NIHR | Maudsley Biomedical Research Centre

Supervised learning: classification and sequence learning

Angus Roberts

Adapted from Sumithra Velupillai

with acknowledgements to Genevieve Gorrell, University of Sheffield

CLINICAL RECORD INTERACTIVE SEARCH

What we will cover

- Introduction to supervised machine learning in the context of NLP
  - Basic concepts
  - Common algorithms
    - We will NOT discuss all algorithms and underlying theories in detail
  - Common tasks

Supervised machine learning

- Supervised ML algorithms
  - learn to map an output variable given some input data
- Supervised
  - the algorithm learns from *existing* output variables/labels
  - iteratively makes predictions
    - corrected according to some criteria until no improvements

Classification tasks in NLP

- Sentiment
- Spam
- Document category
- Language
- Phenotypes
- Patient status
- Entities
- …

Terminology

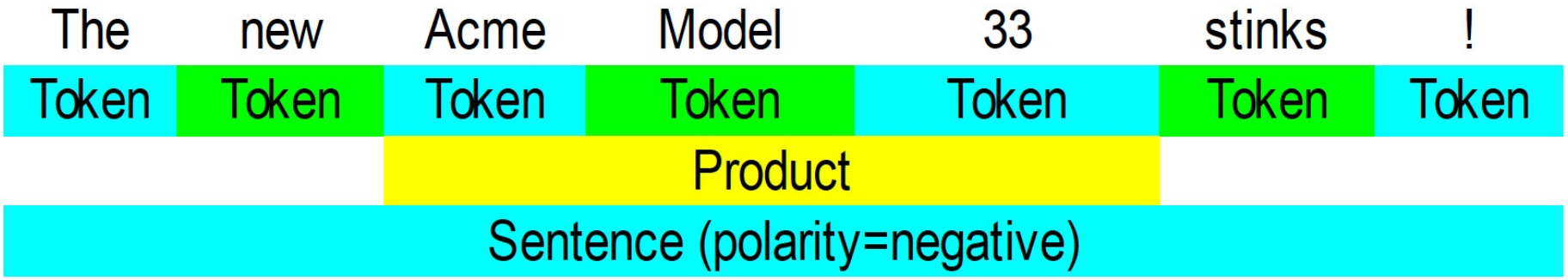
- Instances
- Attributes, or features
- Classes, or labels

Instances

- Cases that can be learned
- Every instance is mapped to an output variable

Instances

- Cases that can be learned
- Every instance is mapped to an output variable



Instances

- Cases that can be learned
- Every instance is mapped to an output variable

Table 1

Table 1 Annotator Training Samples

No.	Sample Sentences	Smoking Status (based on text)
1	She is a past smoker, but quit two years ago when she was found to have right upper lobe nodule, which was resected and found to be positive for TB granuloma, for which she was treated with antibiotics for nine months.	Past Smoker
2	She quit smoking four months ago.	Current Smoker
3	Depression, anxiety, chronic obstructive pulmonary disease/asthma, history of tobacco abuse, chronic headaches, atypical chest pain with 6/97 Dobutamine MIBI revealing no ischemia and a history of tuberculosis exposure.	Smoker
4	No tobacco.	Non-Smoker
5	Most recently, she developed dyspnea two days prior to admission, trigger was felt to be marijuana smoke in the building where she lives where there are many drug dealers.	Unknown

