seaborn, scikit-learn, TensorFlow 2, and others as necessary.

Textbooks and Materials

- Required textbooks for COMP 4432 Machine Learning:
  - Géron, A. (2019). *Hands-on machine learning with scikit-learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

- Assignment datasets are located here: https://github.com/arjayit/cs4432_data. Instructors will provide additional guidance when they assign each deliverable. Please reach out to your instructor if you have any questions or trouble accessing the data sets."

Grading

Assignment/Assessment	Points	Weight on Final Grade
Assignment 1	100	20%
Assignment 2	100	20%
Assignment 3	100	20%
Assignment 4	100	20%
Assignment 5	100	20%

Late assignments policy (10% off per week)

Grading Scale

A	= 93–100
A−	= 90–92.99
B+	= 86–89.99
B	= 83–85.99
B−	= 80–82.99
C+	= 76–79.99
C	= 73–75.99
C−	= 70–72.99
D+	= 66–69.99
D	= 63–65.99
D−	= 60–62.99
F	= <60

Assignment and Assessment Information

- **Assignment 1 (due by midnight MST the day prior to Live Session 2)**
  - **Assignment 1, Part 1**: Data Loading and Preparation. Load the diabetes dataset into two numpy arrays: one for the feature set and one for the target. Pick a single feature to try to predict the target (disease progression). Document the reason you chose the feature you did. Break your single feature and target sets into training and test sets with the last 20 rows being in the test set.
  - **Assignment 1, Part 2**: Model Training. Instantiate a linear regression model, and train it with your single feature and target sets.
  - **Assignment 1, Part 3**: Prediction and Measurement. List the first 10 predictions on your single feature training set. Print out the feature coefficient and the root mean squared error of your model.
  - **Assignment 1, Part 4**: Visualization. Print out a scatter plot with the feature you chose on the x-axis, and progression on the y-axis. Plot the regression line on this same graph with appropriate labels on each axis.

Late assignments policy (10% off per week)

Weekly Schedule

- There will be a graded assignment assigned each odd week and due the following week by midnight the day prior to the live session.

- The schedule also includes many asynchronous exercises in addition to the assignments.

- Please complete each week’s asynchronous exercises 24 hours before each live session.

Weekly Schedule

- Week 1. Machine Learning Basics
  - *Readings: Géron, Chapter 1*
- Week 2. Data Analytics Project and Process Management
  - *Readings: Géron, Chapter 2*
  - *Complete Assignment 1*
- Week 3. Classification
  - *Readings: Géron, Chapter 3*
- Week 4. Model Training
  - *Readings: Géron, Chapter 4*
  - *Complete Assignment 2*
- Week 5. Support Vector Machines
  - *Readings: Géron, Chapter 5*

Weekly Schedule

- Week 6. Decision Trees
  - *Readings: Géron, Chapter 6*
  - *Complete Assignment 3*
- Week 7. Ensemble Trees
  - *Readings: Géron, Chapter 7*
- Week 8. Artificial Neural Nets
  - *Readings: Géron, Chapter 10*
  - *Complete Assignment 4*
- Week 9. Training Deep Neural Networks
  - *Readings: Géron, Chapter 11*
- Week 10. Custom Models in TensorFlow
  - *Readings: Géron, Chapter 12*
  - *Complete Assignment 5*

Attendance Policy & Program Mission

- Attendance at all live session meetings is mandatory.

- Our MS in Data Science provides students with a broad course of study in programming, algorithms, statistics, and data management, as well as a depth of understanding in specific fields such as data mining, machine learning, and parallel systems. Graduates of the data science program go on to work in a wide variety of careers, including business, government, education, and the natural sciences.

Honor Code and Academic Integrity

- All students are expected to abide by the University of Denver Honor Code. These expectations include the application of academic integrity and honesty in your class participation and assignments. Violations of these policies include but are not limited to
  - Plagiarism, including any representation of another's work or ideas as one's own in academic and educational submissions
  - Cheating, including any actual or attempted use of resources not authorized by the instructor(s) for academic submissions
  - Fabrication, including any falsification or creation of data, research, or resources to support academic submissions
  - Violations of the Honor Code may have serious consequences including, but not limited to, a zero for an assignment or exam, a failing grade in the course, and reporting of violations to the Office of Student Conduct.

