Nonextensive Entropic Kernels

André Martins[1,3] Pedro Aguiar[2] Mário Figueiredo[3]
Noah Smith[1] Eric Xing[1]

[1]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA

[2]Instituto de Sistemas e Robótica
Instituto Superior Técnico
Lisboa, Portugal

[3]Instituto de Telecomunicações
Instituto Superior Técnico
Lisboa, Portugal

Summary

Outline

Outline

- We want to classify structured objects (strings, trees, graphs, ...)

Outline

- We want to classify structured objects (strings, trees, graphs, ...)
  - Generative methods allow modeling data generation
  - Discriminative methods directly discriminate data

Outline

- We want to classify structured objects (strings, trees, graphs, ...)
  - Generative methods allow modeling data generation
  - Discriminative methods directly discriminate data
- How to get the best of both worlds?

Outline

- We want to classify structured objects (strings, trees, graphs, ...)
  - Generative methods allow modeling data generation
  - Discriminative methods directly discriminate data
- How to get the best of both worlds?
  - Represent objects x, y as probability distributions $p_x(.), p_y(.)$
  - Use a kernel between distributions, $k(p_x, p_y)$

Outline

- We want to classify structured objects (strings, trees, graphs, ...)
  - Generative methods allow modeling data generation
  - Discriminative methods directly discriminate data
- How to get the best of both worlds?
  - Represent objects x, y as probability distributions $p_x(.), p_y(.)$
  - Use a kernel between distributions, $k(p_x, p_y)$
- This work: A new family of kernels between distributions

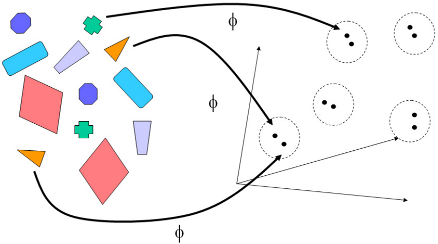
Outline

- We want to classify structured objects (strings, trees, graphs, ...)
  - Generative methods allow modeling data generation
  - Discriminative methods directly discriminate data
- How to get the best of both worlds?
  - Represent objects x, y as probability distributions $p_x(.), p_y(.)$
  - Use a kernel between distributions, $k(p_x, p_y)$
- This work: A new family of kernels between distributions
  - Grounded on nonextensive (Tsallis) information theory
  - Contains known kernels as particular cases
  - Experiments in text classification

Outline

Hilbert space embedding

- Theorem: $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a positive definite (pd) kernel iff there is a feature space $\mathcal{F}$ and a map $\Phi : \mathcal{X} \to \mathcal{F}$, such that $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathcal{F}}$



- A kernel induces a similarity measure

Kernels for structured data

- What if $\mathcal{X}$ is structured?

Kernels for structured data

- What if $\mathcal{X}$ is structured?
  - Extract features and use a linear kernel (Joachims, 1997)
  - Decompose objects into subparts (convolution kernels, Haussler, 1999)
  - Generative approach through Fisher kernel (Jaakkola, 1999)

Kernels for structured data

- What if $\mathcal{X}$ is structured?
  - Extract features and use a linear kernel (Joachims, 1997)
  - Decompose objects into subparts (convolution kernels, Haussler, 1999)
  - Generative approach through Fisher kernel (Jaakkola, 1999)
- Our approach: Map each object to a probability distribution, and devise kernels on probability distributions:

$$\begin{array}{l} x \mapsto p_x(.) \\ y \mapsto p_y(.) \end{array} \quad \Leftrightarrow \quad K(x, y) \triangleq k(p_x, p_y)$$

Kernels on probability distributions

- Inner product kernels (Jebara, Kondor, Howard, 2004)

$$k_{JKH}(\mathbf{p}_1, \mathbf{p}_2) \triangleq \langle \mathbf{p}_1^{\alpha}, \mathbf{p}_2^{\alpha} \rangle$$

- (†) Information geometry of the multinomial (Lafferty, Lebanon, 2005),

$$k_{\text{heat}}(\mathbf{p}_1, \mathbf{p}_2) \approx \exp\left(-\lambda d_g^2(\mathbf{p}_1, \mathbf{p}_2)\right)$$

- (†) KL divergence (Moreno, Ho, Vasconcelos, 2003),

$$k_{MHV}(\mathbf{p}_1, \mathbf{p}_2) \triangleq \exp\left(-\lambda(KL(\mathbf{p}_1, \mathbf{p}_2) + KL(\mathbf{p}_2, \mathbf{p}_1))\right)$$

