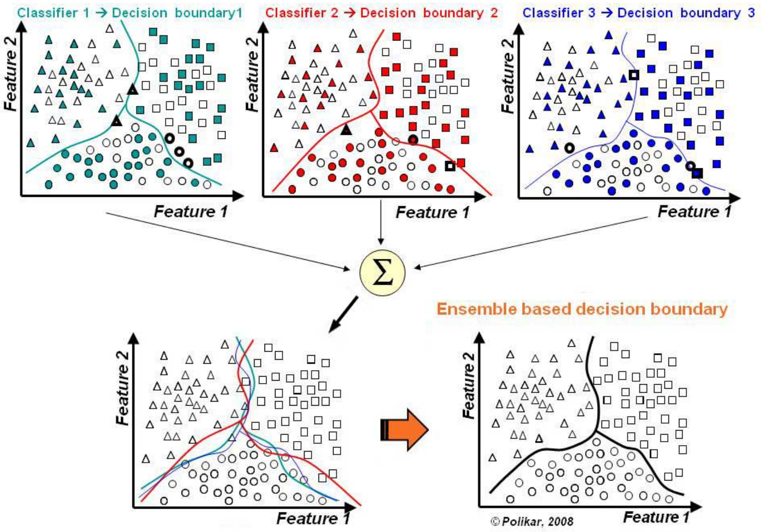
Topical Ensembles for Text Classification

Matti Lyra
Dr. David Weir

Department of Informatics
University of Sussex

November 27, 2015



http://www.scholarpedia.org/article/File:Combining_classifiers2.jpg

Ensemble Models

Ensemble models are meta algorithms that aim to improve performance at a task by aggregating the predictions of many "weak predictors"

- Bagging (Bootstrap Aggregation)
  - Random Patches
  - Random Subspaces
- Random Forests / Extra Trees
- Boosting

Distributional Semantics

- . . . a combined capacity of about 120 megawatts generated by nearly 300 **2348** turbines.
- . . . forecast for Thursday and Friday with higher **2348** gusts to near 50 mph . . .
- Members of the international science community today **2348** up a conference . . .
- . . . the dollar zig-zagging against the mark, only to **2348** up little changed . . .
- I think it is going to take some **2348** out the near-term potential of the stock . . .
- Copper also got the **2348** knocked out when a prominent trader made . . .
- . . . a fund being set up to **2348** up failed mortgage companies . . .

Distributional Semantics

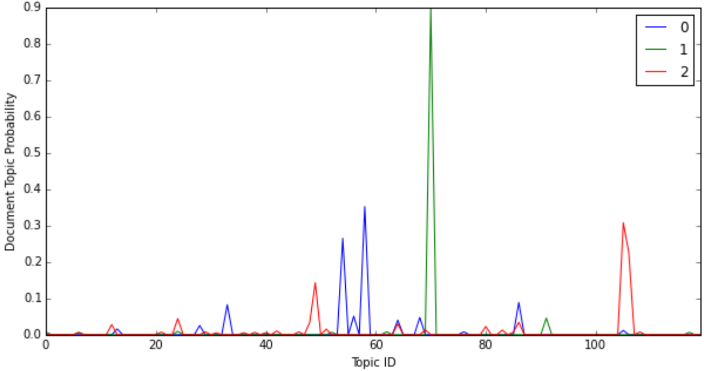
- . . . a combined capacity of about 120 megawatts generated by nearly 300 **wind** turbines.
- . . . forecast for Thursday and Friday with higher **wind** gusts to near 50 mph . . .
- Members of the international science community today **wind** up a conference . . .
- . . . the dollar zig-zagging against the mark, only to **wind** up little changed . . .
- I think it is going to take some **wind** out the near-term potential of the stock . . .
- Copper also got the **wind** knocked out when a prominent trader made . . .
- . . . a fund being set up to **wind** up failed mortgage companies . . .

Topic Modelling

- A topic is a probability distribution over the vocabulary

word	1	2	3	4	5
china	0.025	6.21e-06	2.59e-07	8.70e-07	1.22e-09
market	0.013	1.16e-05	3.55e-06	1.17e-05	1.11e-08
chinese	0.012	2.49e-06	5.71e-07	1.30e-06	5.22e-10
economy	0.008	3.28e-10	9.10e-10	2.21e-06	1.36e-09
currency	0.006	6.46e-11	4.36e-06	6.85e-22	2.03e-09
stock	0.005	1.85e-08	1.99e-06	3.74e-06	4.93e-09
growth	0.005	1.39e-06	6.10e-09	6.02e-09	1.12e-09
global	0.005	2.03e-07	8.43e-07	7.13e-06	7.63e-09
beijing	0.005	6.35e-10	6.85e-22	1.13e-10	6.85e-22
bank	0.004	4.35e-10	1.70e-05	1.87e-06	1.94e-09

Topic Modelling



- A distribution over K topics for each document

Weighted SVM

$$\min_{w,\xi,b} = \frac{1}{2} w^T w + C \sum_{i=1}^{n} W_i \xi_i$$

subject to $\{i = 1, \dots n\}$

$$y_i(\vec{w}\vec{x_i} - b) \geq 1 - \xi_i, \xi_i \geq 0$$

[1]

[1] Learning using privileged information: SVM+ and weighted SVM (Neural Networks, Volume 53, M. Lapin and M. Hein and B. Schiele)

Topical Ensembles using SVM

$$\hat{\theta}_i = lda(x_i)$$

for every $\{j = 1, \ldots K\}$

$$\min_{w,\xi,b} = \frac{1}{2} w^T w + C \sum_{i=1}^{N} \hat{\theta}_{ij} \xi_i$$

