A temporal model of text periodicities using Gaussian Processes

The University Of Sheffield.

Daniel Preoțiuc-Pietro, Trevor Cohn
Department of Computer Science
{daniel,t.cohn}@dcs.shef.ac.uk

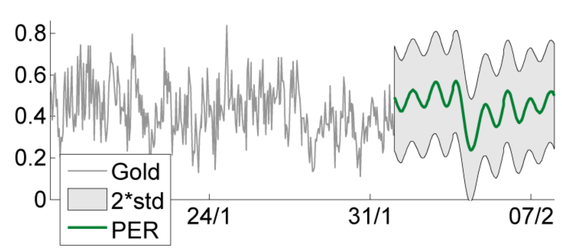
Data

1176 hashtags time series
from 1.1.2011-28.2.2011
~6.5 mil deduplicated tweets
~9.55 voc.tokens/tweet
a proxy for topics on Twitter

AIMS: given hashtag time series, use GPs to: find periodicities, group time series, forecast future values, text classification using periodicity info

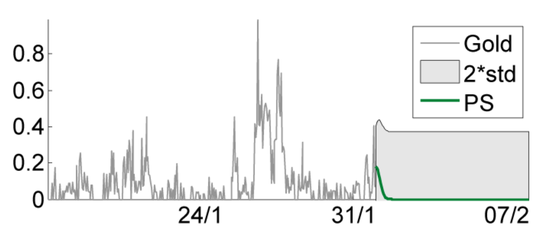
#brb
#coffee
#facebook
#facepalm
#funny
#love
#rock
#running
#xbox
#youtube

#FAIL



$$k_{PER}(t,t') = s^2 \cdot \exp -2 \cdot \left(\frac{\sin^2(2\pi(t-t')^2/p)}{l^2} \right)$$

#SNOW

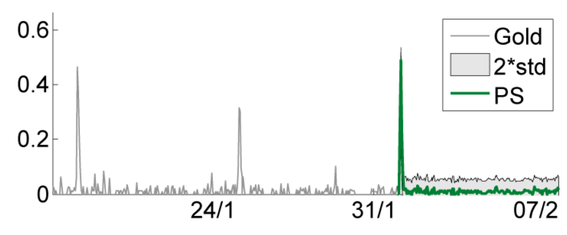


$$k_{SE}(t,t') = s^2 \cdot \exp -\frac{(t-t')^2}{2l^2}$$

#2011
#backintheday
#confessionhour
#februarywish
#haiti
#makeachange
#questionsidontlike
#savelibraries
#snow
#snowday

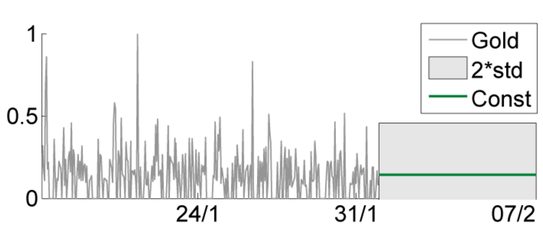
#breakfast
#eastenders
#ff
#followfriday
#goodnight
#jobs
#news
#tgif
#thegame
#ww

#RAW



$$k_{PS}(t,t') = \cos\left(\sin\left(\frac{2\pi \cdot (t-t')^2}{p}\right)\right) \cdot \exp\left(\frac{s\cos(2\pi \cdot (t-t')^2)}{p} - s\right)$$

#FYI



$$k_C(t,t') = c$$

#funny
#lego
#likeaboss
#money
#nbd
#nf
#notetoself
#priorities
#social
#true

Gaussian Processes (GP)

METHOD

- Bayesian non-parametric framework
- regarded as state-of-the-art for regression
- assumes a latent function drawn from a GP prior:

$$f(t) \sim \mathcal{GP}(m, k(t,t'))$$

m - mean, k - kernel

- the predictive posterior can be computed analytically

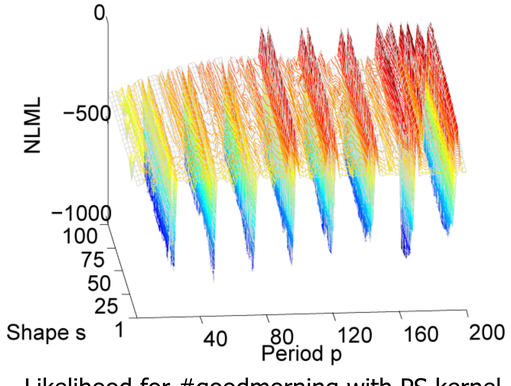
$$x_* \sim \mathcal{N}(k_*^T(K+\sigma_n^2 I)^{-1}\mathbf{t}, k(t_*,t_*) - k_*^T(K+\sigma_n^2 I)^{-1}k_*)$$

$K = \{k(t_i,t_j)\}_{j=1..n}^{i=1..n}$ Gram matrix k_* kernel evals between test and all $\mathbf{t}$ training points

- GPs explicitly incorporate uncertainty

KERNELS

- extrapolation setting: much harder than interpolation
- GP is fully defined by its kernel (assuming 0 mean)
- kernel induces similarities in the response between pairs of data points
- intuitively:
  smooth function -> closer points, high covariance
  periodic function -> points at period length, high covariance
- for extrapolation, kernel choice is paramount



Likelihood for #goodmorning with PS kernel

Model Selection

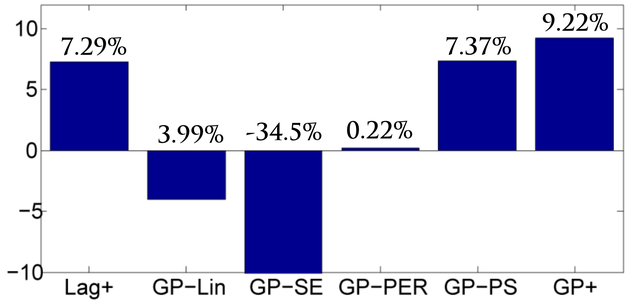
- given a model, compute probability of the data integrating over the parameter space i.e. Bayesian 'evidence'
- conditioned on kernel parameters, the evidence can be computed analytically

$$p(x|\mathcal{D},\theta,\mathcal{H}_i) = \int_f p(x|\mathcal{D},f,\mathcal{H}_i)p(f|\theta,\mathcal{H}_i)$$

$\mathcal{H}_i$ model $\mathcal{D}$ dataset θ kernel parameters

- balances data fit and model complexity (Occam's Razor)
- complex models which can account for many datasets achieve low evidence
- use Negative log Marginal Likelihood (ML-II) for model selection, giving an implicit classification of time series

Forecasting



- train on 1 month, predict 1 month in the future
- performance compared to mean prediction (=GP-Const)
- GP+ performs model selection
- Lag+ AR model that uses the GP determined period

Text Classification

Task: Predict hashtag based on tweet text
Use GP forecast as prior for Naive Bayes