Machine Learning for Language Processing (L101)

Ann Copestake

Computer Laboratory
University of Cambridge

October 2016

Outline of today's lecture

Introduction to the course

Machine learning in NLP

Text classification

Naive Bayes for text classification

About this course

- An introduction to using Machine Learning (ML) in NLP
- part of the NLP ‘theme’, NOT a general introduction to ML
- Prerequisites: L90 (or similar) and L95
- Next term: Advanced Topics in Natural Language Processing (R222)
- Other courses with substantial ML content: Principles of Data Science (L120); Probabilistic Machine Learning (E4F13); Biomedical Information Processing (R214) (next term); Machine Learning and Algorithms for Data Mining (L42) (next term); plus Computer Vision etc

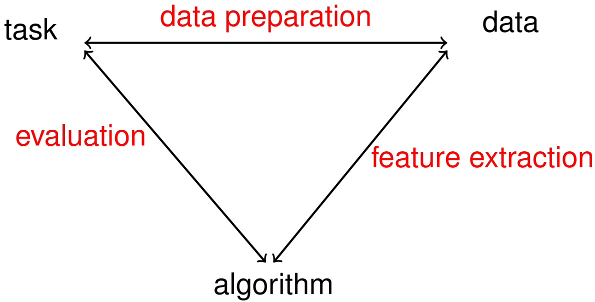
An introduction to using Machine Learning (ML) in NLP

- Lectures
- Seminars (‘ticked’ assessment: 10%)
- Essay/mini-project (main assessment: 90%)

Sources of information

- Course web pages.
- Stephen Clark's lecture notes / slides from last year.
- Textbooks, see syllabus page:
  NB draft/partial third edition of Jurafsky and Martin
  `https://web.stanford.edu/~jurafsky/slp3/`
- L90 notes: overview of NLP.
- Research Skills: November 21 14:00–15:00
  please try and attend.

Machine learning, abstractly



Task

- Usually an abstraction from a real problem, or a piece of a (possible) larger architecture.
- End-user systems vs experimental systems.
- Most research publications concern standard tasks: sentiment classification of movie reviews, document classification, POS tagging etc, etc.

Data

- Used to train and test the ML system:
  - Train
  - Test: no overlap with training data, ideally unseen by experimenter.
  - Development (maybe)
- Supervision:
  - Supervised: training data labelled with desired outcome
  - Unsupervised
  - Semi-supervised, moderately supervised etc
- Annotation:
  - Manually annotated (expert vs crowd-sourced)
  - 'found' annotation (e.g., star ratings for move reviews)

Data acquisition

For NLP, usually a text corpus, possibly with additional material (parallel text, images, etc).

- Realistic data for task?
- Where from?
- How much is needed?
- Annotation?
- Annotation for evaluation?

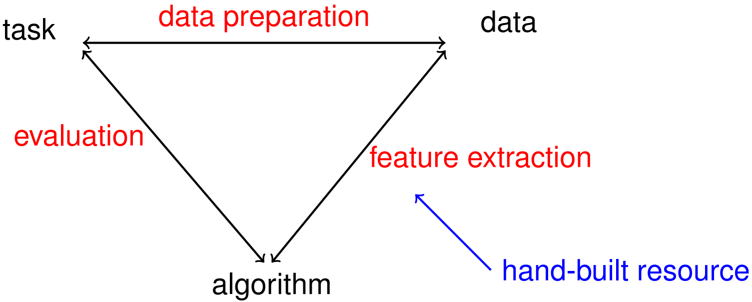
Algorithm

- Choice of algorithm for task, given available data and features.
- Training time and running time.
- Fast and dumb often better than slow and sophisticated.
- Robustness, consistency.

