Project 2: Predictive Modelling

STAT GU4243

Applied Data Science

Cynthia Rush
Columbia University

February 7, 2018

Though we're doing image classification, this is not a project about deep learning.

Project Description

- Carry out model evaluation and selection for predictive analytics on image data.
- Evaluate different modeling/analysis strategies and decide what is the best.
- Present sound evidence in the form of model assessment, validation, and comparison.
- Communicate your decision and supporting evidence clearly and convincingly in an accessible fashion.

Not a competition on prediction accuracy.

Model Comparison

Baseline Model

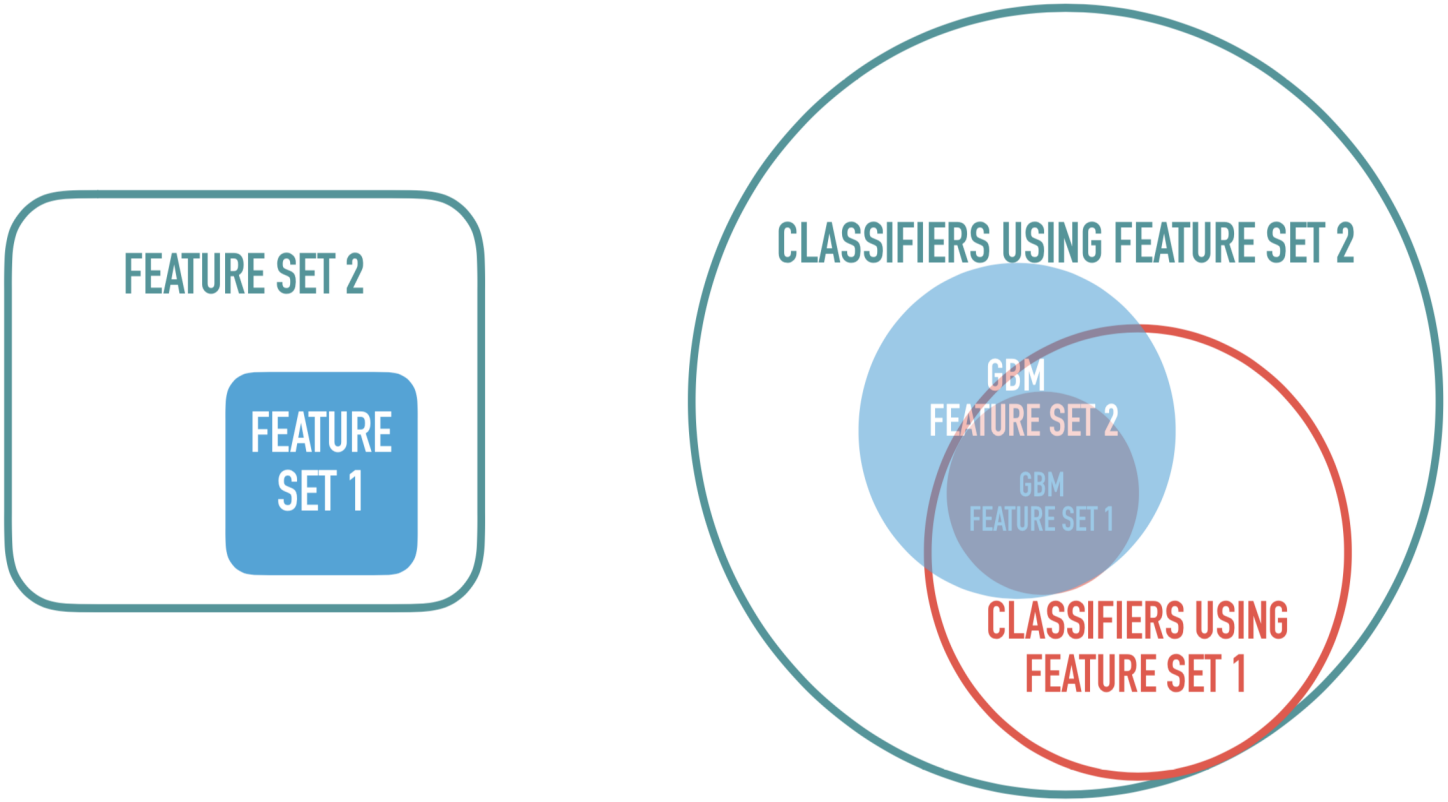
- Use provided SIFT descriptors as features.
- Implement a gradient boosting machine with decision stumps as a classifier.