Diversity, Inclusiveness, Respect

- DU has a core commitment to fostering a diverse learning community that is inclusive and respectful. Our diversity is reflected by differences in race, culture, age, religion, sexual orientation, socioeconomic background, and myriad other social identities and life experiences. The goal of inclusiveness, in a diverse community, encourages and appreciates expressions of different ideas, opinions, and beliefs, so that conversations and interactions that could potentially be divisive turn instead into opportunities for intellectual and personal enrichment.
- A dedication to inclusiveness requires respecting what others say, their right to say it, and the thoughtful consideration of others' communication. Both speaking up AND listening are valuable tools for furthering thoughtful, enlightening dialogue. Respecting one another's individual differences is critical in transforming a collection of diverse individuals into an inclusive, collaborative, and excellent learning community. Our core commitment shapes our core expectation for behavior inside and outside of the classroom.

Instructor biography

My background

- Ph.D. in Computer Science (UTK, 2010)
  - Detect abnormal events using wireless sensor networks and mobile robots
  - Intruder detection (**demo movie**) **[Li** and Parker, *IROS* 2008, and *IROS* 2010]
  - Novel missing data imputation method [**Li** and Parker, *Information Fusion*, 2012]
  - Detect volcano activities and highway accident

  [**Li**, Thomason and Parker, *Human Behavior Understanding in Networked Sensing*, 2013]

- PostDoc in Biomedical Engineering (UTK, 2010-11)
  - Detect toxic nanoparticles using plant-based sensor networks

  [**Li** et al, *Nanotechnology*, 2011] & [Lenaghan, **Li** et al, *IEEE transactions on Nanotechnology*, 2014]
  - Drug delivery using machine learning approach [**Li** et al, *PLoS One*, 2012]

- IRTA PostDoc, Research Fellow, and Staff Scientist in Biostatistics and Computational Biology Branch (NIEHS/NIH, 2011-)
  - Data mining and develop new computational methods
  - Mentor trainees within bioinformatics group

Current work

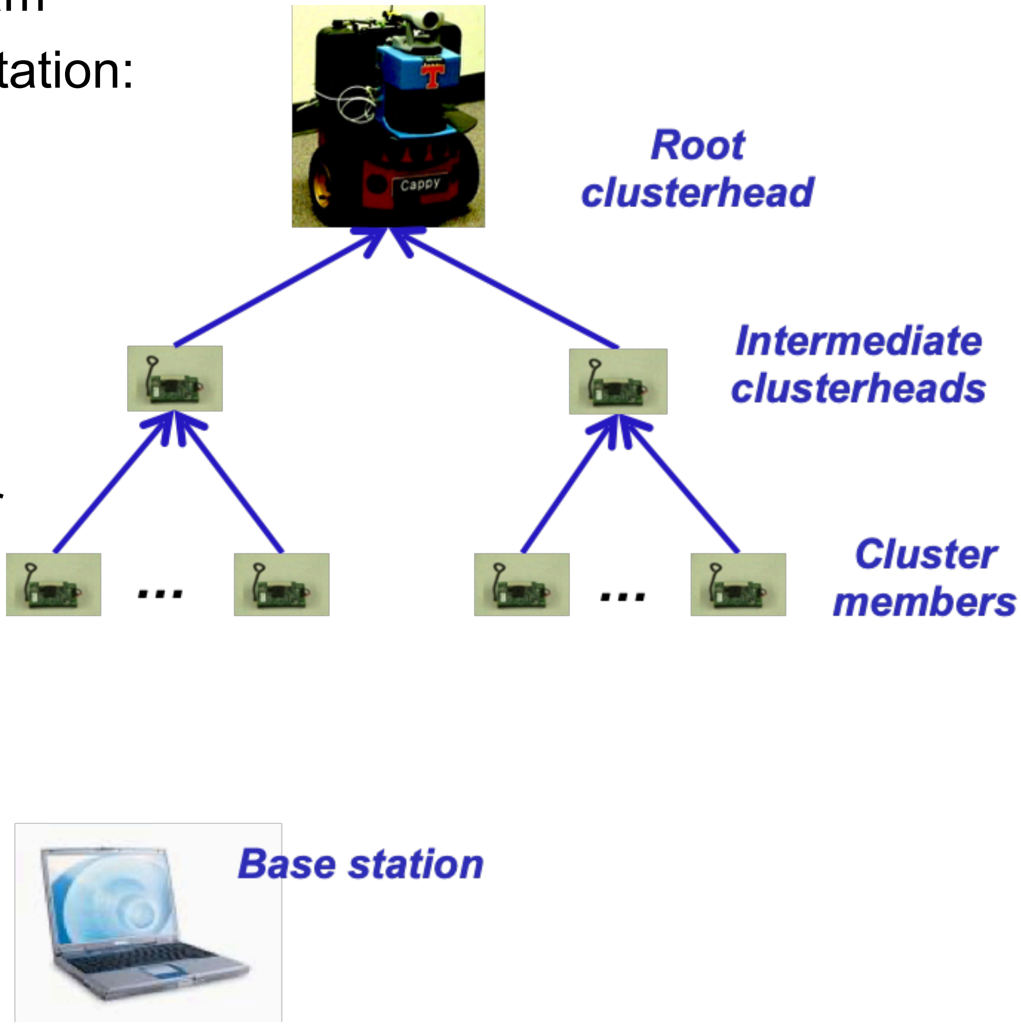
- ChIP-seq data (ENCODE)
  - Characterize constitutive CTCF binding sites [**Li** et al, *BMC Genomics*, 2013]
  - T-KDE method: genome-wide identification of constitutive protein binding sites [**Li** et al, *BMC Genomics*, 2014]
  - Identify p53 binding sites using T-KDE [Nguyen et al, *NAR*, 2018]
- RNA-seq data (TCGA)
  - Pan-cancer classification using gene expression data [**Li** et al, *BMC Genomics*, 2017]
  - Identify genomic characteristic of BRAF V600K vs V600E melanoma [**Li** et al, *Melanoma research*, 2017]
  - GPC6: a putative biomarker for metastatic progression of melanoma [**Li** et al, *Plos one*, 2017]
  - Identify genes that predictive of tumor purity [**Li** et al, *BMC Genomics*, 2019]
  - Identify genes that predictive of drug responses [**Li** et al, *BMC Genomics*, 2021]
- Single-cell data from wet lab component of our group
- Electronic Health Records (EHR) from COVID19 patients in UNC hospitals
- Electroencephalogram (EEG) data from sleep studies in UNC hospitals

