

Lexical alignment: feature-rich models

EM for logistic CPDs

Wilker Aziz

April 12, 2017

Content

Representation

EM

ECG

Feature-rich IBM 1-2

Remarks

IBM 1-2: strong assumptions

Independence assumptions

- ▶ $P(A|M, N)$ does not depend on lexical choices
 $a_1 \text{ cute}_2 \text{ house}_3 \leftrightarrow \text{uma}_1 \text{ bela}_2 \text{ casa}_3$

IBM 1-2: strong assumptions

Independence assumptions

- ▶ $P(A|M, N)$ does not depend on lexical choices

a_1 cute₂ house₃ \leftrightarrow uma₁ bela₂ casa₃

a_1 cosy₂ house₃ \leftrightarrow uma₁ casa₃ aconchegante₂

IBM 1-2: strong assumptions

Independence assumptions

- ▶ $P(A|M, N)$ does not depend on lexical choices
a₁ cute₂ house₃ \leftrightarrow uma₁ bela₂ casa₃
a₁ cosy₂ house₃ \leftrightarrow uma₁ casa₃ aconchegante₂
- ▶ $P(F|E)$ can only reasonably explain one-to-one alignments
I will be leaving soon \leftrightarrow vou embora em breve

IBM 1-2: strong assumptions

Independence assumptions

- ▶ $P(A|M, N)$ does not depend on lexical choices
a₁ cute₂ house₃ \leftrightarrow uma₁ bela₂ casa₃
a₁ cosy₂ house₃ \leftrightarrow uma₁ casa₃ aconchegante₂
- ▶ $P(F|E)$ can only reasonably explain one-to-one alignments
I will be leaving soon \leftrightarrow vou embora em breve

Parameterisation

- ▶ categorical events are unrelated
prefixes/suffixes: normal, normally, abnormally, ...
verb inflections: comer, comi, comia, comeu, ...
gender/number: gato, gatos, gata, gatas, ...

Conditional probability distributions

CPD: condition $c \in \mathcal{C}$, outcome $o \in \mathcal{O}$, and $\theta_c \in \mathbb{R}^{|\mathcal{O}|}$

$$P(O|C = c) = \text{Cat}(\theta_c) \quad (1)$$

- ▶ $P(O = o|C = c) = \theta_{c,o}$
- ▶ $0 \leq \theta_{c,o} \leq 1$
- ▶ $\sum_o \theta_{c,o} = 1$
- ▶ $O(|\mathcal{C}| \times |\mathcal{O}|)$ parameters

How bad is it for IBM model 1?

Probability tables

$$P(F|E)$$

ENGLISH ↓	FRENCH →			
	anormal	normal	normalmente	...
abnormal	0.7	0.1	0.01	...
normal	0.01	0.6	0.2	...
normally	0.001	0.25	0.65	...

- ▶ grows with size of vocabularies
- ▶ no parameter sharing

Logistic CPDs

CPD: condition $c \in \mathcal{C}$ and outcome $o \in \mathcal{O}$

$$P(O = o|C = c) = \frac{\exp(w^\top h(c, o))}{\sum_{o'} \exp(w^\top h(c, o'))} \quad (2)$$

- ▶ $w \in \mathbb{R}^d$ is a weight vector
- ▶ $h : \mathcal{C} \times \mathcal{O} \rightarrow \mathbb{R}^d$ is a feature function
- ▶ d parameters
- ▶ computing CPD requires $O(|\mathcal{C}| \times |\mathcal{O}| \times d)$ operations

How bad is it for IBM model 1?

CPDs as functions

$$h : \mathcal{E} \times \mathcal{F} \rightarrow R^d$$

EVENTS ↓		FEATURES →				
ENGLISH	FRENCH	normal normal	<i>normal</i> - <i>normal</i> -	<u>-normal</u> <u>-normal</u>	ab - a -	- ly - mente
abnormal	a <u>normal</u>	0	0	1	1	0
	<u>normal</u>	0	0	1	0	0
	<i>normal</i> / <i>mente</i>	0	1	0	0	0
normal	<u>a</u> normal	0	0	1	0	0
	normal	1	0	0	0	0
	<i>normal</i> / <i>mente</i>	0	1	0	0	0
normally	<u>a</u> normal	0	0	1	0	0
	<i>normal</i>	0	1	0	0	0
	<i>normal</i> mente	0	1	0	0	1
WEIGHTS →		1.5	0.3	0.3	0.8	1.1