(†) not pd

Kernels on probability distributions (c'ed)

- Jensen-Shannon (JS) divergence (Burbea, Rao, 1982; Lin, 1991)

$$\begin{aligned} JS(\mathbf{p}_1, \mathbf{p}_2) &\triangleq \frac{1}{2} KL\left(\mathbf{p}_1, \frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) + \frac{1}{2} KL\left(\mathbf{p}_2, \frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) \\ &= H\left(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) - \frac{H(\mathbf{p}_1) + H(\mathbf{p}_2)}{2} \end{aligned}$$

  - Replace KL by JS divergence $\Rightarrow$ pd (Cuturi, Fukumizu, Vert, 2005; Hein, Bousquet, 2005):

$$\begin{aligned} k_{CFV}(\mathbf{p}_1, \mathbf{p}_2) &= \exp\left(-\lambda JS(\mathbf{p}_1, \mathbf{p}_2)\right) \\ k_{HB}(\mathbf{p}_1, \mathbf{p}_2) &= \ln 2 - JS(\mathbf{p}_1, \mathbf{p}_2) \end{aligned}$$

- We subsume some of these kernels by going from classic to nonextensive (Tsallis) information theory!

Outline

Shannon entropy (1948)

- Random variable $X \in \mathcal{X} = \{x_1, \ldots, x_n\}$

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} P(x_i) \ln P(x_i) \\ &= -\mathbb{E}[\ln P(X)] \end{aligned}$$

- Extensivity: for X and Y independent,

$$H(X, Y) = H(X) + H(Y)$$

- “Independent systems add their entropies”—cf. Boltzmann-Gibbs entropy in statistical thermodynamics

Rényi entropies (1961)

- A family parameterized by $q \geq 0$,

$$R_q(X) = \frac{1}{1-q} \ln \sum_{i=1}^{n} P(x_i)^q$$

- Shannon's entropy as a limit:

$$\lim_{q \to 1} R_q(X) = H(X)$$

- Still extensive: for X and Y independent,

$$R_q(X, Y) = R_q(X) + R_q(Y)$$

Tsallis entropies (1988)

- A family parameterized by $q \geq 0$ (the entropic index),

$$S_q(X) = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} P(x_i)^q\right)$$

- Shannon's entropy as a limit:

$$\lim_{q \to 1} S_q(X) = H(X)$$

- Not extensive! For X and Y independent,

$$S_q(X, Y) = S_q(X) + S_q(Y) - (q-1)S_q(X)S_q(Y)$$

- Nonextensive thermodynamics—claimed to better model some physical phenomena (*e.g.* long range interactions, heavy-tailed distributions)

Tsallis entropies

- Tsallis entropies can be written as:

$$S_q(X) = -\mathbb{E}_q[\ln_q P(X)]$$

Tsallis entropies

- Tsallis entropies can be written as:

$$S_q(X) = -\mathbb{E}_q[\ln_q P(X)]$$

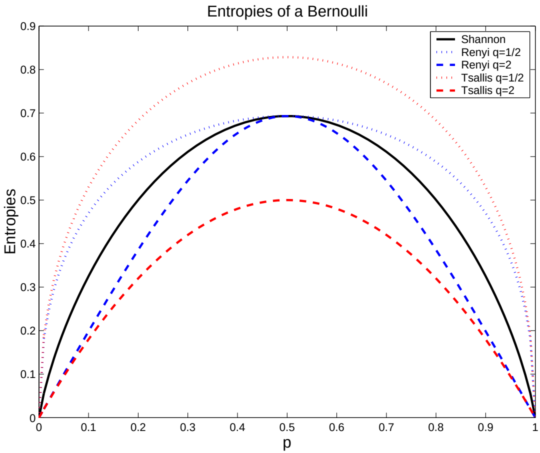
- *q*-expectation:

$$\mathbb{E}_q\left[f(X)\right] = \sum_i P(x_i)^q f(x_i), \qquad \lim_{q \to 1} \mathbb{E}_q\left[f(X)\right] = \mathbb{E}\left[f(X)\right]$$

- *q*-logarithm:

$$\ln_q(x) = \frac{x^{1-q} - 1}{1 - q}, \qquad \lim_{q \to 1} \ln_q(x) = \ln(x)$$

