Brief Intro to ML and Language Identification

Winston Wu

10/1/2020

Question

- What language is this?

冇問題，我可以去

- What are the giveaways?
  - 冇 is not used in Mandarin
  - 問題 is in traditional script

Language ID

- How does a computer do it?
  - See which words/characters/… are associated with what language
  - We'll implement it today!

- First, how do we get a computer to learn how to do it?

Question

What is learning?

What is learning?

From Mitchell (1997):

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

Task: addition, machine translation, language modeling, ...

Performance measure: accuracy, F-score, BLEU, perplexity, ...

Experience: data

Learning: An Analogy

You know nothing about math.

The instructor gives you a list of example problems: data

```
2 + 5 = 7
3 - 4 = -1
1 + 2 = 3
...
```

And says you will be tested on similar problems later. domain

How do you study?

Study Strategies

The consistent (lazy) student

- You glance at the problems, but you just can’t be bothered to learn simple addition and subtraction.
- You decide you will write ‘5’ as the answer to every question.
- You’re bound to get at least some correct, right?

You did not learn from the data. This is called underfitting.

Study Strategies

The memorizer

- You memorize the answer to every practice problem.
- You've seen "2 + 5" so you know the answer.
- The test has "2 + 6". You've never seen this problem, so you don't know how to solve it.

You do well on problems you have seen but not on problems you haven't seen. This is called overfitting.

You don't generalize (apply what you've learned to new data).

Study Strategies

The generalizing learner

development set
validation set

- You set aside some of the example problems as a pretest.
- You work through the practice problems, then test yourself with the pretest to see how you're doing.

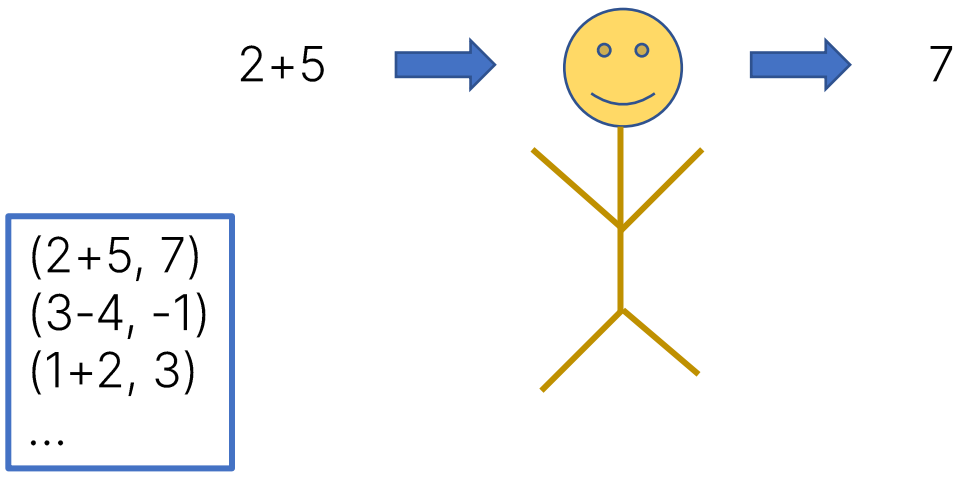
training set

- If you didn't do well, work through the practice problems again, then test yourself again.
  - You could shuffle the problems around