Features

- Information extracted from data.
- e.g., individual words from the text (using spaces as boundaries): bag of words
- automatically annotated with part of speech tags, syntactic dependencies ...
- additional data sources: e.g., WordNet, Wikipedia
- complex systems may involve several layers of annotation
- feature selection in ML algorithm

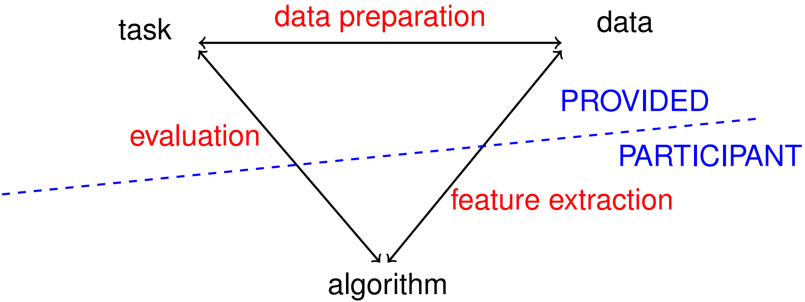
Machine learning, variants



Evaluation

- Different metrics/approaches for different problems
  - Human evaluation (possibly using crowdsourcing)
  - Standardized test sets / metrics
- Notion of 'best' performance depends on details of task: e.g., spam filtering — can the user see the mail marked as spam?
- Sensible choice of baseline: don't fool yourself about your system . . .
- Significance testing: almost never done correctly in NLP!

Standard/shared tasks



Different ML paradigms for NP

- Classification (predefined categories):
  - document → category
  - word sequence → category sequence (e.g., POS tagging)
  - word pairs → binary values (anaphora resolution, see L90)
- Clustering (no predefined categories):
  e.g., document set → document groups
- word sequence → word sequence
  - Statistical MT
  - Sentence compression etc (regeneration)
- word sequence → structured representation
- word sequence (or speech) → action
- structured representation → word sequence

Some types of document classification

- Topic:
  - source: ad hoc (e.g., email) vs conventionalized (e.g., library categories)
  - organisation of classes: flat vs hierarchical
  - class membership: one-of vs any-of
- Sentiment: positive, negative or neutral.
  Whole texts or text fragments.
- Spam, adult content (safe search), etc: binary (possibly with score).

What's the topic? English Wikipedia sentences.

Document A:
Thus, what started as an effort to translate between languages evolved into an entire discipline devoted to understanding how to represent and process natural languages using computers.

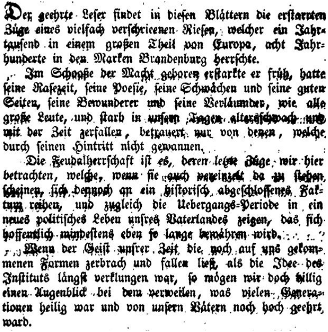
Document B:
An extreme example is the alien species, the Vulcans, who had a violent past but learned to control their emotions.

What’s the topic? German Wikipedia sentences

Document A:
Es umschließt die Mündungen des Hudson River und des East River in den Atlantischen Ozean und erhebt sich durchschnittlich sechs Meter über den Meeresspiegel.

Document B:
Ein weiteres Vorbild war der britische Aufklärungsoffizier, Vogelkundler und Hochstapler Richard Meinertzhagen. Schließlich bediente sich Ian Fleming auch der Geschichten und des Charakters des serbischen Doppelagenten Duško Popov aus dem Zweiten Weltkrieg.

What’s the topic?


Der geehrte Leser findet in diesen Blättern die erstarrten Züge eines vielfach verschrieenen Riesen, welcher ein Jahrtausend in einem großen Theil von Europa, acht Jahrhunderte in den Marken Brandenburg herrschte.

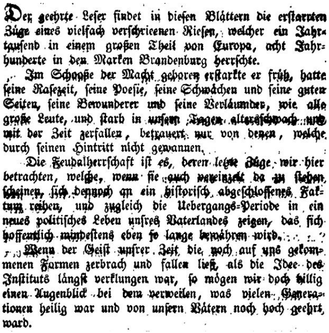
Im Schooße der Macht geboren erstarkte er früh, hatte seine Rasezeit, seine Poesie, seine Schwächen und seine guten Seiten, seine Bewunderer und seine Verläumder, wie alle große Leute, und starb in unsern Tagen altersschwach und mit der Zeit zerfallen, betrauert nur von denen, welche durch seinen Hintritt nicht gewannen.

Die Feudalherrschaft ist es, deren letzte Züge wir hier betrachten, welche, wenn sie auch vereinzelt da zu stehen scheinen, sich dennoch an ein historisch abgeschlossenes Faktum reihen, und zugleich die Uebergangs-Periode in ein neues politisches Leben unsres Vaterlandes zeigen, das sich hoffentlich mindestens eben so lange bewähren wird.