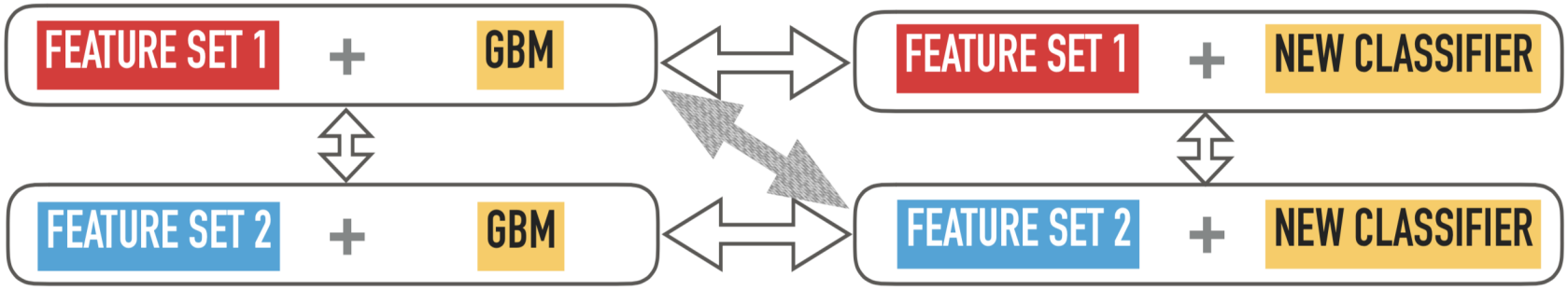
Task 1 will be implementing the above strategy and tuning it correctly

Proposed Strategy

- Consider better features and better models
- Implement structured comparison to establish the value added by new features and new methods

Nested model comparison structure.





Submission of Project 3

- A well-documented GitHub repo (following instruction given in the starter codes).
- A ‘main.rmd‘ that carries out the project.

Submission of Project 3

‘main.rmd‘ uses

(1) A file of feature processing codes (‘feature.R‘) that:

- Takes as input folder of images
- Outputs a folder of “feature” objects with features for the images
  - Format is RData, or other R readable file
  - Make sure you keep track of the file names of the images.

Submission of Project 3

‘main.rmd‘ uses

(2) A file of training codes (‘train.R‘) that:

- Inputs a path for training image features
- Inputs a CSV file containing training image names and labels
- Outputs trained classifiers (in the form of RData, or other R readable file); One for the baseline model and one for the new model.

Note that model training should include any necessary parameter tuning.

Submission of Project 3

‘main.rmd‘ uses

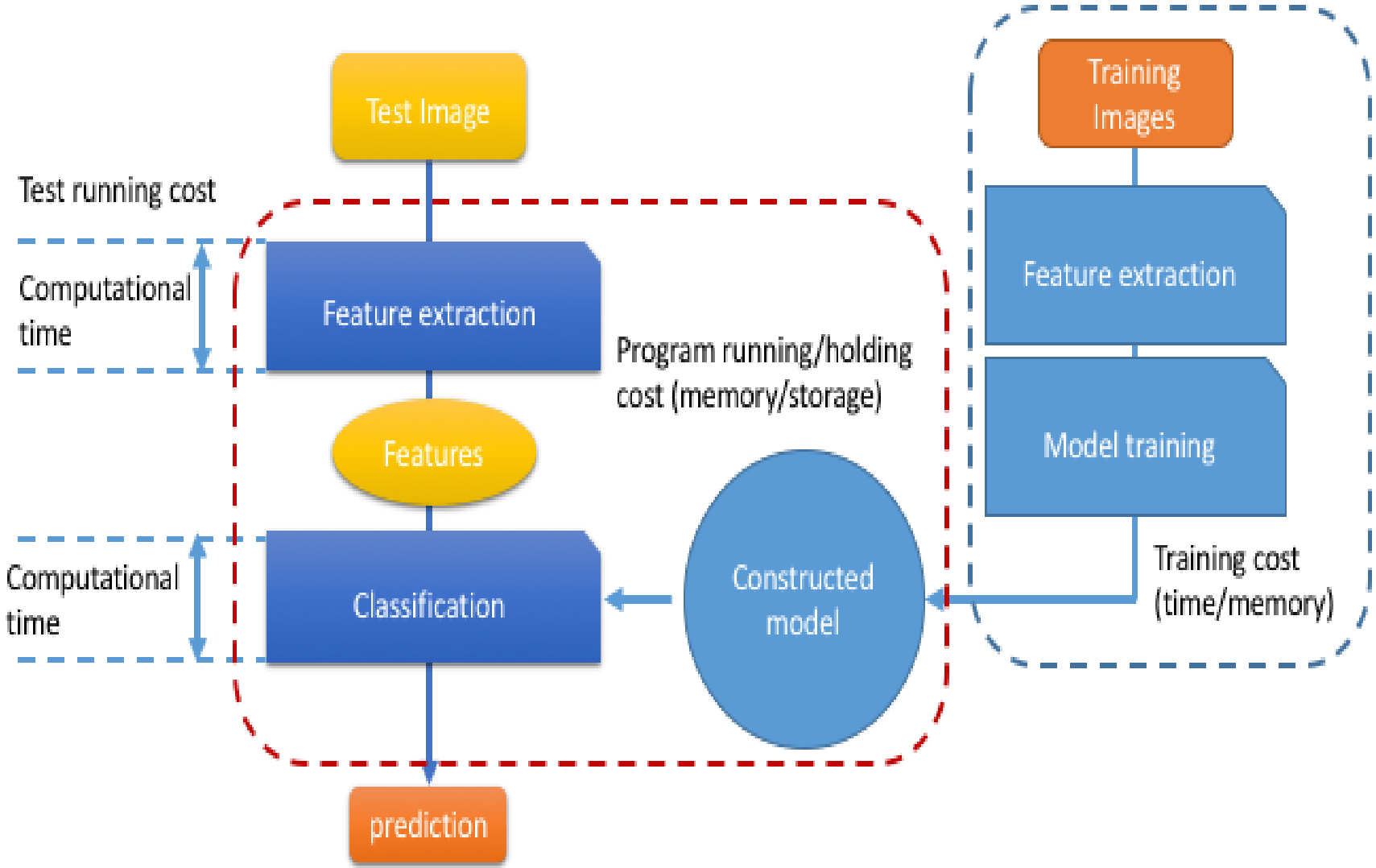
(3) A file of testing codes (‘test.R‘) that:

