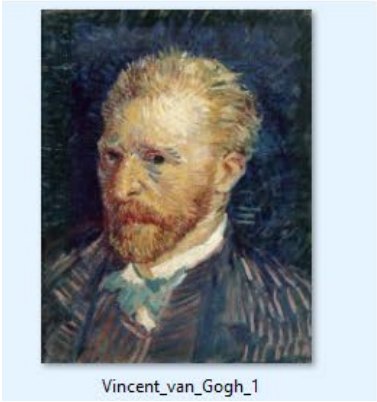
Introduction to Machine Learning

Final Project

- Requirements
- Timeline
- Demo and Reports

Application

- Artwork classification
  - https://www.kaggle.com/ikarus777/best-artworks-of-all-time


Vincent_van_Gogh_1


Vincent_van_Gogh_2


Vincent_van_Gogh_3


Vincent_van_Gogh_4


Vincent_van_Gogh_5


Pierre-Auguste_Renoir_11


Pierre-Auguste_Renoir_12


Pierre-Auguste_Renoir_13


Pierre-Auguste_Renoir_14


Pierre-Auguste_Renoir_15

- Only 8 classes (artists) will be considered
- Testing:
  - Given the feature of a new painting (not used in training), predict the artist: one of the 8 classes or other (total of 9 classes)
    - "other class"
      - Based on predicted probability (or decision value)
      - Train with C+1 classes
- Training data:
  - We selected a subset of images and mapped them to features (using CNN)

Minimum requirements

- 2 options for data normalization
- PCA with 2 options for number of components (dimensions)
- Data mapping using clustering (2 options)
- 4 different types of classifiers
- Ensemble methods: *Bagging* and *AdaBoost*
- Use k-fold cross validation with k=4 for all validations (nested 4-fold if needed)
- Use *Pipelines* and *GridSearch*
- Base your algorithm/parameter selection on accuracy, AUC of ROC, and F1-measure

Minimum requirements

- DO NOT
  - Combine all options and parameters in one giant GridSearch!
  - Use any algorithm/technique that was not covered in class
- DO
  - Consider few options at a time
  - Analyze the results
  - Justify your next set of options

Report 1

- Due April 18 (there will be penalty for late submission: **5points off per day**)
- Worth **30 points**, graded based on
  - Experiments, results, and analysis
  - Discussion/justification of remaining experiments
  - This report should include about 50% of all your experiments

Final report

- Due April 25 (no late submission)
  - *NO REPORTS WILL BE ACCEPTED OR GRADED AFTER 04/25*
- Worth 40 points, graded based on
  - Experiments, results, considered options, etc.
  - Discussion of the most important parameters that affect the results
  - Performance of the classifier (based on cross-validation of the provided data)
  - Analysis of the results
    - Visualization of correct samples, confused samples, etc.
    - Possible justification for misclassified samples

Demo and Test (of new images)

- 10 min demo (for each student) will be scheduled on **April 26 & 27**
- You will be asked to share your screen and run your code on new test images (*one of the C classes* or *others*)
- Worth max of 30 points.
  - Depending on all available results (including ours), we will grade your accuracy as
    - **Excellent** (30 pts)
    - **Good** (20 pts)
    - **Average** (10 pts)
    - **Don't make sense** (e.g. all assigned to the same class) or code doesn't run (0 pts)
- Bonus points
  - Best: + 10points Second: +5points Third: + 2points