Tsallis entropies



Tsallis entropies

- Joint Tsallis entropy:

$$S_q(X, Y) = -\mathbb{E}_q[\ln_q P(X, Y)]$$

- Conditional Tsallis entropy:

$$S_q(X|Y) = -\mathbb{E}_q[\ln_q P(X|Y)]$$

- Chain rule:

$$S_q(X, Y) = S_q(X|Y) + S_q(Y)$$

- Tsallis mutual information (Furuichi, 2006):

$$I_q(X; Y) = S_q(X) - S_q(X|Y)$$

Outline

Jensen differences

- Jensen's inequality: for a concave function f (*e.g.*, Shannon, Rényi, or Tsallis entropies),

$$f(\mathbb{E}[Z]) \geq \mathbb{E}[f(Z)]$$

- Weighted Jensen-Shannon divergence of m distributions

$$\begin{aligned} J_H^{\pi}(\mathbf{p}_1, \ldots, \mathbf{p}_m) &\triangleq \underbrace{H\left(\sum_{j=1}^{m} \pi_j \, \mathbf{p}_j\right)}_{H(X)} - \underbrace{\sum_{j=1}^{m} \pi_j H(\mathbf{p}_j)}_{H(X|Y)} \\ &= I(X;Y), \end{aligned}$$

where $Y \sim (\pi_1, ..., \pi_m)$ and $P(X|Y=j) = \mathbf{p}_j$

Jensen differences

- Jensen-Rényi divergences:

$$J^{\boldsymbol{\pi}}_{R_q}(\mathbf{p}_1, \ldots, \mathbf{p}_m) \triangleq R_q\left(\sum_{j=1}^{m} \pi_j \, \mathbf{p}_j\right) - \sum_{j=1}^{m} \pi_j R_q(\mathbf{p}_j)$$

- Jensen-Tsallis divergences:

$$J^{\boldsymbol{\pi}}_{S_q}(\mathbf{p}_1, \ldots, \mathbf{p}_m) \triangleq S_q\left(\sum_{j=1}^{m} \pi_j \, \mathbf{p}_j\right) - \sum_{j=1}^{m} \pi_j S_q(\mathbf{p}_j)$$

Jensen differences

- Jensen-Rényi divergences:

$$J^{\boldsymbol{\pi}}_{R_q}(\mathbf{p}_1, \ldots, \mathbf{p}_m) \triangleq R_q\left(\sum_{j=1}^{m} \pi_j \, \mathbf{p}_j\right) - \sum_{j=1}^{m} \pi_j R_q(\mathbf{p}_j)$$

- Jensen-Tsallis divergences:

$$J^{\boldsymbol{\pi}}_{S_q}(\mathbf{p}_1, \ldots, \mathbf{p}_m) \triangleq S_q\left(\sum_{j=1}^{m} \pi_j \, \mathbf{p}_j\right) - \sum_{j=1}^{m} \pi_j S_q(\mathbf{p}_j)$$

- No mutual information interpretation!

Outline

Jensen q-differences

- $f : \mathcal{X} \to \mathbb{R}$ is q-convex iff, for any $x, y \in \mathcal{X}$ and $\lambda \in [0,1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda^q f(x) + (1-\lambda)^q f(y)$$

- q-Jensen's inequality: for f q-concave (*e.g.*, $f = S_q$, $q \geq 1$),

$$f(\mathbb{E}[Z]) \geq \mathbb{E}_q[f(Z)]$$

- Jensen-Tsallis q-difference:

$$\begin{aligned} T_q^{\boldsymbol{\pi}}(\mathbf{p}_1, \ldots, \mathbf{p}_m) &\triangleq \underbrace{S_q\left(\sum_{j=1}^m \pi_j \, \mathbf{p}_j\right)}_{S_q(X)} - \underbrace{\sum_{j=1}^m \pi_j^q S_q(\mathbf{p}_j)}_{S_q(X|Y)} \\ &= I_q(X;Y) \end{aligned}$$

