INF5830 – 2013 FALL
NATURAL LANGUAGE PROCESSING

Jan Tore Lønning, Lecture 6, 24.9

Today & Thursday

- Word Sense Disambiguation (WSD)
- Classification
- Naive Bayes
- NB applied to WSD
  - Two approaches
- Evaluation
- Feature selection
- Smoothing

Word Sense Disambiguation

Word Senses

Word Net:

Noun

- S: (n) bass (the lowest part of the musical range)
- S: (n) bass, bass part (the lowest part in polyphonic music)
- S: (n) bass, basso (an adult male singer with the lowest voice)
- S: (n) sea bass, bass (the lean flesh of a saltwater fish of the family Serranidae)
- S: (n) freshwater bass, bass (any of various North American freshwater fish with lean flesh (especially of the genus Micropterus))
- S: (n) bass, bass voice, basso (the lowest adult male singing voice)
- S: (n) bass (the member with the lowest range of a family of musical instruments)
- S: (n) bass (nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes)

Word Sense Disambiguation

Why?

- Dialogue system: fish vs. musical instrument
- Translation: Han bestilte en salt-bakt bass.
- Search: relevant hits only
- Speech synthesis: (Norw.) tapet, urte

How?

- Corpus-based methods
  - Unsupervised – clustering
  - Supervised – classification
- Dictionaries and Thesauri

Corpus-based

- Professor Emeritus Dag Østerberg doserer i klassisk byteori , og Knut Halvorsen fra Oslo Teknopol i synergieffektene mellom byer og forskningsmiljøer .
- Her sitter jeg på en bar , tenkte Harry , og hører på en transvestitt som doserer om australsk politikk .
- Han doserer teen etter tempoet i fortellingen hennes , ser ikke ned på den for å forvisse seg om at koppen er på rett kjøl , ser ufravendt på henne , henger ved hennes lepper - teen kommer av seg selv .
- Som doserer morfinen og holder en knivskarp balanse mellom bevissthet og smerteterskel .

- How to disambiguate?
- Use context

7

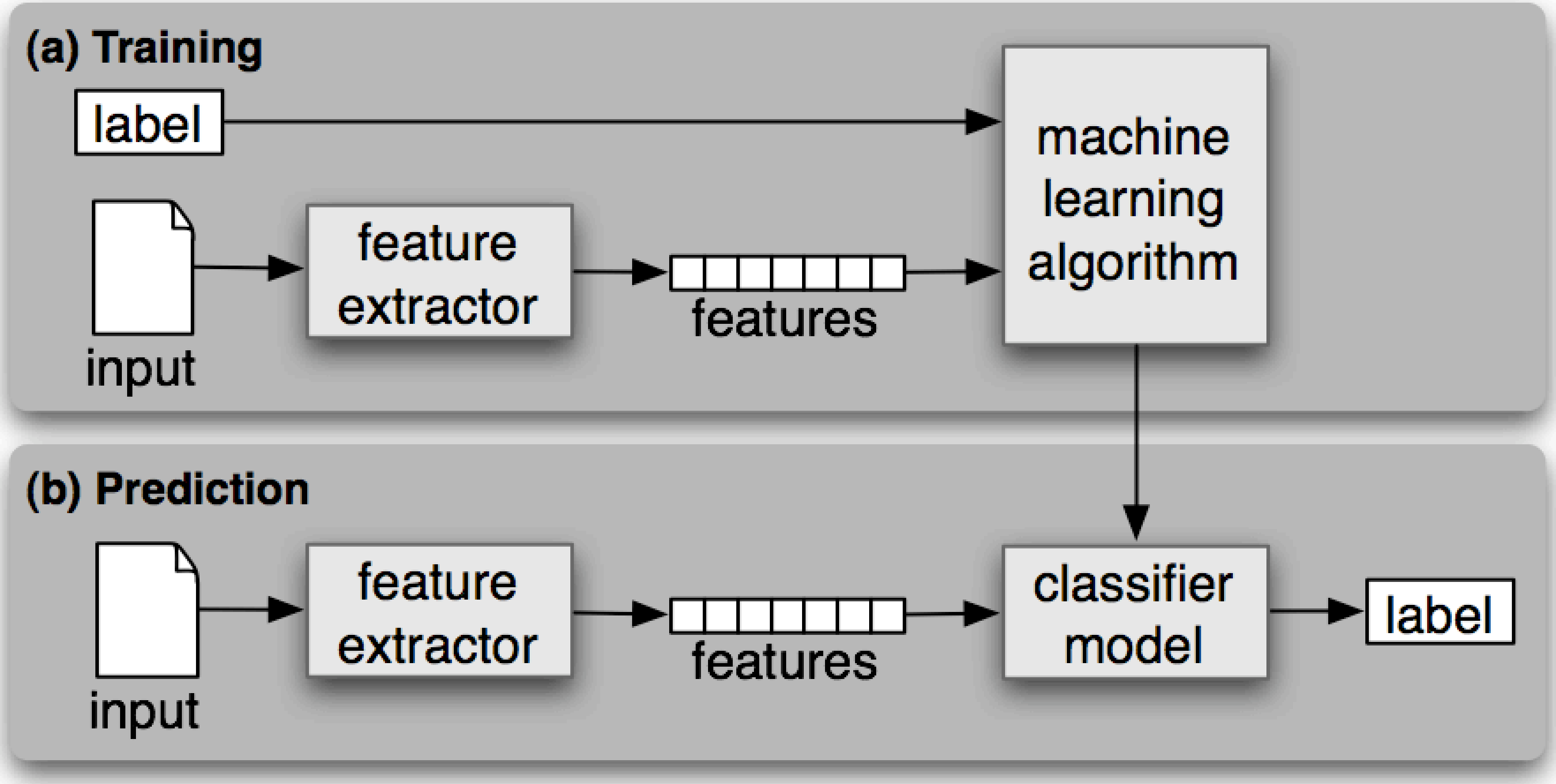
Classification

Classification

- Supervised:
  - Given classes
  - Examples
- Unsupervised:
  - Construct classes

