Recap

LING572
Advanced Statistical Methods for NLP
January 23, 2020

Outline

- Summary of the material so far
- Reading materials
- Math formulas

So far

- Introduction:
  - Course overview
  - Information theory
  - Overview of classification task

- Basic classification algorithms:
  - Decision tree
  - Naïve Bayes
  - kNN

- Feature selection, chi-square test and recap
- Hw1-Hw3

Main steps for solving a classification task

- Prepare the data:
  - Reformulate the task into a learning problem
  - Define features
  - Feature selection
  - Form feature vectors
- Train a classifier with the training data
- Run the classifier on the test data
- Evaluation

Comparison of 3 Learners

	kNN	Decision Tree	Naïve Bayes
Modeling	Vote by your neighbors	Vote by your groups	Choose the c that max P(c | x)
Training	None	Build a decision tree	Learn P(c) and P(f | c)
Decoding	Find neighbors	Traverse the tree	Calculate P(c)P(x | c)
Hyper parameters	K Similarity fn	Max depth Split function Thresholds	Delta for smoothing

Implementation issues

- Taking the log:

$$\log(P(c)\prod_i P(f_i\,|\,c)) = \log P(c) + \sum_i \log P(f_i\,|\,c)$$

- Ignoring some constants:

$$P(d_i\,|\,c) = P(\,|\,d_i\,|\,)\,|\,d_i\,|\,!\prod_{k=1}^{|V|}\frac{P(w_k\,|\,c)^{N_{ik}}}{N_{ik}!}$$

- Increasing small numbers before dividing

$$\log P(x, c_1) = -200;\ \log P(x, c_2) = -201$$

Implementation issues (cont)

- Reformulate the formulas:

$$P(d_i, c) = P(c) \prod_{w_k \in d_i} P(w_k \mid c) \prod_{w_k \notin d_i} (1 - P(w_k \mid c))$$

$$= P(c) \prod_{w_k \in d_i} \frac{P(w_k \mid c)}{1 - P(w_k \mid c)} \prod_{w_k} (1 - P(w_k \mid c))$$

- Store useful intermediate results: $\prod_{w_k} 1 - P(w_k \mid c)$

- Vectorize! (e.g. entropy)

Lessons learned

- Don’t follow the formulas blindly. Vectorize when possible.
  - Ex1: Multinomial NB

$$P(c)\prod_{k=1}^{|V|}P(w_k\,|\,c)^{N_{ik}}$$

  - Ex2: cosine function for kNN

$$\cos(d_i, d_j) = \frac{\sum_k d_{i,k}d_{j,k}}{\sqrt{\sum_k d_{i,k}^2}\sqrt{\sum_k a_{j,k}^2}}$$

Next

- Next unit (2.5 weeks): two more advanced methods:
  - MaxEnt (aka multinomial logistic regression)
  - CRF (Conditional Random Fields)
- Focus:
  - Main intuition, final formulas used for training and testing
  - Mathematical foundation
  - Implementation issues

Reading material

The purpose of having reading material

- Something to rely on besides the slides
- Reading before class could be beneficial
- Papers (not textbooks; some blog posts) could be the main source of information in the future

Problems with the reading material

- The authors assume that you know the algorithm already:
  - Little background info
  - Page limit
  - Style

- The notation problem

➔ It could take a long time to understand everything

Some tips

- Look at several papers and slides at the same time
  - Skim through the papers first to get the main idea
  - Go to class and understand the slides
  - Then go back to the papers (if you have time)

- Focus on the main ideas. It’s ok if you don’t understand all the details in the paper.

Math formulas

The goal of LING572

- Understand ML algorithms
  - The core of the algorithms
  - Implementation: e.g., efficiency issues

- Learn how to use the algorithms:
  - Reformulate a task into a learning problem
  - Select features
  - Write pre- and post-processing modules

Understanding ML methods

- 1: never heard about it
- 2: know very little
- 3: know the basics
- 4: understand the algorithm (modeling, training, testing)
- 5: have implemented the algorithm
- 6: know how to modify/extend the algorithm

➔ Our goal: kNN, DT, NB: 5

MaxEnt, CRF, SVM, NN: 3-4

Math is important for 4-6, especially for 6.

Why are math formulas hard?

- Notation, notation, notation.
  - Same meaning, different notation: f_k, w_k, t_k
- Calculus, probability, statistics, optimization theory, linear programming, …
- People often have typos in their formulas.
- A lot of formulas to digest in a short period of time.

Some tips

- No need to memorize the formulas

- Determine which part of the formulas matters

$$P(d_i \mid c) = P(\mid d_i \mid) \mid d_i \mid ! \prod_{k=1}^{\mid V \mid} \frac{P(w_k \mid c)^{N_{ik}}}{N_{ik}!}$$

$$\textsf{classify}(d_i) = \arg\max_c P(c) P(d_i \mid c)$$

$$\textsf{classify}(d_i) = \arg\max_c P(c) \prod_{k=1}^{\mid V \mid} P(w_k \mid c)^{N_{ik}}$$

- It's normal if you do not understand it the 1st/2nd time around.

Understanding a formula

$$P(w_t \mid c_j) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j \mid d_i)}$$

$$P(w_t \mid c_j) = \frac{\sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j \mid d_i)}$$

$$= \frac{\sum_{i=1}^{|D|} N_{it} P(c_j \mid d_i)}{Z(c_j)}$$

$$= \frac{\sum_{d_i \in D(c_j)} N_{it}}{Z(c_j)}$$

Next Week

- On to MaxEnt! Don’t forget: reading assignment due Tuesday at 11AM!