Collaborative work

- Find maternal urinary metabolite concentrations that are predictive of fetal growth restriction [Clinton et al, *Scientific Reports*, 2020]
- Complete deconvolution for gene expression data from bulk tissues [Kang et al, *PLoS Computational Biology*, 2019] and [Kang et al, *BMC Bioinformatics*, 2021]
- Knockoff boosted tree for model-free variable selection [Jiang, Li, and Motsinger-Reif, *Bioinformatics*, 2020]
- HNF4α confers sensitivity to methionine restriction through regulation of sulfur amino acid metabolism in human hepatocellular carcinoma [Xu et al, *Nature Communication*, 2020]
- Detect multi-SNP epistasis using nuclear families [Nodzenski et al, *Bioinformatics*, 2021]
- Evaluating the Diversity Outbred mice as a model for human obesity and metabolic syndrome using a machine learning approach
- Find dietary variables that predictive of Antinuclear Antibodies (NHANES)

Intruder detection: hierarchical networking & learning

- All nodes run the same program
- Sample commands for base station:
  - Configure cluster setting
  - Set learning parameters
  - Clear learned prototypes
  - Pause and start the learning
  - Display the learning
  - Set radio transmission power
  - Set microphone sensitivity
  - Change sampling rate



Root clusterhead

Intermediate clusterheads

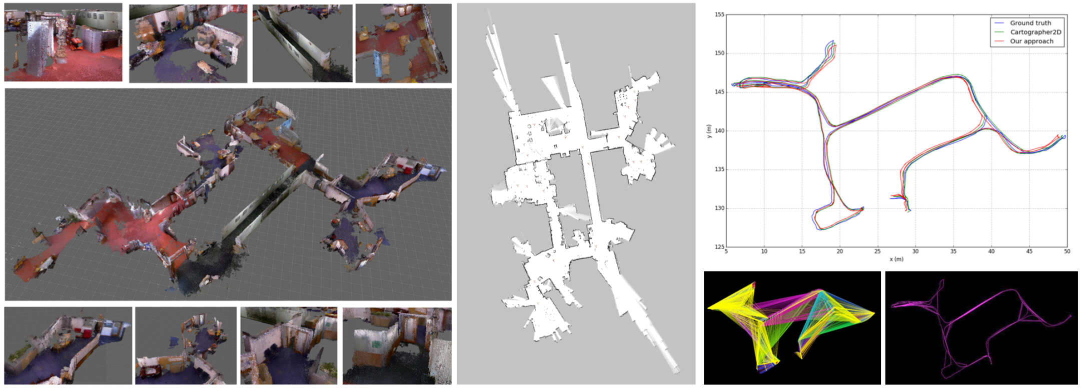
Cluster members

Base station

Key ML algorithms used in intruder detection

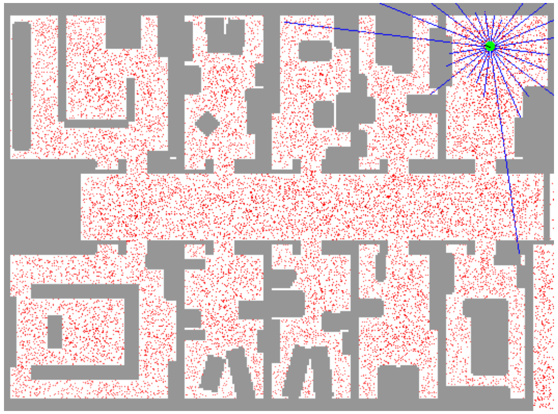
- Offline learning
  - Map building: Simultaneous Localization and Mapping (SLAM)
- Online learning
  - Anomaly detection: Fuzzy ART neural network
  - Robot localization: Monte-Carlo localization
  - Robot path planning: Wave-front path planning

