LDA, Decision Trees, and Extra Trees

on the MNIST and Yale B Datasets

Kudiyar Orazymbetov, Nico Casale

NCSU

2017/04/23

outline

motivation

-
- applications:

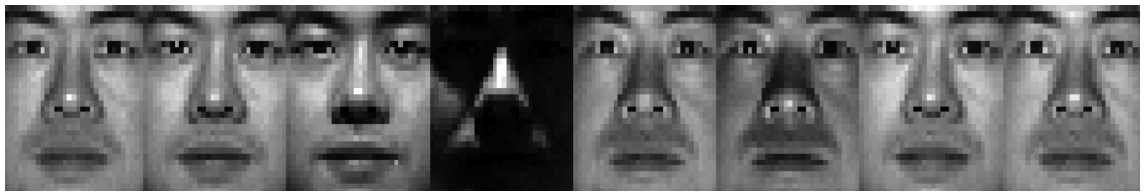
datasets

Modified Nat'l Institute of Standards and Technology (MNIST) database

- source: Yann LeCun et al. [1]
-

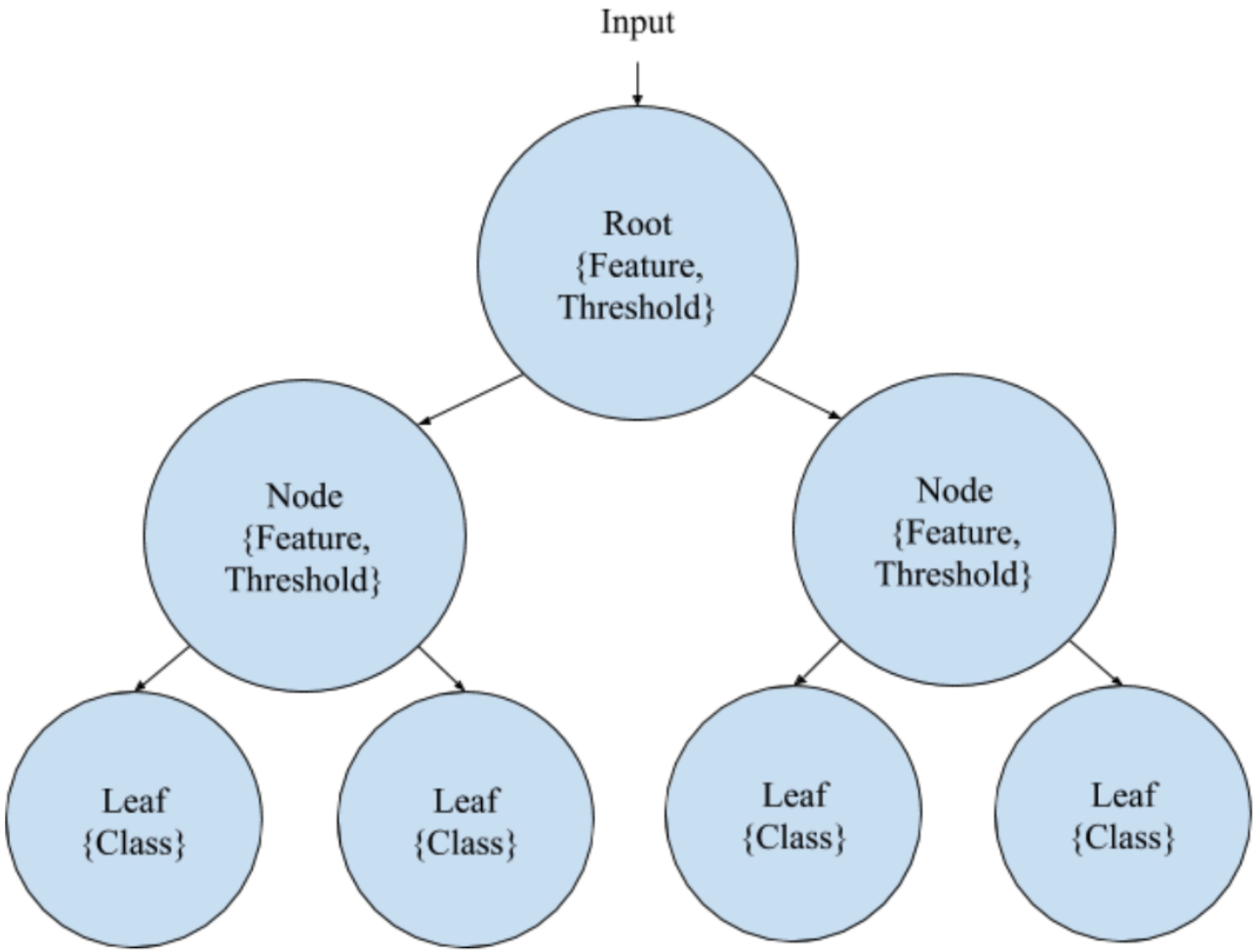
Yale Exended Face Database B

- source: Yale University [2]
-



progress

decision trees



training decision trees

recursive training algorithm:

1. check stopping conditions
   - no more features
   - set is smaller than `minLeaf`
   - all samples in the same class
   - no feature improves information gain (IG)
2. iterate over each available feature, perform a line search to approximate the highest IG
3. recur over the subsets given by splitting at the feature and threshold with the highest IG

$$\mathrm{IG}(X) = \mathrm{H}(X) - \sum_{i=1}^{2} \frac{|S_i|}{|X|} \mathrm{H}(S_i) \tag{1}$$

$$\mathrm{H}(X) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_2 \mathrm{P}(x_i) \tag{2}$$

progress

extra trees

recursive random tree

1. check stopping conditions
   - no more features
   - set is smaller than `minLeaf`
   - all samples in the same class
2. choose random feature. simply use the raw pixels as features.
3. find the mean and variance of this feature across the set. generate a random value from a normal distribution with this mean and variance.
4. recur on the subsets obtained by splitting the parent set on the randomly chosen feature and threshold

ensemble of random trees votes on test data to build extra-tree classifier

progress

progress

results

best performance

algorithm	MNIST	Yale B
LDA	13.5%	6.8%
decision tree	16.6%	57.9%
extra-trees	4.9%	34%

5-fold cross-validated performance

algorithm	MNIST	Yale B
LDA	14.5%	2.5%
decision Tree	17.9%	74.9%
extra-trees	5.4%	36.3%

conclusion

- intuition and heuristics needed for each algorithm
- choice of algorithm depends on features of interest
- future work:
  - alternative distance metrics
  - normalize input data
  - parallelize algorithms

references


Yann LeCun and Corinna Cortes.
MNIST handwritten digit database.
2010.


Yale face database b.

thanks!