- Supervised examples:
  - Spam
  - Word sense disambiguation
  - Genre of text
  - Language classification

Classification



A variety of classifiers

- k-Nearest Neighbors
- Rocchio
- Decision Trees
- Naive Bayes
- Maximum entropy (Logistic regression)
- Support Vector Machines

INF4820

Supervised classification

- Given a well-defined set of objects, O
- A given set of classes, $S=\{s_1, s_2, \ldots, s_k\}$
- For training: a set of pairs from OxS
- Goal: a classifier, γ, a mapping from O to S

Task	O	S
Word sense disambiguation	Occurrences of "bass"	Sense1, …, sense8
Spam classification	E-mails	Spam, no-spam
Language clssification	Pieces of text	Arabian, Chinese, English, Norwegian, …

Features

- To represent the objects in O, extract a set of features

Object: person
Features:
- height
- weight
- hair color
- eye color
- …

Object: email
Features:
- length
- sender
- contained words
- language
- …

- Be explicit:
  - Which features
  - The value space for each feature

Supervised classification

- A given set of classes, $S=\{s_1, s_2, \ldots, s_k\}$
- A well defined class of objects, O

- Some features $f_1, f_2, \ldots, f_n$
- For each feature: a set of possible values $V_1, V_2, \ldots, V_n$
- The set of feature vectors: $V= V_1 \times V_2 \times \ldots \times V_n$
- Each object in O is represented by some member of V:
  - Written $(v_1, v_2, \ldots, v_n)$, or
  - $(f_1=v_1, f_2=v_2, \ldots, f_n=v_n)$
- A classifier, γ, can be considered a mapping from V to S

Examples

Language classifier

- C = {English, Norwegian,…}
- O is the set of strings of letters
- f_1 is last letter of o
- V_1= {a, b, c,…, å}
- f_2 is the last two letters
- V_2 are all two letter combinations
- f_3 is the length of o,
- V_3 is 1, 2, 3, 4, …

Word sense disambiguation

- C = {fish, music}
- O: all occurrences of "bass"
- $f_i = f_{wi}$: word w_i occurs in same sentence as "bass", where
  - w_1 = *fishing*, w_2 = *big*, …,
  - w_{11} = *guitar*, w_{12} = *band*
- $V_1=V_2=\ldots=V_{12}=\{1,0\}$
- Example:
  - o = (0,0,0,1,0,0,0,0,0,0,1,0)
  - o = (f_{fish}=1, …, f_{guitar}=1, f_{band}=0)

Naive Bayes

Naive Bayes: Decision

- Given an object
  - $\langle f_1 = v_1, f_2 = v_2, \dots, f_n = v_n \rangle$
- Consider
  - $P\left(s_m \mid \langle f_1 = v_1, f_2 = v_2, \dots, f_n = v_n \rangle\right)$ for each class sm
- Choose the class with the largest value, in symbols

$$\underset{s_m \in S}{\arg\max}\, P\left(s_m \mid \langle f_1 = v_1, f_2 = v_2, \dots, f_n = v_n \rangle\right)$$

Naive Bayes: Model

- Bayes formula
  - $$P\left(s_m \mid \left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle\right) = \frac{P\left(\left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle \mid s_m\right) P(s_m)}{P\left(\left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle\right)}$$
- Sparse data, we may not even have seen
  - $$\left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle$$
- Assume (wrongly) independence
  - $$P\left(\left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle \mid s_m\right) \approx \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right)$$
- Putting together
  - $$\arg\max_{s_m \in S} P\left(s_m \mid \left\langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \right\rangle\right) \approx \arg\max_{s_m \in S} P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right)$$

Naive Bayes: Calculation

- $$\arg\max_{s_m \in S} P\left(s_m \mid \langle f_1 = v_1, f_2 = v_2, \dots, f_n = v_n \rangle\right) \approx \arg\max_{s_m \in S} P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right)$$

- For calculations
  - avoid underflow, use logarithms
  - $$\arg\max_{s_m \in S} P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right) = \arg\max_{s_m \in S} \left(\log \left(P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right) \right) \right)$$

Naive Bayes: Calculation

- $$\arg\max_{s_m \in S} P\left(s_m \mid \langle f_1 = v_1, f_2 = v_2, \ldots, f_n = v_n \rangle\right) \approx \arg\max_{s_m \in S} P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right)$$

- For calculations
  - avoid underflow, use logarithms
  - $$\arg\max_{s_m \in S} P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right) = \arg\max_{s_m \in S} \left(\log \left(P(s_m) \prod_{i=1}^{n} P\left(f_i = v_i \mid s_m\right) \right) \right)$$

$$= \arg\max_{s_m \in S} \left(\log(P(s_m)) + \sum_{i=1}^{n} \log(P\left(f_i = v_i \mid s_m\right)) \right)$$

Naive Bayes: Training

- Maximum Likelihood
- $$\hat{P}(s_m) = \frac{C(s_m, o)}{C(o)}$$
  - where $C(s_m, o)$ are the number of occurrences of objects o in class s_m
- $$\hat{P}(f_i = v_i \mid s_m) = \frac{C(f_i = v_i, s_m)}{C(s_m)}$$
  - where $C(f_i=v_i, s_m)$ are the number of occurrences of objects o
    - where the object o belongs to class s_m
    - and the feature f_i takes the value v_i
  - $C(s_m)$ are the number of occurrences belonging to class s_m
- Observe: this is different from FSNLP
- + smoothing: Thursday

Data