Jensen q-differences

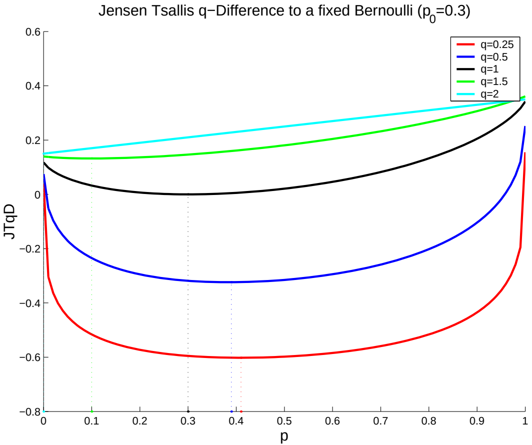
- Let's focus on $m = 2$, $\boldsymbol{\pi} = (\frac{1}{2}, \frac{1}{2})$ ("fair coin").
- Balanced JS divergence

$$JS(\mathbf{p}_1, \mathbf{p}_2) \triangleq H\left(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) - \frac{H(\mathbf{p}_1) + H(\mathbf{p}_2)}{2}.$$

- Balanced JT q-difference

$$T_q(\mathbf{p}_1, \mathbf{p}_2) \triangleq S_q\left(\frac{\mathbf{p}_1 + \mathbf{p}_2}{2}\right) - \frac{S_q(\mathbf{p}_1) + S_q(\mathbf{p}_2)}{2^q}.$$

Jensen-Tsallis q-differences



Outline

Jensen-Tsallis kernels

- Definition:

$$
\begin{aligned}
k_q(\mathbf{p}_1, \mathbf{p}_2) &\triangleq \ln_q(2) - T_q(\mathbf{p}_1, \mathbf{p}_2) \\
&= \frac{1}{2^q(q-1)} \sum_i \left((p_{1i} + p_{i2})^q - p_{1i}^q - p_{2i}^q \right)
\end{aligned}
$$

- These kernels can be extended to unnormalized distributions (see paper)
- Proposition: the kernel k_q is pd for $q \in [0, 2]$.

Special cases

- $q = 0$: Boolean kernel,

$$k_{Bool}(\mathbf{p}_1, \mathbf{p}_2) = \|\mathbf{p}_1 \odot \mathbf{p}_2\|_0$$

- $q = 1$: JS kernel (Hein & Bousquet, 2005),

$$k_{JS}(\mathbf{p}_1, \mathbf{p}_2) = \ln(2) - JS(\mathbf{p}_1, \mathbf{p}_2)$$

- $q = 2$: linear kernel,

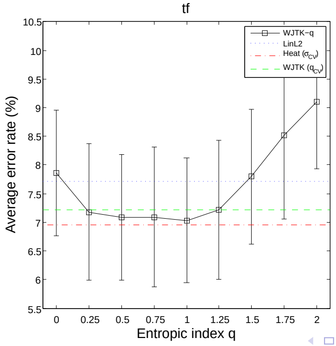
$$k_{lin}(\mathbf{p}_1, \mathbf{p}_2) = \frac{1}{2} \langle \mathbf{p}_1, \mathbf{p}_2 \rangle$$

- Corollary: All these kernels are pd.

Outline

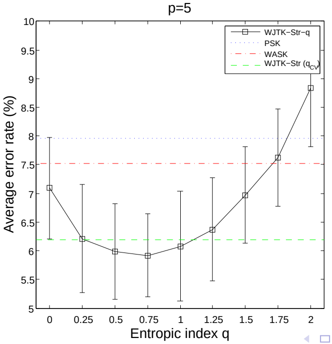
Text classification experiments

- WebKB dataset (student vs faculty homepages)
  - 400 documents for training, 450 for testing
- Each document mapped into bag-of-words unigram model
- Baselines: linear kernel with ℓ_2 normalization (Joachims, 1999) and heat kernel (not pd, Lafferty, Lebanon, 2004)



Text classification experiments (c'ed)

- WebKB dataset (student vs faculty homepages)
  - 400 documents for training, 450 for testing
- Each document mapped into bag-of-5-grams (string kernel)
- Baselines: *p*-spectrum kernel (Leslie, 2002) and all-substrings kernel (Vishwanathan, Smola, 2003) with ℓ_2 normalization



Outline

Conclusions and future work

- A new family of kernels on distributions
- based on nonextensive (Tsallis) information theory
- contains some previously known kernels
- defined on possibly unnormalized distributions (see paper)
- shown to be positive definite
- proofs, kernels between stationary stochastic processes, etc.: `http://www.cs.cmu.edu/~afm/Home_files/CMU-ML-08-106.pdf`
- preliminary experiments on text classification
- future work: exploit nonextensivity in other problems
- future work: when is $q < 1$ best? When is $q > 1$ best?
- future work: multi-kernels