SLAM: build map



Source: https://github.com/shannon112/AMIR-SLAM

Monte-Carlo localization



Source: https://www.cs.cmu.edu/~thrun/tutorial/sld041.htm

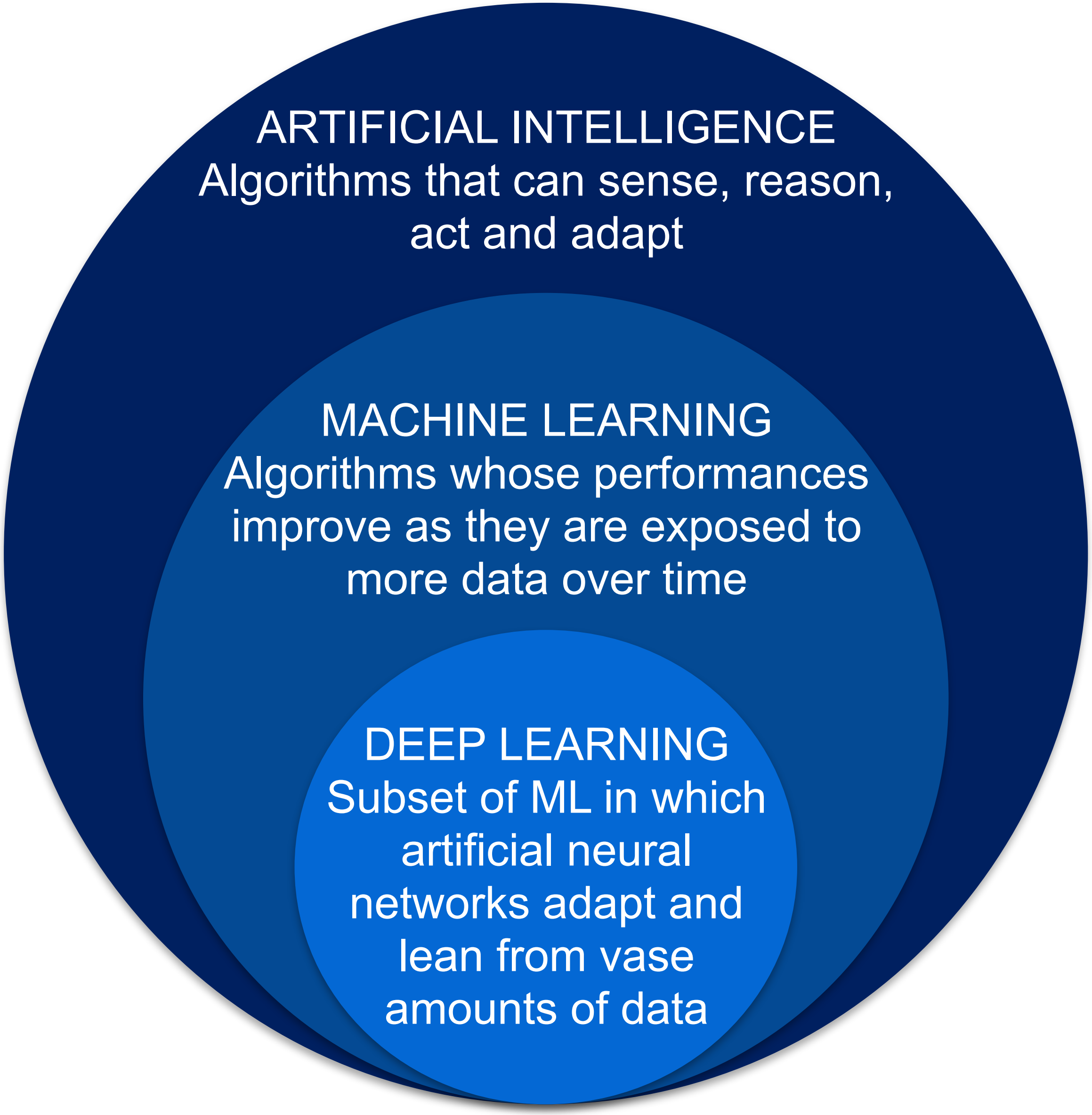
Intruder detection demo

Introductions

Please tell us about yourself

- Your name and what we should call you
- Where are you from?
- What was your undergrad major?
- What do you hope to learn from this class?
- What is your experience with Python?
- What o/s do you use?

