Cat or Dog: Predictive Modeling

STAT GU4243 - Applied Data Science

Group 6

Columbia University

March 5, 2018

Group members

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

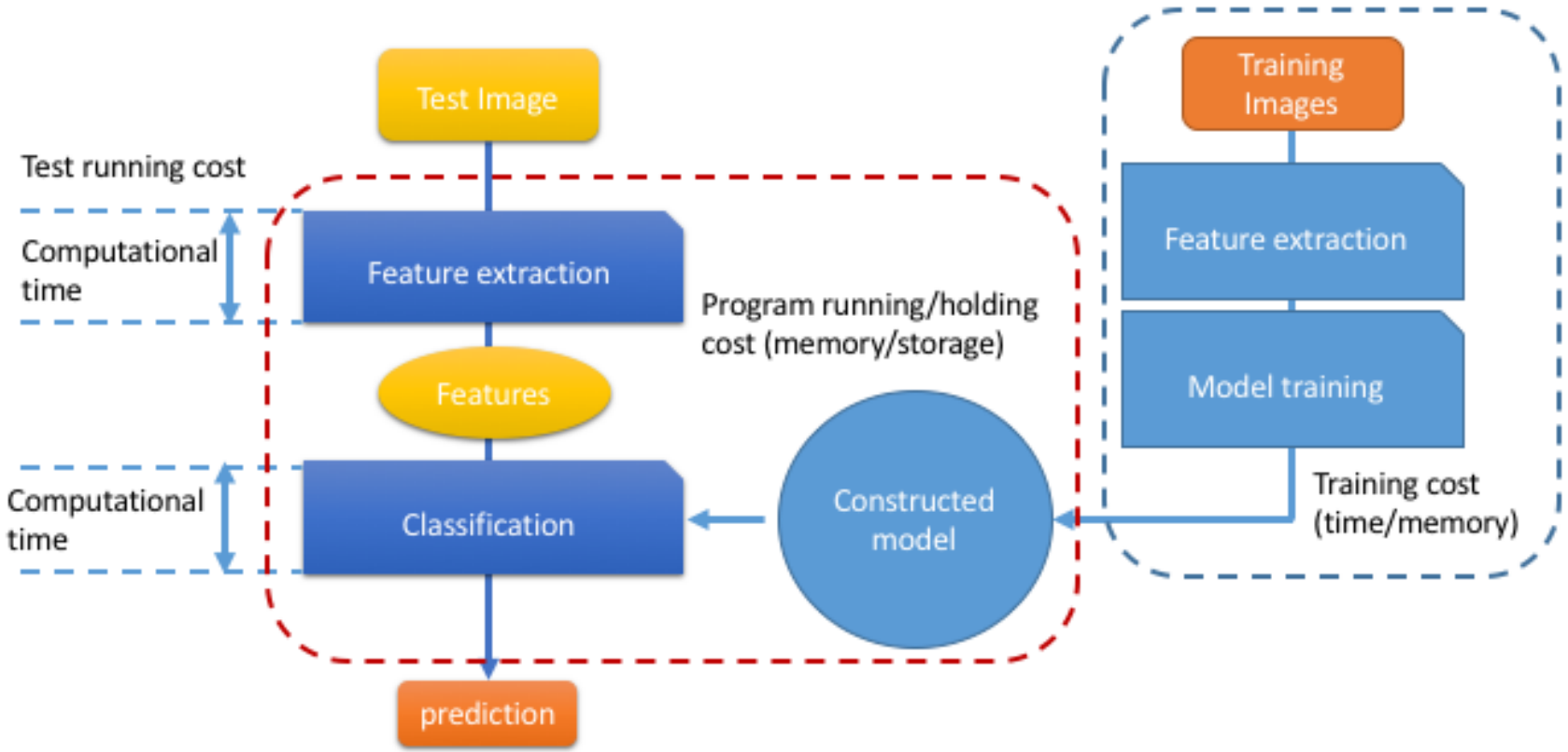
Why do this?—Motivation

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
Features
Computational time
Classification
prediction
Program running/holding cost (memory/storage)
Constructed model
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").

