Intro to Machine Learning
• Intro to Machine Learning - what is learning?
• Data - Matrix type
  - algebraic notations
• Heuristics and Quantitative rules
• Error measurement
  - training VS testing error, Cross Validation
  - overfitting
What is machine learning? Supervised learning

- data given with labels
- but learning setup up to us
- ML algorithm up to us
- error measurement has to be meaningful
What is machine learning? Graph learning

- data defined by links or analogies or connections
- for example social networks, or web links
- task: identify object properties from links
- tasks: detect graph patterns
What is machine learning? Clustering

- data given without labels
- task: group similar data points

13 States clustered into 51 Custom Ecoregions
What is machine learning? Time series analysis

- data that evolves with time
- like stocks or patient records
- task: predict future behavior
- task: detect anomalies
Matrix data

- $m$ datapoints/objects $X=(x_1, x_2, \ldots, x_d)$
- $d$ features/columns $f_1, f_2, \ldots, f_d$
Heuristic rules / decisional

- If fever > 100, patient has flu
- If email contains words “free” or “porn”, it is spam
- If a web page contains ngram “Michael Jackson”, it is relevant to the user
- If age < 22 and sex = F and highschool_diploma = Yes, then eligible for application
- If income_per_capita <$1000, region prone to civil war
- If romantic = Yes and comedy = Yes and Orlando_Bloom = Yes, then movie success among females aged 20-40
- If Nasdaq_Computer_Index = Gain and Apple announces new Ipad, then AAPL_Stock = Buy
• if $3 \times \text{exam\_grade} + 2 \times \text{HW\_grade} > 55$, then student can pass
• if $\frac{\text{blood\_pressure}}{\log(\text{age})} > 3$, recommend medicine
• if $\text{rent} + \text{food} + \text{bills} < \frac{1}{2} \text{salary}$, loan for $\frac{1}{2}$ salary possible
### Matrix data / training VS testing

| AUTO | REL | BUL | CYR | CZE | DEN | EST | FIN | FRA | GBR | GRC | HUN | ITA | LAT | LIT | LUX | MLT | NLD | POL | POR | ROM | SVE | SLO | ESP | SWE | GER |
|------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
| 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 | 0.03 |
| 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 | 0.04 |

- **testing set has to be independent of training set**
  - or else testing result is inconclusive
  - and not reliable
- usually the data is partitioned before running any ML algorithm
Overfitting

- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful: the purpose of the learning model is to applicable to new data (testing)
Overfitting

- might be capable to create a model that essentially memorizes all training dataset
  - for example a decision tree deep enough
- that is not useful: the purpose of the learning model is to applicable to new data (testing)
Overfitting

- as we keep training (insisting on ability to classify training set), the performance on the training set (green) becomes unrealistically small
  - model becomes more complex
- but at the same time ability to predict/classify new data (pink) worsens
Cross Validation Setup

- **Split Data in K Folds**
- **Execute K Independent Learning Trials**:
  - Train on K-1 folds
  - Test on remaining fold
- **Measure Testing Performance**
- **Average Results across K Trials**
for objects like text documents or images:
- extract features (to obtain matrix form)
- annotate (to obtain labels)
• about 4000 emails
• 54 features numerical
• two classes: spam / no_spam
Housing dataset

- 1300 houses
- 13 features (numerical)
- label: purchase prices (quantitative)
Digits dataset

- 60000 images of scanned digits
- 26x26 pixel per image, black or white
- features not extracted
- 10 classes: 0, 1, 2, ..., 9
Documents dataset

- 20,000 news articles (text)
- features not extracted
- 20 categories: religion, music, computers, sports, etc.
main focus: learning algorithms
main focus: hands-on practice on datasets
secondary focus: analysis, error measurement
secondary focus: features, representation
typical module subtasks / objectives

• THEORY
  - explain/understand fundamental mechanism
  - proof (math, intuition)
  - pseudocode

• CODE
  - run existing code
  - implement and demo your code
  - data handling: features, dimensionality, scale, missing values, normalization
  - computational issues: memory, cache, CPU, disk

• EVALUATION
  - setup
  - performance measurement, comparison
  - analysis/failure of procedure behavior

• HOWTO
  - practical advise, hacks, heuristics
  - communicate on topic well: email, forums
  - where to look online