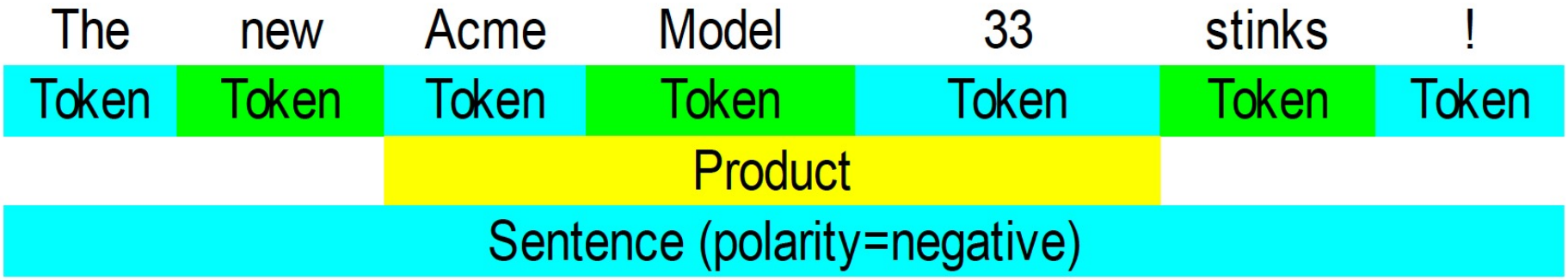
Uzuner O, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records.
J Am Med Inform Assoc. 2008;15(1):14–24. doi:10.1197/jamia.M2408

Attributes/features

- Pieces of information about an instance

Attributes/features

- Pieces of information about an instance



Attributes/features - representations

- Bag-of-words

	text	takes	with	in	today	pain	past	free	the	for	she	has	aspirin	had	patient	abdominal	taken	no
0	patient with abdominal pain. she has taken aspirin.	0	1	0	0	1	0	0	0	0	1	1	1	0	1	1	1	0
1	she has had abdominal pain in the past. pain free today.	0	0	1	1	2	1	1	1	0	1	1	0	1	0	1	0	0
2	no abdominal pain.	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	1
3	takes aspirin for pain. has no pain today.	1	0	0	1	2	0	0	0	1	0	1	1	0	0	0	0	1

Attributes/features - representations

- Bag-of-words – no stopwords

	text	takes	today	pain	past	free	patient	abdominal	taken	aspirin
0	patient with abdominal pain. she has taken aspirin.	0	0	1	0	0	1	1	1	1
1	she has had abdominal pain in the past. pain free today.	0	1	2	1	1	0	1	0	0
2	no abdominal pain.	0	0	1	0	0	0	1	0	0
3	takes aspirin for pain. has no pain today.	1	1	2	0	0	0	0	0	1

Attributes/features - representations

- Tfidf

	text	takes	today	pain	past	free	patient	abdominal	taken	aspirin
0	patient with abdominal pain. she has taken aspirin.	0.000	0.000	0.0	0.000	0.000	0.173	0.036	0.173	0.087
1	she has had abdominal pain in the past. pain free today.	0.000	0.063	0.0	0.126	0.126	0.000	0.026	0.000	0.000
2	no abdominal pain.	0.000	0.000	0.0	0.000	0.000	0.000	0.096	0.000	0.000
3	takes aspirin for pain. has no pain today.	0.173	0.087	0.0	0.000	0.000	0.000	0.000	0.000	0.087

