Mach Learn (2013) 93:5–29
DOI 10.1007/s10994-013-5391-2

Probabilistic topic models for sequence data

Nicola Barbieri · Giuseppe Manco · Ettore Ritacco · Marco Carnuccio · Antonio Bevacqua

Received: 9 September 2012 / Accepted: 4 June 2013 / Published online: 3 July 2013

Abstract Probabilistic topic models are widely used in different contexts to uncover the hidden structure in large text corpora. One of the main (and perhaps strong) assumption of these models is that generative process follows a bag-of-words assumption, i.e. each token is independent from the previous one. We extend the popular Latent Dirichlet Allocation model by exploiting three different conditional Markovian assumptions: (i) the token generation depends on the current topic and on the previous token; (ii) the topic associated with each observation depends on topic associated with the previous one; (iii) the token generation depends on the current and previous topic. For each of these modeling assumptions we present a Gibbs Sampling procedure for parameter estimation. Experimental evaluation over real-word data shows the performance advantages, in terms of recall and precision, of the sequence-modeling approaches.

Editors: Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen, and Filip Železný.

N. Barbieri (✉)
Yahoo Research, Av. Diagonal 177, Barcelona, Spain
e-mail: barbieri@yahoo-inc.com

G. Manco · E. Ritacco
Institute for High Performance Computing and Networks (ICAR), Italian National Research Council,
via Bucci 41c, 87036 Rende, CS, Italy

G. Manco
e-mail: manco@icar.cnr.it

E. Ritacco
e-mail: ritacco@icar.cnr.it

M. Carnuccio · A. Bevacqua
Department of Electronics, Informatics and Systems, University of Calabria, via Bucci 41c,
87036 Rende, CS, Italy

M. Carnuccio
e-mail: mcarnuccio@deis.unical.it

A. Bevacqua
e-mail: abevacqua@deis.unical.it