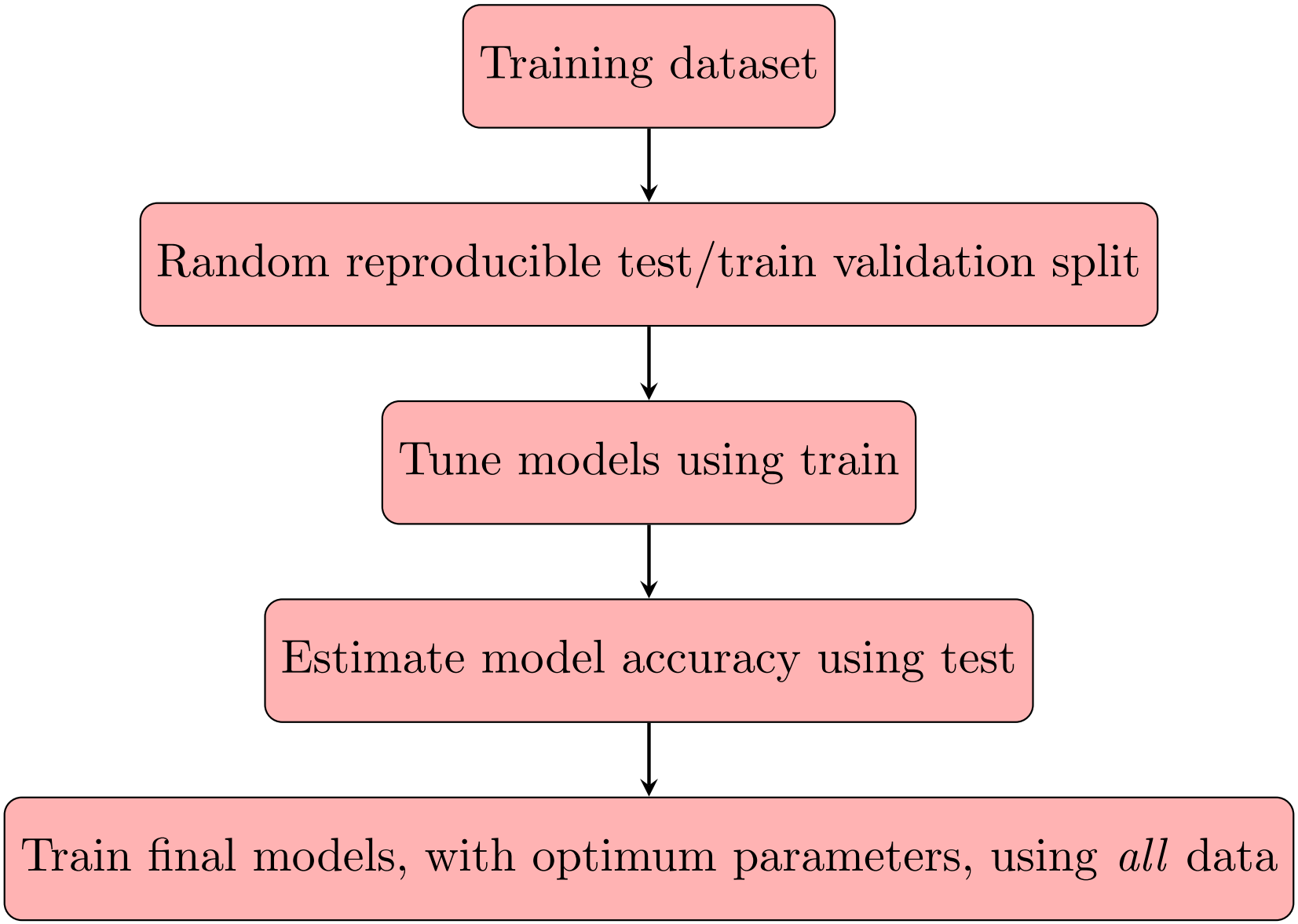
Statistical Machine Learning Models

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").