- ▶ computation still grows with size of vocabularies
- ▶ but far less parameters to estimate

Content

Representation

EM

ECG

Feature-rich IBM 1-2

Remarks

Expectation Maximisation

Coordinate ascent in \mathcal{F}

[Neal and Hinton, 1998]

$$\mathcal{L}(\theta) \equiv \log P(X|\theta) \geq \mathbb{E}_{Q(Z|X,\psi)} [\log P(X, Z|\theta)] + H(Q) \quad (3)$$

$$\equiv \mathcal{F}(Q, \theta) \quad (4)$$

Expectation Maximisation

Coordinate ascent in \mathcal{F}

[Neal and Hinton, 1998]

$$\mathcal{L}(\theta) \equiv \log P(X|\theta) \geq \mathbb{E}_{Q(Z|X,\psi)} [\log P(X, Z|\theta)] + H(Q) \quad (3)$$

$$\equiv \mathcal{F}(Q, \theta) \quad (4)$$

E-step: choose $Q^{(t+1)}$ that maximises \mathcal{F} for fixed $\theta^{(t)}$

problem $Q^{(t+1)} = \arg \max_Q F(Q, \theta^{(t)})$

solution $Q^{(t+1)}(z|x, \psi) = P(z|x, \theta^{(t)})$
which means using the **exact posterior**

Expectation Maximisation

Coordinate ascent in \mathcal{F}

[Neal and Hinton, 1998]

$$\mathcal{L}(\theta) \equiv \log P(X|\theta) \geq \mathbb{E}_{Q(Z|X,\psi)} [\log P(X, Z|\theta)] + H(Q) \quad (3)$$

$$\equiv \mathcal{F}(Q, \theta) \quad (4)$$

E-step: choose $Q^{(t+1)}$ that maximises \mathcal{F} for fixed $\theta^{(t)}$

problem $Q^{(t+1)} = \arg \max_Q F(Q, \theta^{(t)})$

solution $Q^{(t+1)}(z|x, \psi) = P(z|x, \theta^{(t)})$

which means using the **exact posterior**

M-step: choose $\theta^{(t+1)}$ that maximises \mathcal{F} for fixed $Q^{(t+1)}$

problem $\theta^{(t+1)} = \arg \max_{\theta} \mathcal{F}(Q^{(t+1)}, \theta)$

Gradient-based M-step for logistic CPDs

For each distribution t , with context c and outcome o

$$\theta_{t,c,o}(w) = \frac{\exp(w^\top h(t, c, o))}{\sum_{o'} \exp(w^\top h(t, c, o'))} \quad (5)$$

Gradient-based M-step for logistic CPDs

For each distribution t , with context c and outcome o

$$\theta_{t,c,o}(w) = \frac{\exp(w^\top h(t, c, o))}{\sum_{o'} \exp(w^\top h(t, c, o'))} \quad (5)$$

Expected counts

$$\mu_{t,c,o} = \mathbb{E} [n(t : c \rightarrow o | Z)] \quad (6)$$

Gradient-based M-step for logistic CPDs

For each distribution t , with context c and outcome o

$$\theta_{t,c,o}(w) = \frac{\exp(w^\top h(t, c, o))}{\sum_{o'} \exp(w^\top h(t, c, o'))} \quad (5)$$

Expected counts

$$\mu_{t,c,o} = \mathbb{E}[n(t : c \rightarrow o | Z)] \quad (6)$$

Expected complete log likelihood

$$\ell(w | \mu) = \sum_{t,c,o} \mu_{t,c,o} \log \theta_{t,c,o}(w) \quad (7)$$

Gradient-based M-step for logistic CPDs

For each distribution t , with context c and outcome o

$$\theta_{t,c,o}(w) = \frac{\exp(w^\top h(t, c, o))}{\sum_{o'} \exp(w^\top h(t, c, o'))} \quad (5)$$

Expected counts

$$\mu_{t,c,o} = \mathbb{E}[n(t : c \rightarrow o | Z)] \quad (6)$$

Expected complete log likelihood

$$\ell(w | \mu) = \sum_{t,c,o} \mu_{t,c,o} \log \theta_{t,c,o}(w) \quad (7)$$

Gradient wrt w (for fixed μ)

$$\nabla_w \ell(w | \mu) = \sum_{t,d,o} \mu_{t,d,o} \Delta_{t,c,o}(w) \quad (8)$$

$$\Delta_{t,c,o}(w) = h(t, c, o) - \sum_{o'} \theta_{t,c,o'}(w) h(t, c, o') \quad (9)$$