Wenn der Geist unsrer Zeit die noch auf uns gekommenen Formen zerbrach und fallen ließ, als die Idee des Instituts längst verklungen war, so mögen wir doch billig einen Augenblick bei dem verweilen, was vielen Generationen heilig war und von unsern Vätern noch hoch geehrt ward.

Bitcoin?

What's the topic?



Der geehrte Leser findet in diesen Blättern die erstarrten Züge eines vielfach verschrieenen Riesen, welcher ein Jahrtausend in einem großen Theil von Europa, acht Jahrhunderte in den Marken Brandenburg herrschte.

Im Schooße der Macht geboren erstarkte er früh, hatte seine Rasezeit, seine Poesie, seine Schwächen und seine guten Seiten, seine Bewunderer und seine Verläumder, wie alle große Leute, und starb in unsern Tagen altersschwach und mit der Zeit zerfallen, betrauert nur von denen, welche durch seinen Hintritt nicht gewannen.

Die Feudalherrschaft ist es, deren letzte Züge wir hier betrachten, welche, wenn sie auch vereinzelt da zu stehen scheinen, sich dennoch an ein historisch abgeschlossenes Faktum reihen, und zugleich die Uebergangs-Periode in ein neues politisches Leben unsres Vaterlandes zeigen, das sich hoffentlich mindestens eben so lange bewähren wird.

Wenn der Geist unsrer Zeit die noch auf uns gekommenen Formen zerbrach und fallen ließ, als die Idee des Instituts längst verklungen war, so mögen wir doch billig einen Augenblick bei dem verweilen, was vielen Generationen heilig war und von unsern Vätern noch hoch geehrt ward.

Bitcoin?

Observations

- Full text understanding isn't always/usually necessary for classification.
- Individual words can be very good cues, especially when classes are very different.
- Some words are more useful than others (class titles especially!)
- Metadata etc: but ignore that here.

Statistical classification

Choose most probable class from set of classes C given a feature vector $\vec{f}$ ($\hat{c}$ means “estimate of c”):

$$\hat{c} = \operatorname*{argmax}_{c \in C} P(c|\vec{f})$$

Apply Bayes Theorem:

$$P(c|\vec{f}) = \frac{P(\vec{f}|c)P(c)}{P(\vec{f})}$$

Constant denominator:

$$\hat{c} = \operatorname*{argmax}_{c \in C} P(\vec{f}|c)P(c)$$

Naive Bayes Classifier

Rather than considering all the features together:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, P(\vec{f}|c)P(c)$$

We make the ('naive') independent feature assumption:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, P(c) \prod_{i=1}^{n} P(f_i|c)$$

NB with binary-valued word features (Bernoulli)

- Vocabulary is a list of all words in the documents (excluding any in a stop list).
- Feature vector $\vec{f}$ for document d: for each item w_i in the vocabulary, generate 1 if w_i is in d, 0 otherwise.
- Estimate $P(f_i|c)$ as the fraction of documents of class c that contain w_i.
- Estimate $P(c)$ as the proportion of documents which have class c.
- Alternatively, Multinomial Naive Bayes: uses frequencies of words in documents.

Some properties of Naive Bayes

- The independence assumption is wrong!
  e.g., consider 'Hong' and 'Kong'.
- Very bad probability estimates but is a reasonably good (and robust) classifier.
- Optimal in terms of efficiency (linear).
- A good baseline for classification experiments.
- Usually: multinomial NB better for topic classification (especially for long documents), binary-valued better for sentiment analysis.

Generative models

- NB is a generative model: we train a model of the joint distribution of observations and classes, $P(\vec{f}, c)$.
- Hence, for multinomial NB, this is equivalent to a unigram model.
- Contrast discriminative models, where we train the posterior distribution of the class given the observation $P(c|\vec{f})$
- Also: discriminant functions — we just train a mapping from the observation to the class label without the probability.

Further reading

- Much more in the readings for the seminars: see course web page.
- Chapter 7 of `https://web.stanford.edu/~jurafsky/slp3/`
- Chapter 13 of Manning et al.

Next time

- Monday, October 10, 15:00
- POS tagging
- READ the notes for Lecture 3 from L90 before the lecture