Predictions:

- majority voting on binary predictions
- majority voting on topic proportion
- majority voting on SVM confidence
- majority voting on SVM confidence + topic proportion

20 Newsgroups

talk.politics.misc	talk.politics.mideast
talk.politics.guns	talk.religion.misc
sci.med	sci.space
sci.crypt	sci.electronics
rec.autos	rec.sport.hockey
rec.motorcycles	rec.sport.baseball
comp.graphics	comp.os.ms-windows.misc
comp.windows.x	comp.sys.mac.hardware
comp.sys.ibm.pc.hardware	
alt.atheism	misc.forsale
soc.religion.christian	

- each category has 1000 documents
- 20 x 1 vs ALL (40x random splits)
- LDA model only sees training data

20 Newsgroups

	F1-score				
Category	LDA+SVM	LDA+SVM θ	SVM	bag	SVM θ
talk					
.politics.misc	0.68	**0.86**	0.84	0.85	0.22
.politics.guns	0.75	**0.90**	0.88	**0.90**	0.39
.politics.mideast	0.82	**0.94**	**0.94**	**0.95**	0.41
.religion.misc	0.46	**0.80**	0.78	0.77	0.19
sci					
.med	0.76	**0.93**	0.92	**0.93**	0.45
.crypt	0.88	**0.95**	0.94	**0.95**	0.47
.space	0.82	**0.94**	0.92	0.93	0.47
.electronics	0.69	**0.86**	0.83	0.85	0.23

20 Newsgroups

	F1-score				
Category	LDA+SVM	LDA+SVM θ	SVM	bag	SVM θ
rec					
.autos	0.75	0.89	0.89	**0.90**	0.36
.motorcycles	0.84	**0.94**	0.93	**0.94**	0.36
.sport.hockey	0.80	0.94	0.94	**0.95**	0.55
.sport.baseball	0.75	0.92	0.91	**0.93**	0.45
comp					
.graphics	0.65	**0.82**	0.79	**0.83**	0.30
.windows.x	0.74	**0.89**	0.86	**0.88**	0.46
.os.ms-windows.misc	0.70	**0.83**	0.81	**0.82**	0.33
.sys.mac.hardware	0.72	**0.86**	0.84	**0.86**	0.29
.sys.ibm.pc.hardware	0.64	**0.77**	0.75	0.76	0.33

20 Newsgroups

	F1-score				
Category	LDA+SVM	LDA+SVM θ	SVM	bag	SVM θ
alt.atheism	0.64	**0.86**	0.83	**0.85**	0.26
misc.forsale	0.79	**0.85**	0.84	**0.85**	0.34
soc.religion.christian	0.70	**0.88**	0.87	0.87	0.43

TREC - Filtering News for Relevant Stuff

R114, Effects of global warming
Description: *Evidence of effects of global warming or the greenhouse effect on climate and environment.*
Narrative: *Only articles that describe actual changes due to global warming or the greenhouse effect are relevant. Current evidence that points to future effects is relevant.*

TREC - Filtering News for Relevant Stuff

R137, Sea turtle deaths
Description: *Identify any information relevant to the deaths of sea turtles.*
Narrative: *Relevant documents will provide any information with information on the deaths of sea turtles including where and reasons for their death.*

TREC - Filtering News for Relevant Stuff

R143, Improving aircraft safety
Description: *What is being done by U.S. airplane manufacturers to improve the safety of their passenger aircraft?.*
Narrative: *Relevant documents reflect independent actions taken by airlines, under their own initiative, to improve the safety of their passenger aircraft. Documents citing actions taken by the manufacturers as a result of safety mandates imposed by Federal regulations are not relevant.*

RCV1 / TREC

topic ID	#P (train)	#N (train)	#P (test)	#N (test)
R102	135	64	204	662
R104	120	74	98	805
R105	16	21	157	1110
R109	20	20	77	737
R113	12	56	100	1353
R116	16	30	96	1115
R121	14	67	95	1316
R126	19	10	586	583
R129	17	55	71	1302
R141	24	32	89	1268

RCV1 / TREC (> 10)

- in 34 out of 50 cases LDA+SVM significantly outperforms other methods

	F1-score				
Category	LDA+SVM	LDA+SVM θ	SVM	bag	SVM θ
R102	**0.50**	0.47	0.46	**0.51**	0.32
R104	**0.42**	0.28	0.31	0.39	0.15
R105	**0.29**	0.26	0.26	0.26	0.25
R109	0.24	0.27	0.29	**0.31**	0.14
R113	0.004	**0.15**	0.13	**0.15**	0.12
R116	**0.19**	0.18	0.18	**0.19**	0.17
R121	**0.22**	**0.21**	0.19	0.20	0.19
R126	0.57	**0.72**	0.71	**0.73**	0.47
R129	0.08	0.13	**0.18**	0.15	0.08

RCV1 / TREC (< 10)

	F1-score				
Category	LDA+SVM	LDA+SVM θ	SVM	bag	SVM θ
R101	**0.61**	0.41	0.46	0.43	0.33
R106	0.00	**0.08**	**0.08**	**0.08**	0.06
R107	0.00	**0.12**	0.06	0.07	0.07
R108	0.00	**0.11**	0.07	0.08	0.03
R114	0.11	0.16	0.22	**0.24**	0.10
R120	0.00	**0.32**	0.27	0.28	0.29
R131	**0.33**	0.21	0.23	0.20	0.07
R137	**0.36**	0.03	0.04	0.04	0.01
R138	0.00	**0.21**	0.13	0.14	0.10
R143	0.00	0.002	0.03	0.02	0.03

Thank you

Matti Lyra
@mattilyra