import.io: ML Test

April Shen
12 September 2015

Project Euler: Problem 67

Problem Statement

- Find the maximum path sum of a given triangle, where a path moves from the top to bottom and each step moves to adjacent numbers in the next row
- Small example: max sum is 3+7+4+9 = 23

$$3$$
$$7 \quad 4$$
$$2 \quad 4 \quad 6$$
$$8 \quad 5 \quad 9 \quad 3$$

- Actual triangle has 100 rows

Approach

- Brute-force enumeration of paths is too slow
- Fortunately, perfect for dynamic programming
  - Overlapping subproblems
  - Optimal substructure
- Model triangle as DAG
  - (in your head, not necessary explicit in program…)

Algorithm

- Recursion:

max-sum-to(node) = node.value if node is root

max{node.value + max-sum-to(p)} for parents p

Answer = max{max-sum-to(l)} for leaves l

- Compute via DP table

Answer = 7273 for the large triangle

Predict the Missing Grade

Problem Statement

- Given students' grade records for several subjects, predict what grade they received in Mathematics
  - Each student has taken 5 out of 10 possible subjects, and grades are numeric levels between 1 and 8
- Provided with training and test dataset
  - Training: 79, 465 records
 - Test: 69,530 records (and their actual Math grades)

Approach

- Classification problem
  - Features are grades for each subject (or -1 if subject not taken)
  - Classes are grades for Mathematics
- Use AdaBoost with decision stumps as base learners
  - Good performance [Dietterich 2000]
  - Robust to overfitting (without too many rounds)
  - Interpretable model

Results

accuracy vs. number of models
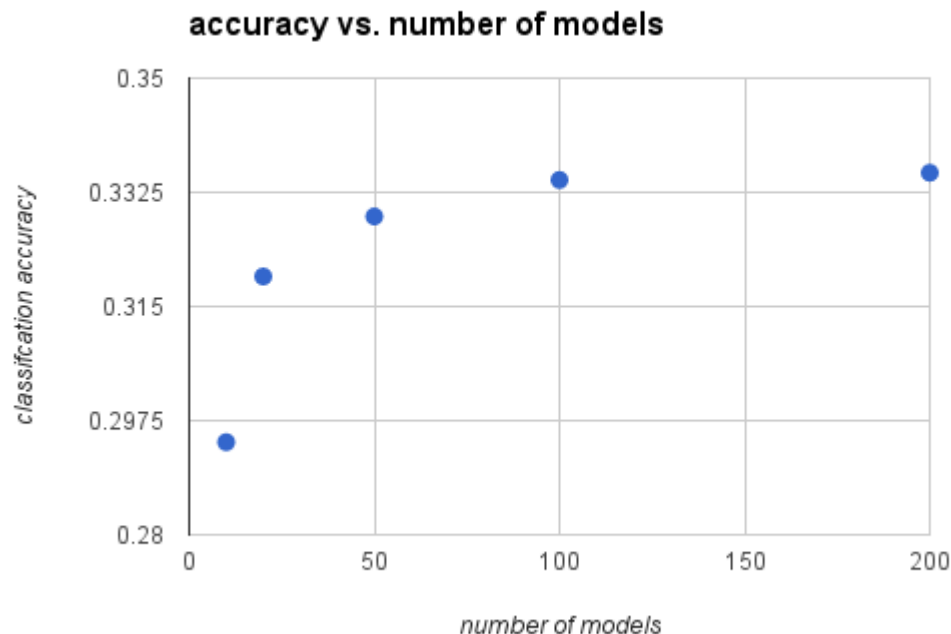
number of models	classifcation accuracy	hackerrank score
10	0.2942	17.09
20	0.3196	28.84
50	0.3288	30.8
100	0.3344	33.25
200	0.3355	32.71

hackerrank score = 100 * ((C-W)/N)
C = # correct predictions
        (<= 1 grade level away from actual)
W = # wrong predictions
N = Total number of data instances

(cf. best score on HackerRank = 45.33,
though on a different test set)

Discussion

- Can see effects of overfitting at 200 models
- Imbalances in training data
  - First 10 models are all rooted with either Physics or Chemistry
  - Besides Math itself (and English), Physics and Chemistry are by far the most frequent subjects appearing in the training data
    - >65,000 / 79,465 students for both
  - Possibly English just isn't as informative…
- Might get better ensemble by enforcing diversity in the models

The Punctuation Corrector

Problem Statement

- Given sentences with either "it's" or "its" removed, determine which should appear in the sentence
- Example:
  - In: "This restaurant is known for ??? emphasis on spicy cooking."
  - Out: "This restaurant is known for its emphasis on spicy cooking."
- Also provided with a corpus of text
- Test set consists of 32 sentences with a total of 36 missing it's/its

Approach

- Similar to problem of lexical substitution
  - Given a marked word in context, choose an alternative word to replace it that preserves the meaning
- … Except the only alternatives are "it's" and "its"
- Use criterion of syntagmatic coherence to rank candidates [Giuliano et al. 2007]

Algorithm

- Idea: given a sentence, search corpus for n-grams containing a candidate substitute (it's/its)
- Prefer longer n-grams, with ties broken in favor of higher frequency
  - E.g., prefer w1 that appears once in a 5-gram to w2 that appears 1000 times in 4-grams.
- Use this to rank candidates and choose the best

Results

hackerrank score = 100 * ((C-W)/N)
C = # correct predictions
W = # wrong predictions
N = Total number of data instances

accuracy	hackerrank score
0.7778	55.56

(cf. best score on HackerRank = 78.46, though on a different test set)

Discussion

- Almost exclusively, the mistakes use "its" instead of "it's", rather than vice versa
  - Possibly "its" appears more in the corpus
  - Example: "It's one thing to engineer a great car but its another to produce it."
- Errors could be fixed by basic knowledge of grammar/syntax
- Note ranking is unbiased as to whether the n-grams precede or follow, but usage of it's/its depends much more on what follows than what precedes

Future Work

(Or things I would do differently if I had more time…)

- More thorough experiments
  - Average performance using 10-fold cross-validation
  - Compare with baseline algorithms
- Optimize code efficiency
- Grade predictor:
  - Try regression rather than classification; since the grade levels are just bins of continuous values, the discreteness is really superficial
- Punctuation corrector:
  - Since "it's" and "its" are different parts of speech, could really benefit from part-of-speech tagging and use of grammar models

References

Dietterich, T. G. (2000). An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, *40*(2), 139-157.

Giuliano, C., Gliozzo, A., & Strapparava, C. (2007, June). Fbk-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 145-148). Association for Computational Linguistics.

Zhu, J., Zou, H., Rosset, S., & Hastie, T. (2009). Multi-class adaboost. *Statistics and its Interface*, *2*(3), 349-360.