Attributes/features - representations

- Embeddings

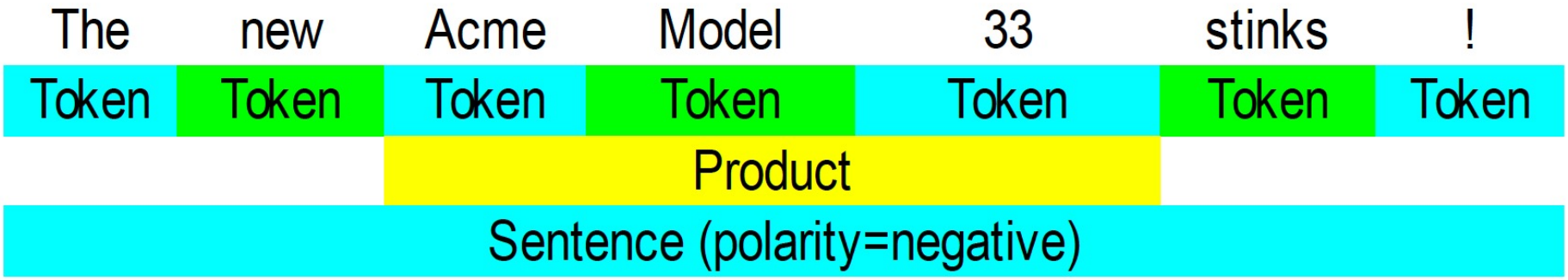
	text	0	1	2	3	4	5	6	7	8	...	15	16	17	18
0	patient with abdominal pain. she has taken aspirin.	-0.605661	0.270313	-0.243304	0.102912	0.362055	-0.057770	0.930624	-0.107427	-0.066948	...	0.222401	0.457482	0.127095	0.258677
1	she has had abdominal pain in the past. pain free today.	-0.252845	0.281251	-0.192978	0.204936	0.046332	-0.046926	0.897162	-0.537238	-0.140003	...	0.260515	0.364818	0.242338	-0.026832
2	no abdominal pain.	0.430410	-0.019595	-0.390710	0.655580	0.810515	0.479990	0.143103	0.725345	0.399600	...	0.242515	-0.187829	0.488414	1.387550
3	takes aspirin for pain. has no pain today.	-0.062846	0.320588	-0.266946	0.262927	0.009739	-0.287657	0.592511	-0.510897	-0.106132	...	0.031052	0.138355	0.044606	0.442873

Classes/labels

- The output variable we want to learn

Classes/labels

- The output variable we want to learn



Classes/labels

- The output variable we want to learn

	text	takes	free	patient	taken	pain	past	abdominal	aspirin	today	label
0	patient with abdominal pain. she has taken aspirin.	0.000	0.000	0.173	0.173	0.0	0.000	0.036	0.087	0.000	current pain
1	she has had abdominal pain in the past. pain free today.	0.000	0.126	0.000	0.000	0.0	0.126	0.026	0.000	0.063	past pain
2	no abdominal pain.	0.000	0.000	0.000	0.000	0.0	0.000	0.096	0.000	0.000	no current pain
3	takes aspirin for pain. has no pain today.	0.173	0.000	0.000	0.000	0.0	0.000	0.000	0.087	0.087	no current pain

Sequence learning tasks in NLP

Anna Larsson PERSON is a famous author from Sweden GPE who now lives in New York GPE .

Her recent book Shadows in the Dark WORK_OF_ART was an international success.

Yesterday DATE at 9 a.m. the TIME IKEA ORG stock went up 30% PERCENT

because of their upcoming launch in New Zealand GPE .

Sequence learning tasks in NLP

What are the instances?

Anna Larsson PERSON is a famous author from Sweden GPE who now lives in New York GPE .

Her recent book Shadows in the Dark WORK_OF_ART was an international success.

Yesterday DATE at 9 a.m. the TIME IKEA ORG stock went up 30% PERCENT

because of their upcoming launch in New Zealand GPE .

Representation - IOB

TEXT	IOB	ENTITY TYPE	DESCRIPTION
Anna	B	PERSON	beginning of an entity
Larsson	I	PERSON	inside an entity
is	O	""	outside an entity
a	O	""	outside an entity
famous	O	""	outside an entity
author	O	""	outside an entity
from	O	""	outside an entity
Sweden	B	GPE	beginning of an entity

Representation - BILOU

TEXT	IOB	ENTITY TYPE	DESCRIPTION
Anna	B	PERSON	beginning of an entity
Larsson	L	PERSON	last of an entity
is	O	""	outside an entity
a	O	""	outside an entity
famous	O	""	outside an entity
author	O	""	outside an entity
from	O	""	outside an entity
Sweden	U	GPE	unit entity

Classification tasks in NLP

- Common ‘old-school’ algorithms:
  - Support Vector Machines
  - Naive Bayes
  - Random Forest
  - Conditional Random Fields (sequence learning)
  - Hidden Markov Models (sequence learning)
- CNN, RNN
- Transformer models