Content

Representation

EM

ECG

Feature-rich IBM 1-2

Remarks

Expectation Conjugate Gradient (ECG)

Direct marginal likelihood optimisation [Salakhutdinov et al., 2003]

$$\nabla_{\theta} \log P(X|\theta) = \mathbb{E}_{P(Z|X,\theta)} [\nabla_{\theta} \log P(X, Z|\theta)] \quad (10)$$

EM: until convergence

1. compute expected counts μ
2. repeat until convergence
 - ▶ compute $l(w|\mu)$
 - ▶ compute $\nabla \ell(w|\mu)$
 - ▶ $w \leftarrow \text{climb}(w, \ell(w|\mu), \nabla \ell(w|\mu))$

ECG: until convergence

1. compute expected counts μ
2. compute $\mathcal{L}(w)$
3. compute $\nabla \ell(w|\mu)$
4. $w \leftarrow \text{climb}(w, \ell(w|\mu), \nabla \ell(w|\mu))$

Content

Representation

EM

ECG

Feature-rich IBM 1-2

Remarks

Lexical distribution in IBM model 1

$$P(F = f|E = e) = \frac{\exp(w_{\text{lex}}^{\top} h_{\text{lex}}(e, f))}{\sum_{f'} \exp(w_{\text{lex}}^{\top} h_{\text{lex}}(e, f'))} \quad (11)$$

Features

- ▶ prefixes/suffixes
- ▶ character n -grams
- ▶ POS tags

Extension: lexicalised jump distribution

$$P(\Delta = \delta | E = e) = \frac{\exp(w_{\text{dist}}^{\top} h_{\text{dist}}(e, \delta))}{\sum_{\delta'} \exp(w_{\text{dist}}^{\top} h_{\text{dist}}(e, \delta'))} \quad (12)$$

Features

- ▶ POS tags
- ▶ suffixes/prefixes
- ▶ lemma
- ▶ jump values
- ▶ m, n, j, i (values used to compute jump)

Extension: nonlinear models

Nothing prevents us from using more expressive functions

[Kočíský et al., 2014]

- ▶ $P(O|C = c) = \text{softmax}(f_{\theta}(c))$
- ▶ $P(O = o|C = c) = \frac{\exp(f_{\theta}(c,o))}{\sum_{o'} \exp(f_{\theta}(c,o'))}$

where $f_{\theta}(\cdot)$ is a neural network with parameters θ

Features

- ▶ induce features (word-level, char-level, n -gram level)
- ▶ pre-trained embeddings

Content

Representation

EM

ECG

Feature-rich IBM 1-2

Remarks

Limitations

Local normalisation may be expensive
but see [Gutmann and Hyvärinen, 2012]

Limitations

Local normalisation may be expensive
but see [Gutmann and Hyvärinen, 2012]

E-step takes $O(|\mathcal{D}| \times m \times n)$

- ▶ EM: reuses expected counts
- ▶ ECG: always recomputes expected counts

References I

- Taylor Berg-Kirkpatrick, Alexandre Bouchard-Côté, John DeNero, and Dan Klein. Painless unsupervised learning with features. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 582–590, Los Angeles, California, June 2010. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/N10-1083>.
- Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *J. Mach. Learn. Res.*, 13(1): 307–361, February 2012. ISSN 1532-4435. URL <http://dl.acm.org/citation.cfm?id=2503308.2188396>.

References II

- Tomáš Kočiský, Karl Moritz Hermann, and Phil Blunsom. Learning bilingual word representations by marginalizing alignments. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 224–229, Baltimore, Maryland, June 2014. Association for Computational Linguistics. URL <http://www.aclweb.org/anthology/P14-2037>.
- Radford M. Neal and Geoffrey E. Hinton. *A View of the EM Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht, 1998. ISBN 978-94-011-5014-9. doi: 10.1007/978-94-011-5014-9_12. URL http://dx.doi.org/10.1007/978-94-011-5014-9_12.

References III

Ruslan Salakhutdinov, Sam Roweis, and Zoubin Ghahramani.
Optimization with em and expectation-conjugate-gradient. In
*Proceedings of the Twentieth International Conference on
International Conference on Machine Learning*, ICML'03, pages
672–679. AAAI Press, 2003. ISBN 1-57735-189-4. URL
<http://dl.acm.org/citation.cfm?id=3041838.3041923>.