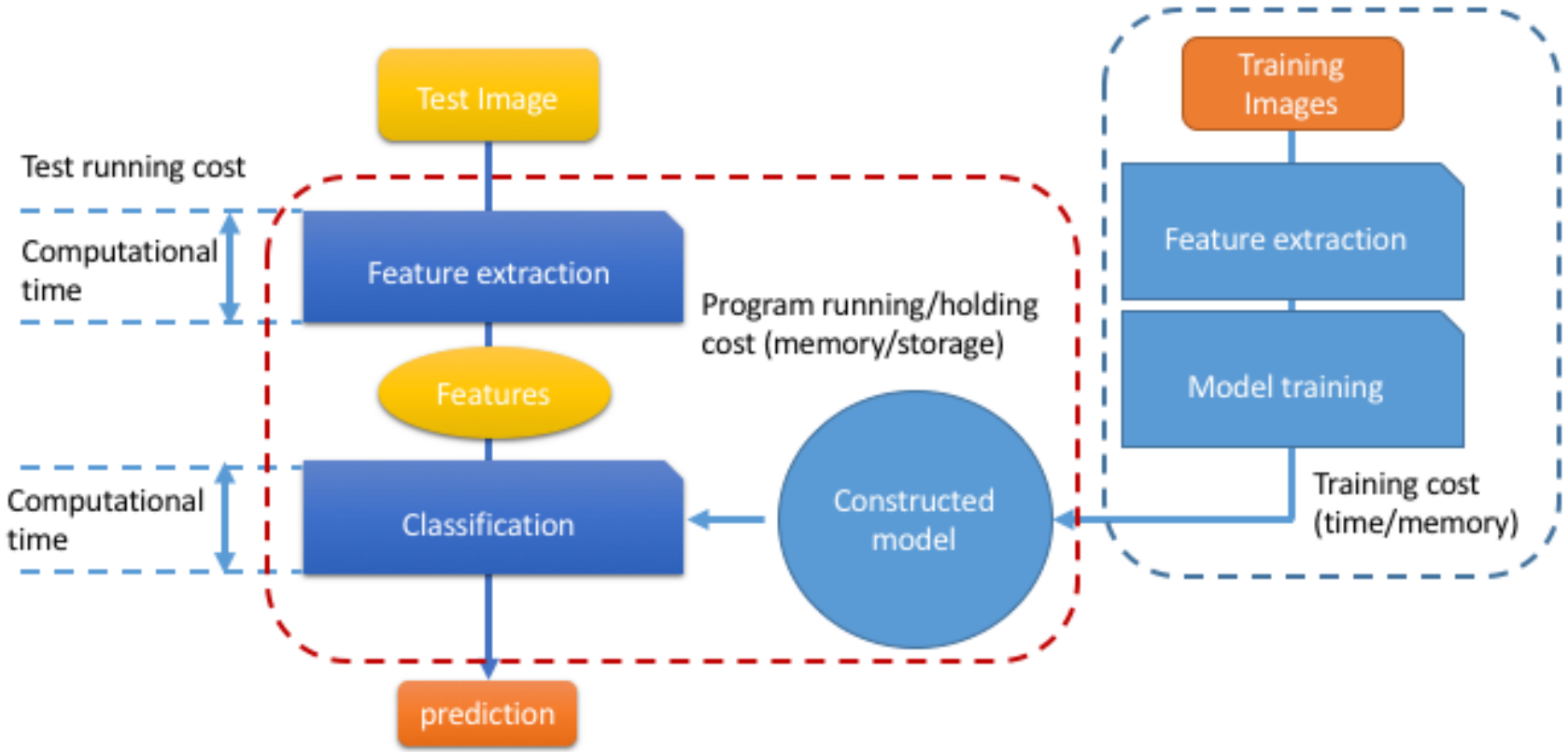
Tuning and Training

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

How we finalized models—Flowchart



Results


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

Feature extraction time

Feature type	Color	HOG	LBP
Time (m)	7	5	20

Table: Image processing time by feature, in minutes

Training Times—Computational Cost

Model	SIFT	Color	HOG	LBP
GBM	13.452	89.036	116.964	2.74
RF	174.1	905.883	1999.23	24.842
NN	31	65.81	61.58	28.12
SVM	3.484		33.099	0.805
XGBoost	3.829	16.851	27.005	1.67
AdaBoost	16.34	103.61	142.45	3.25

Table: Training time per model, in seconds

Prediction Accuracy

Model	SIFT	Color	HOG	LBP
GBM	73.25	69.5	75.25	69.
RF	72.25	73.	74.25	69.
NN	75.75	64.5	76.75	69.
SVM	77.5		77.5	69.75
XGBoost	72	72	77.25	66.5
AdaBoost	72.75	69.5	71.75	69.75

Table: Prediction accuracy by model, in percentage

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!