Cat or Dog: Predictive Modeling

STAT GU4243 - Applied Data Science

Group 6

Columbia University

March 5, 2018

1. Outline
2. Introduction
    - Us
    - Motivation
    - Scope
3. Method
    - Exploratory analysis
    - Feature extraction
    - Statistical machine learning models
    - Tuning and training
4. Results
5. Discussion

Group members

Wanting Cheng, Mingkai Deng, Jiongjiong Li, Kai Li, Daniel Parker

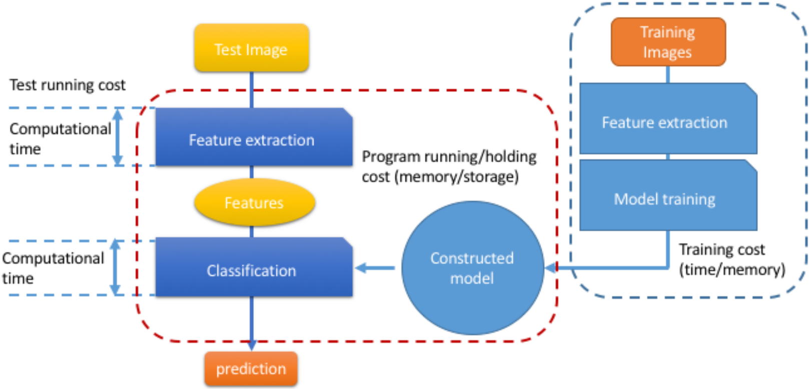
Why do this?—Motivation

Why do this?—Motivation

Spec & Scope

[C]*arry out model evaluation and selection for predictive analytics on image data …* [using] *a set of 4387 labeled images of cats and dogs …* *creat*[e] *a mobile AI program that accurately distinguishes between* [them] *… balance between the complexity of variables/features/models used and the predictive performance.*

Spec & Scope


Test Image
Test running cost
Computational time
Feature extraction
Program running/holding cost (memory/storage)
Features
Computational time
Classification
Constructed model
prediction
Training Images
Feature extraction
Model training
Training cost (time/memory)

Exploratory Analysis

What makes one animal different from another? [Intuition]
What approaches did previous semesters' groups employ? [Research]

Feature Extraction

1. SIFT = scale-invariant feature transformation.

Feature Extraction

1. SIFT = scale-invariant feature transformation.
2. HOG = histogram of oriented gradients.

Feature Extraction

1. SIFT = scale-invariant feature transformation.
2. HOG = histogram of oriented gradients.
3. LBP = local binary patterns.

Feature Extraction

1. SIFT = scale-invariant feature transformation.
2. HOG = histogram of oriented gradients.
3. LBP = local binary patterns.
4. HSV = hue, saturation, value.

Feature Extraction

1. SIFT = scale-invariant feature transformation.
2. HOG = histogram of oriented gradients.
3. LBP = local binary patterns.
4. HSV = hue, saturation, value.
5. RGB = red, green, blue.

Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.

Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.
2. Random forests.

Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.
2. Random forests.
3. TensorFlow/Keras neural network.

Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.
2. Random forests.
3. TensorFlow/Keras neural network.
4. Support vector machine.

Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.
2. Random forests.
3. TensorFlow/Keras neural network.
4. Support vector machine.
5. Adaptive boosting (“AdaBoost”).

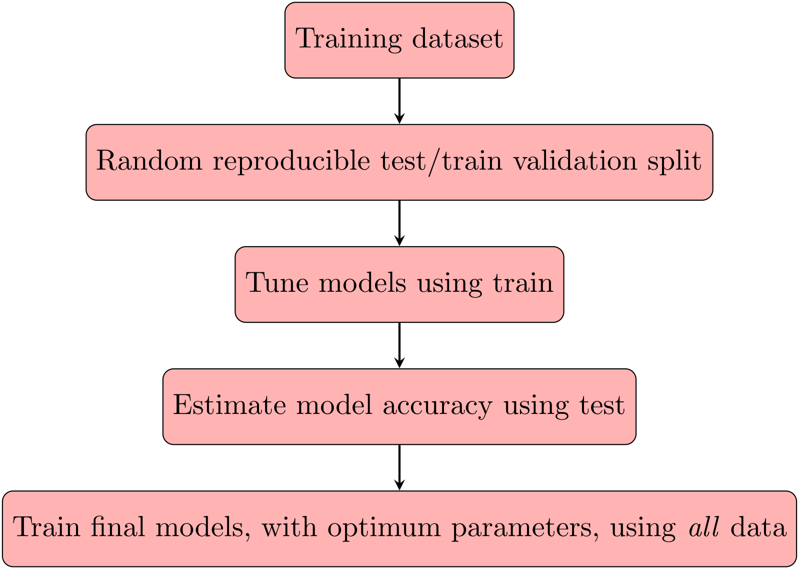
Statistical Machine Learning Models

1. Gradient boosting machine—the baseline.
2. Random forests.
3. TensorFlow/Keras neural network.
4. Support vector machine.
5. Adaptive boosting (“AdaBoost”).
6. Extreme gradient boosting (“XGBoost”).

Tuning and Training

Simplifying heuristic: use *all* features, rather than subsets. Preference for built-in package functions, rather than a generalized syntax.

How we finalized models—Flowchart



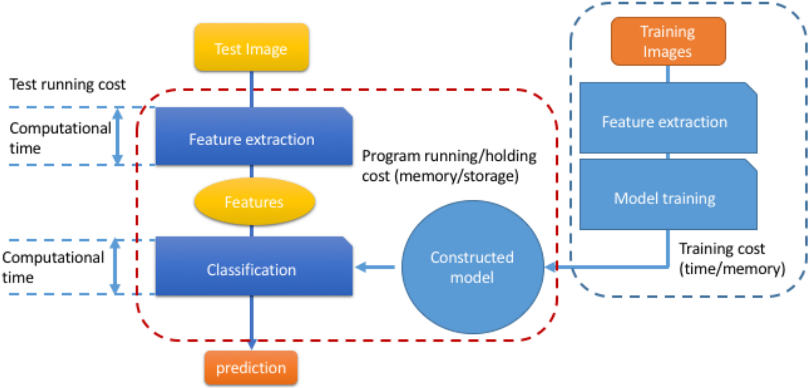
Training dataset

Random reproducible test/train validation split

Tune models using train

Estimate model accuracy using test

Train final models, with optimum parameters, using *all* data


Test Image
Test running cost
Computational time
Feature extraction
Program running/holding cost (memory/storage)
Features
Computational time
Classification
Constructed model
prediction
Training Images
Feature extraction
Model training
Training cost (time/memory)

Training Cost

Computation time and memory use for: 1. Feature extraction 2. Model training

Test / Use Cost

Computation time for: 1. Feature extraction 2. Classification

Accuracy

Comparison

Further directions to explore

1. Other extractions and combinations thereof.

Further directions to explore

1. Other extractions and combinations thereof.
2. Dataset manipulation to “grow” more training data for free.

Further directions to explore

1. Other extractions and combinations thereof.
2. Dataset manipulation to “grow” more training data for free.
3. Other models.

Further directions to explore

1. Other extractions and combinations thereof.
2. Dataset manipulation to “grow” more training data for free.
3. Other models.
4. Ensembling.

Further directions to explore

1. Other extractions and combinations thereof.
2. Dataset manipulation to “grow” more training data for free.
3. Other models.
4. Ensembling.
5. . . .

Thank you!