Introduction to machine learning



ARTIFICIAL INTELLIGENCE
Algorithms that can sense, reason, act and adapt

MACHINE LEARNING
Algorithms whose performances improve as they are exposed to more data over time

DEEP LEARNING
Subset of ML in which artificial neural networks adapt and lean from vase amounts of data

Machine learning emphasizes high dimensional prediction problems. [Wasserman, 2012]

Figure modified from: https://www.codeproject.com/articles/1185501/How-to-Get-Started-as-a-Developer-in-AI

Types of machine learning

- Features engineering
  - Feature selection
  - Feature extraction

- Unsupervised: outcome is unknown
  - Clustering
  - Kernel density estimation

- Supervised: outcome is known
  - Classification
  - Regression

Feature Engineering

Feature selection

- Select a subset of features from the original feature set
- Suitable for high dimensional data
  - Image processing
  - Text processing
  - Biomedical data
- Find the optimal subset is NP-hard
- Main methods
  - Filtering
  - Wrap around a classifier/regressor
  - Embedded within a classifier/regressor

Filter feature selection method

- Example method
  - Statistical testing (T-test, F-test, ANOVA)
  - Correlation (Pearson, Spearsman)
  - Information gain
  - Variance threshold
- Is independent of the machine learning model
- Is generally univariate or low variate
- Is simple and fast
- Scales well to high dimensional data

Filter: statistical hypothesis testing

- The null hypothesis is that a given feature is not different between the two classes (normal vs disease).

- The alternative hypothesis is that the feature is different between the two classes.

- The hypothesis testing is performed by calculating a statistic (eg, the t-statistic) on the values of the feature of interest measured in the two classes.

- The computed value of the statistic is then compared with a threshold, calculated from a model (eg, the t-distribution) & a desired significance level (eg, 5% or 1%).

Filter: statistical hypothesis testing

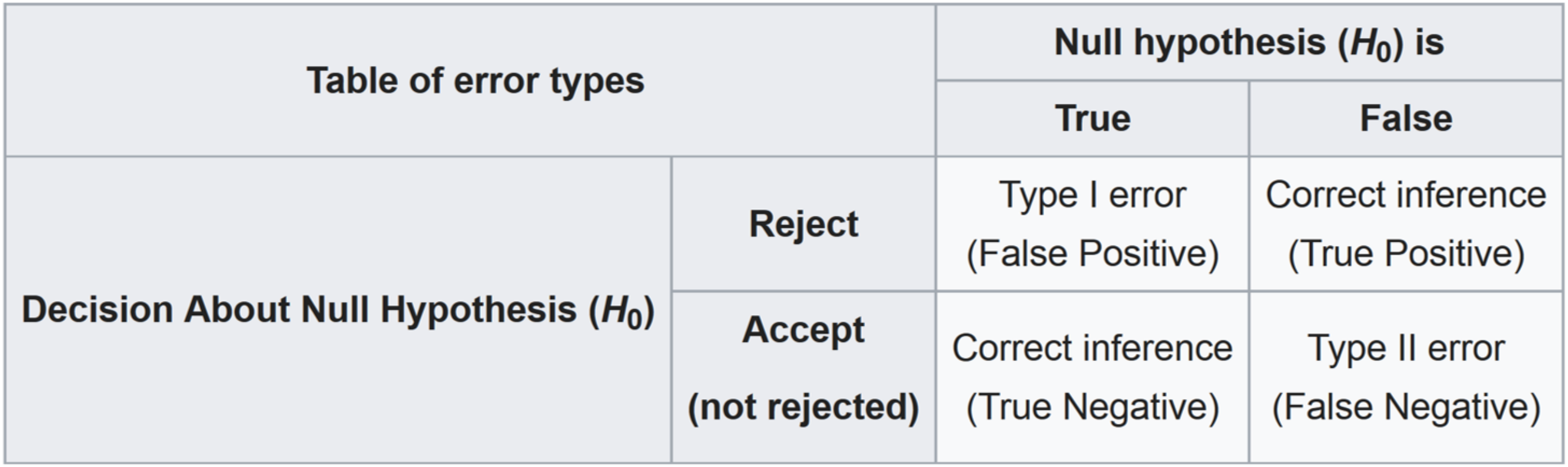


Table of error types		Null hypothesis (H_0) is	
		True	False
Decision About Null Hypothesis (H_0)	Reject	Type I error (False Positive)	Correct inference (True Positive)
	Accept (not rejected)	Correct inference (True Negative)	Type II error (False Negative)

