

Recap

LING572

Fei Xia

Outline

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

So far

- Introduction (0.5 week):
 - Course overview
 - Information theory
 - Overview of classification task
- Basic classification algorithms (2 weeks):
 - Decision tree
 - Naïve Bayes
 - kNN
- Feature selection and chi-square test (1 week)
- Hw1-Hw4

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 three learners

	kNN	Decision Tree	Naïve Bayes
Modeling	Vote by your neighbors	Vote by your groups	Choose the c that $\max P(c x)$
Training	No	Build a decision tree	Learn $P(c)$ and $P(f c)$
Decoding	Find neighbors	Traverse the tree	Calculate $P(c)P(x c)$
Tuned parameters	K Similarity func	Max depth Split function Thresholds	delta for smoothing

Implementation issue

- Take the log:

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

- Ignore some constants:

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

- Increase small numbers before dividing

$$\log P(x, c_1) \text{ is } -200, \log P(x, c_2) \text{ is } -201.$$

Implementation issue (cont)

- Reformulate the formulas: e.g., entropy calc

$$P(d_i, c)$$

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

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

- Store the useful intermediate results

$$\prod_{w_k} (1 - P(w_k | c))$$

Lessons learned

- Don't follow the formulas blindly.
 - Ex1: Multinomial NB

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

- Ex2: cosine function for kNN

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

Next

- Next unit (2.5 weeks): two more advanced methods:
 - MaxEnt
 - CRF
- 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) could be the main source of information in the future

Problems with the reading material

- The authors assume that you have known 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 is ok if you don't understand all the detail 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: have 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, TBL: 3-4

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

Why are math formulas hard?

- Notation, notation, notation.
 - Same meaning, different notations: 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

$$P(d_i|c_j) = P(|d_i|)|d_i|! \prod_{t=1}^{|V|} \frac{P(w_t|c_j)^{N_{it}}}{N_{it}!}$$

$$\text{classify}(d_i) = \operatorname{argmax}_c P(c) \prod_{k=1}^{|V|} P(w_k|c)^{N_{ik}}$$

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

Understanding a formula

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

$$\begin{aligned} P(w_t | c_j) &= \frac{\sum_{i=1}^{|D|} N_{it} P(c_j | d_i)}{\sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j | d_i)} \\ &= \frac{\sum_{i=1}^{|D|} N_{it} P(c_j | d_i)}{Z(c_j)} \\ &= \frac{\sum_{d_i \in D(c_j)} N_{it}}{Z(c_j)} \end{aligned}$$