- Inputs a path for testing image features
- Inputs a trained classifier from the output of ‘train.R‘
- Output predicted labels

Submission of Project 3

- You can use any methods to generate features.
- On March 5th, we will first fork all project repos to save a time-stamped version of all your codes.
- On a new set of images and SIFT descriptors, each team will have 30 minutes to process them into features chosen.
- Submit the processed features as a folder of feature objects file. The feature objects should be readable by ‘train.R‘ and ‘test.R‘.
- We (the instruction team) will then run your ‘main.rmd‘ file from all submissions.
- For comparison, you are also required to submit predictions on the test images WITHOUT retraining your classifiers.

Submission of Project 3



Submission of Project 3

You should also prepare a presentation for this project

- Methodology details of the proposed solution
- Evaluation results as supporting evidence
  - Prediction performance comparison between baseline and new models.
  - Time/cost analysis.

Evaluation of Project 3

Ease of reproducibility by the client (5 points)

- Are codes for the proposed methods well-annotated and documented?
- Can the analysis be re-run nearly automatically using the 'main.rmd'?

Level of reproducibility (5 points)

- Can the client derive the same evaluation conclusion as presented in the team's final presentation?
- How close are the reported performances (presentation and online) to the reproduced performances?

Evaluation of Project 3

Portability of proposed strategies (5 points)

- Computational speed for feature extraction and model training.
- Computational speed for prediction.
- Memory use for model training and prediction.

Presentation and organization (5 points)

- Is the the intuition behind the proposed strategies convincing?
- Is it supported by adequate and appropriate evidence?
- Is the GitHub organized and prepared so that it's easy to understand the proposed strategies and their advantages and limitations?

- **Next Week:** Basic image analysis in R and image features
- **Two Weeks from Now:** GBM, Classification, Cross-validation

Group Projects

Channels of Communication

During class time

- Brainstorm
- Ask questions during tutorial

Before and after classes

- Piazza

If you have questions

- Piazza
- As a last resort, email

Working together

- You don't have to be in the same room at the same time to work together.
- You will work together in this course in the following ways:
  1. Face-to-face brainstorming
  2. Online discussion in a group forum
  3. Online video chat (say, via Google Hangout) with screen share
  4. GitHub collaboration

Project assignment on Classroom for Github

- I have created a Project 2 starter code folder
- I have assigned groups with group numbers (1-7) and shared the group info (including group number) on Piazza
- I will send assignment invitation links with instructions:
  - First, check whether your teammate already created a team for your group from the "Join an existing group".
  - If you cannot find your group's name (as shown on Piazza), please create the team using precisely the group name specified.