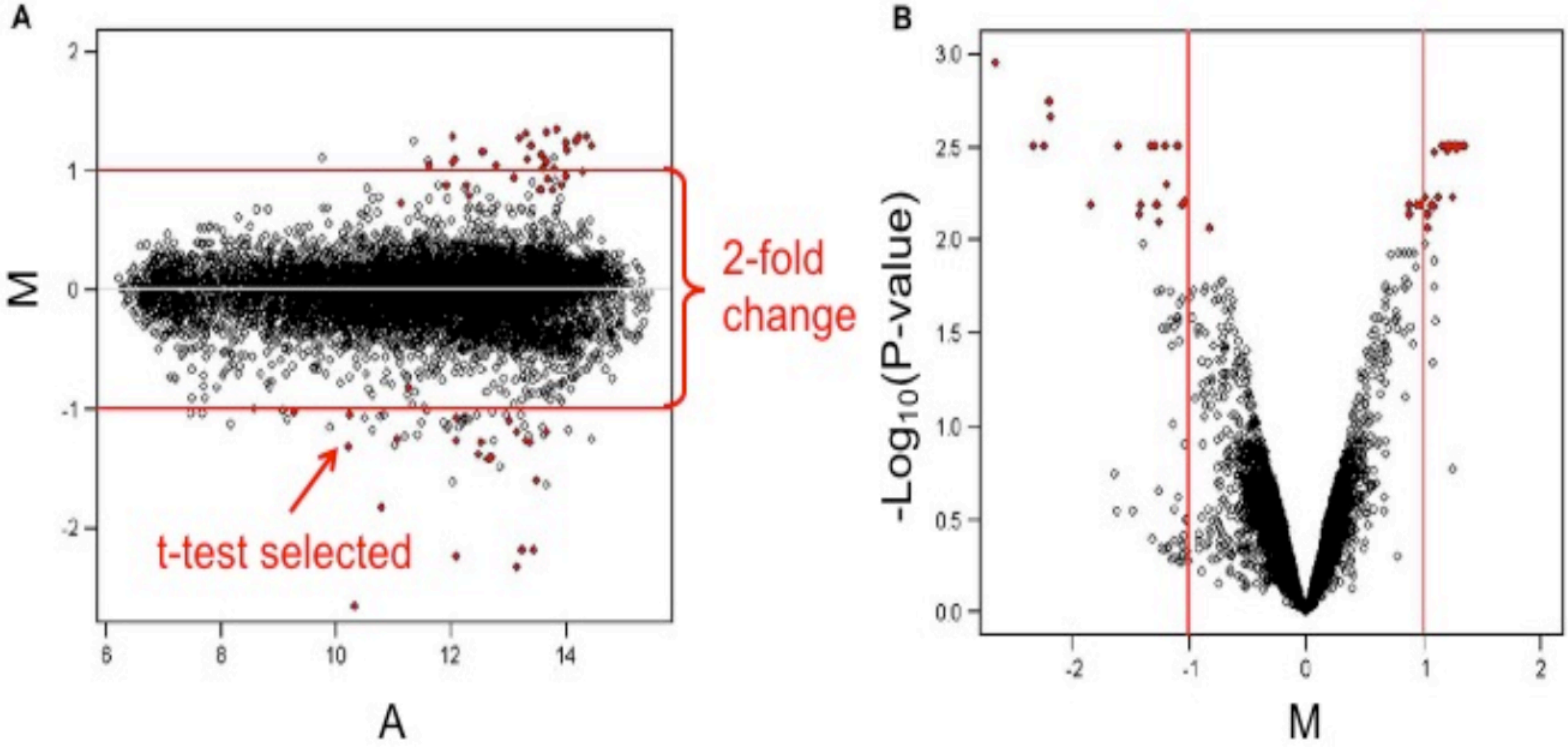
- The significance level is chosen before the test and represents % of Type I error that we are prepared to accept.
  - Eg, a significance level of 1% means that, on average, there will be one false positive gene for every 100 genes identified as differentially expressed.
- The statistical power of a technique is a measure of its ability to identify true positives.