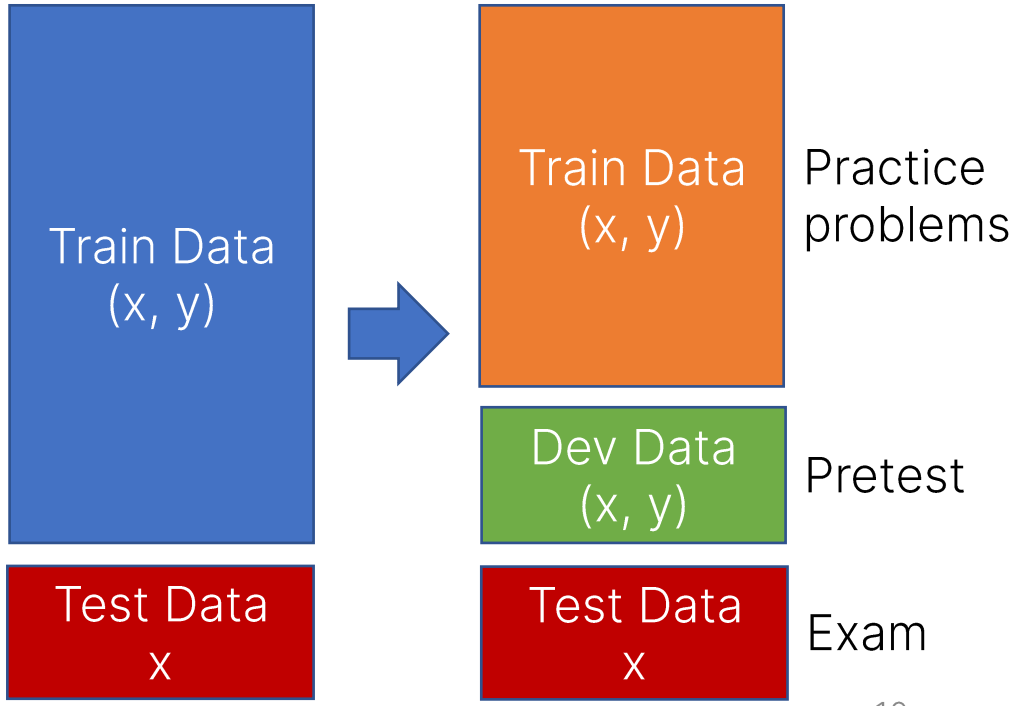
This is the typical setup in supervised machine learning.

Supervised Learning

- Data: (x, y) pairs
- Goal: predict Y from X
- Model: f(x)



- Data Splitting
  - 70/15/15 or 80/10/10 are common
  - Only see the test data once!



Evaluation Metrics

- How well you do on the exam?
- Accuracy: how many did you get right?
  - That's kinda harsh
  - What about partial credit?
    - Regression
    - Machine translation

- What does the metric tell you?
  - Model gets 98% accuracy! Is it a good model?
  - Model:

```
function is_it_spam(email)
    return false
end
```

Precision and Recall

- Data: `NNNNN NNN SS`
- Your model: `NNNNN SSS SN`

Red is incorrect prediction

- Precision (aka positive predictive value)
  - Out of all the ones you labeled as spam, how many actually are? 1/4
- Recall (aka sensitivity, true positive rate, detection rate)
  - Out of all the spam messages, how many did you label as spam? 1/2
- F-score = 2 * P * R / (P + R)
- There's usually a tradeoff

Language Identification

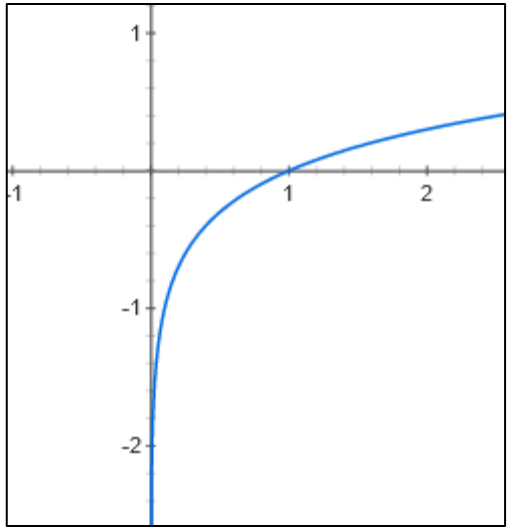
- Scenario: we are given text from an unknown language

有問題，我可以去

- Task: identify the language!
- Data: (sentence, language) pairs
- Metric: accuracy
- Model?
  - Hint: use what we learned from language models last week

Dealing with Underflow

- Underflow: 0.00000000001 * 0.0000000001 = 0
  - Multiplying a lot of small probabilities can eventually become zero!
  - Solution: do calculations in log space



$$\prod x_i = \exp\left(\sum \ln(x_i)\right)$$

$$0.3 * 0.4 = e^{\ln(0.3)+\ln(0.4)} = 0.12$$

Softmax

- We usually normalize probabilities like this:

```
function normalize(seq)
    total = sum(seq)
    return seq ./ total
end
```

$$\frac{x_i}{\sum_i x_i}$$

- But what if we use log probabilities?

```
function normalize(seq)
    denom = sum(exp.(seq))
    return exp.(seq) ./ denom
end
```

$$\frac{e^{x_i}}{\sum_i e^{x_i}}$$

Language Identification

- Let’s go try it out!

Survey Questions

1. What is the point of having a development/validation set?
2. What kind of features do you think would be good for spam classification?
3. Questions, comments, concerns, suggestions?