Filter: statistical hypothesis testing

- Formulate a null & alternative hypothesis for every feature
- T-test: difference in means between 2 classes divided by the standard deviation
  - compares the difference in the mean values between 2 classes
  - taking into account the variability of the data
- F-test: between-group variability divided by within-group variability considering a decomposition of the variability in a collection of data in terms of sums of squares (SS)
- These SS are constructed so that the statistic tends to be greater when the null hypothesis is not true.
- ANOVA F-test: used when comparing > 2 groups

Adjust for multiple testing

T-test for gene selection application



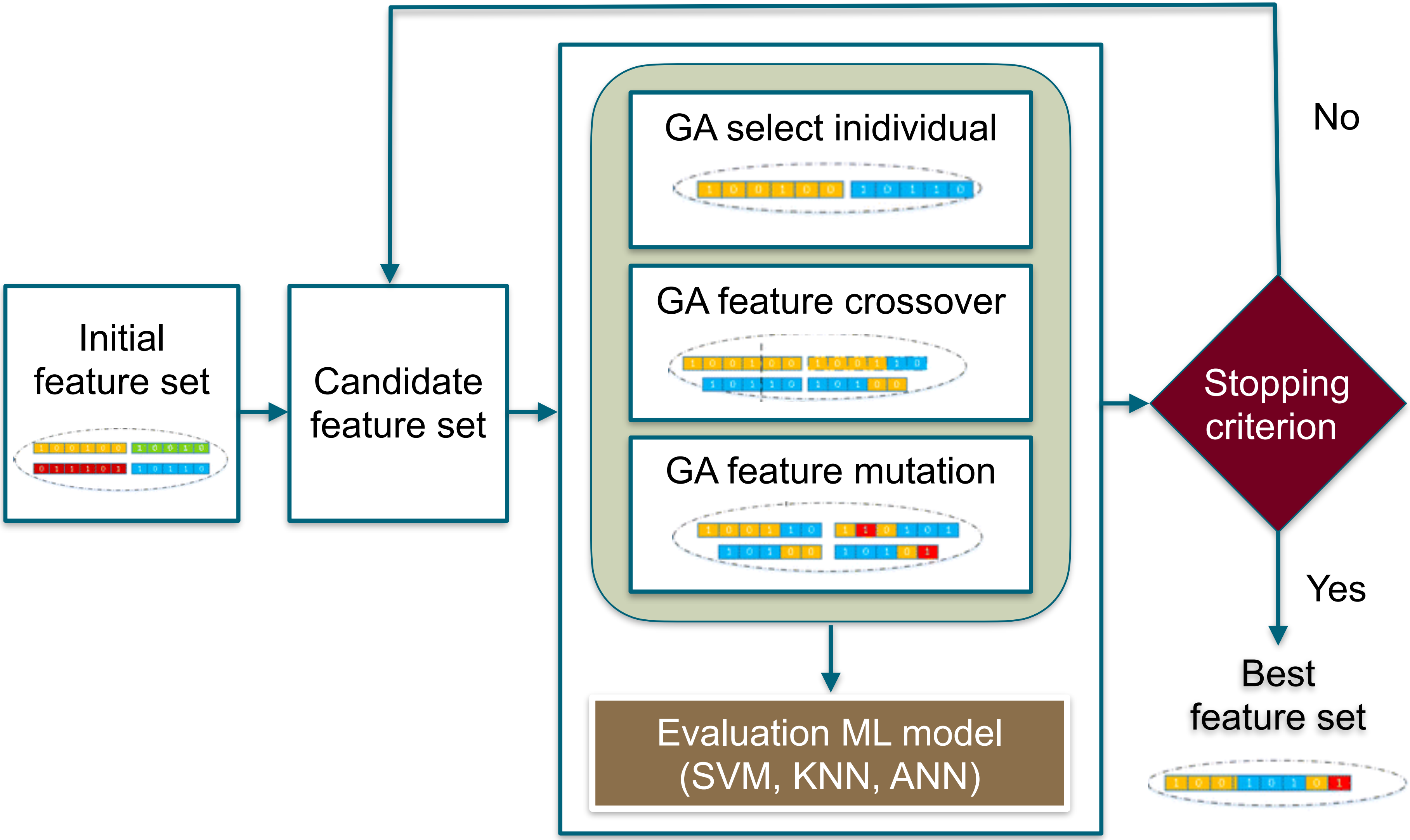
MA plot: the difference of the log-intensity of the two channels (log-ratio) against the average log-intensities (mean average).

Source: Cowley & Ying, 2011

Wrapper feature selection method

- Example method
- Sequential Forward Selection (SFS)
    - Genetic Algorithm (GA)
- Combines feature selection algorithm with a classifier/ regressor to evaluate selected feature(s)
    - First, a selected subset of feature(s), then pass down to a machine learning model for evaluation
- Is a search-based algorithm
    - Forward search (iteratively add features)
    - Backward search (iteratively remove features)
- Finding the optimal subset is NP-hard
- Selected features directly tie with model performances

Wrapper: Genetic algorithm (GA)



Initial feature set

Candidate feature set

GA select inidividual

GA feature crossover

GA feature mutation

Evaluation ML model (SVM, KNN, ANN)

Stopping criterion

No

Yes

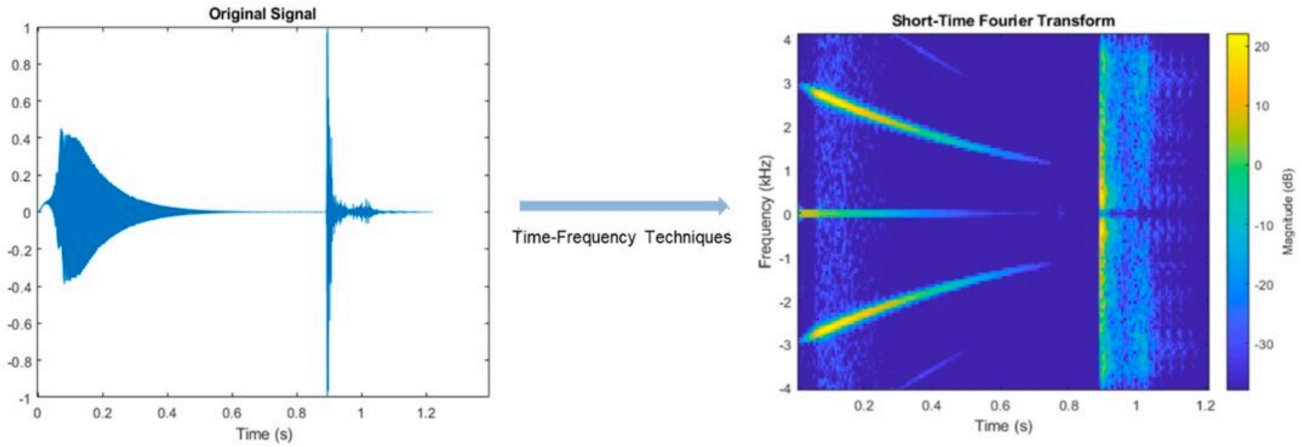
Best feature set

Embedded feature selection method

- Main methods
  - L1 (LASSO) regularization
  - Decision trees
- The machine learning model itself has feature selection procedure inside

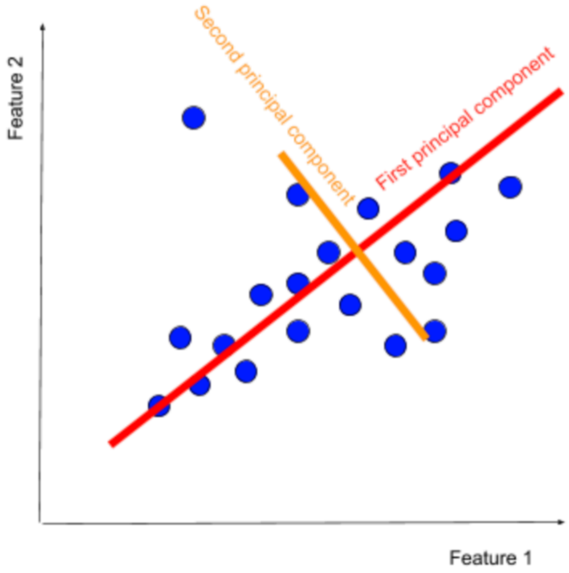
Feature extraction

- Generate new features from existing set
- Add additional helpful information for prediction
  - Image processing
  - Audio/speech processing
  - Medical data like EEG, EKG, and EMG
- Principal component analysis (PCA)
- Independent component analysis (ICA)



Principal component analysis (PCA)

- The principal components are vectors.
- The original data can be represented as feature vectors.
- PCA allows us to go a step further and represent the data as linear combinations of principal components.
- Principal components remove noise by reducing a large number of features to just a couple of principal components.
  - Principal components are orthogonal projections of data onto lower-dimensional space.
- In theory, PCA produces the same number of principal components as there are features in the training dataset. In practice, picking just a few of the first components sufficiently approximates the original dataset.
- The result is a new set of features in the form of principal components, which have many practical applications.



Source: https://www.keboola.com/blog/pca-machine-learning

Principal component analysis (PCA)

1. Feature standardization (a mean of 0 and a variance of 1).
2. Obtain the covariance matrix computation.
3. Calculate the eigendecomposition of the covariance matrix. We calculate the eigenvectors (unit vectors) and their associated eigenvalues (scalars by which we multiply the eigenvector) of the covariance matrix.
4. Sort the eigenvectors from the highest eigenvalue to the lowest.
5. Select the number of principal components. Select the top N eigenvectors (based on their eigenvalues) to become the N principal components. The optimal number of principal components is both subjective and problem-dependent.

Source: https://www.keboola.com/blog/pca-machine-learning

Principal component analysis (PCA)

- Assumes a correlation between features
- Is sensitive to the scale of the features
- Is not robust against outliers
- Assumes a linear relationship between features
- Cannot handle missing values