MASARYKOVA UNIVERZITA

Bregmanove divergencie

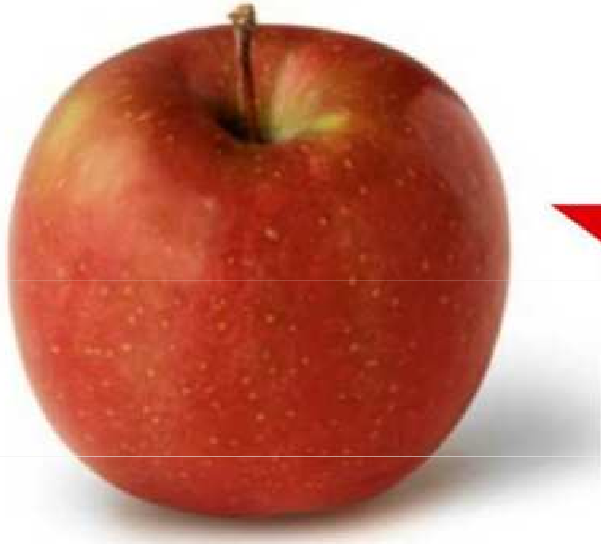
Využitie indexovacích štruktúr pre efektívne podobnostné vyhľadávanie

Lukáš Holecy

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE esf MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ


Ľudské vnímanie podobnosti často nie je metrické

Distribúcia pravdepodobnosti
Rozdiel signálov
Porovnanie rôznych atribútov obrázkov (farby, textúry, tvary)
Sledovanie pohybu
A dalšie

Bregmanová divergencia

$$D_f(p,q) = f(p) - f(q) - \langle \nabla f(q), p - q \rangle$$

- Každá Bregmanova divergencia je založená na nejakej rýdzo konvexnej funkcii f

Kullback-Leibler divergence

Založené na funkcii: $\sum_{i=1}^{d} x_i \cdot \log(x_i)$

$$D_f(p,q) = \sum_{i=1}^{d} p_i \cdot \log\left(\frac{p_i}{q_i}\right)$$

Itakura-Saito divergence

Založené na funkcii: $-\sum_{i=1}^{d} \log(x)$

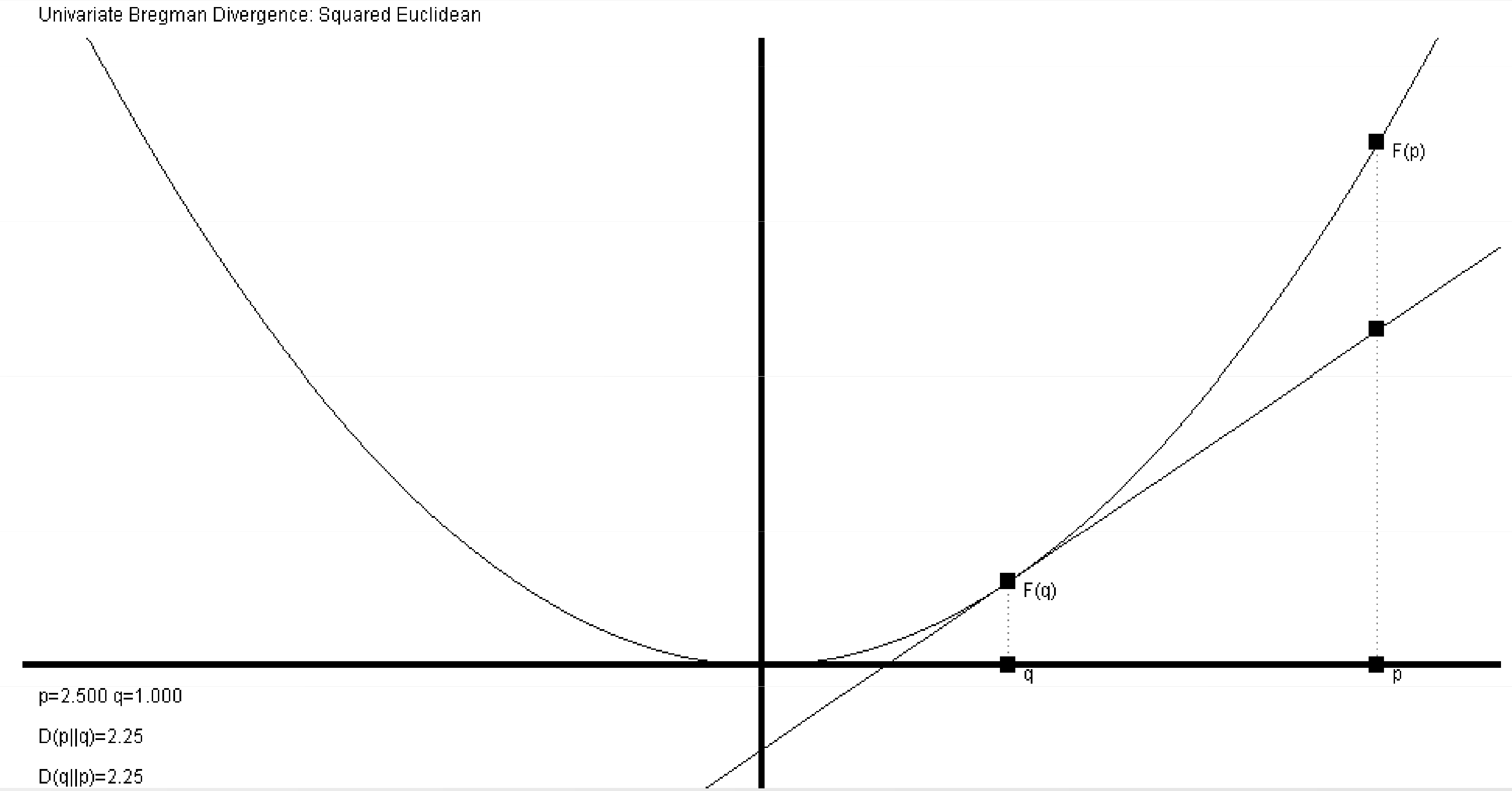
$$D_f(p,q) = \sum_{i=1}^{d} \left(\frac{p_i}{q_i} - \log\left(\frac{p_i}{q_i}\right) - 1\right)$$

Squared Euclidean

Založené na funkcii: $\|x\|^2 = \sum_{i=1}^{d} (x^2)$

$$D_f(p,q) = \|p-q\|^2 = \sum_{i=1}^{d} (p_i - q_i)^2$$

Grafická reprezentácia



http://www.sonycsl.co.jp/person/nielsen/BregmanDivergence/

Problém

Bregmanové divergencie nie sú symetrické

Má ľahké riešenie: $D_f(x,y) = \min_z \left(\frac{1}{2} \left(D_f(x,z) + D_f(y,z) \right) \right)$

V Bregmanových divergenciách neplatí trojuholníková nerovnosť

Nemá ľahké riešenie

Špeciálne indexovacie štruktúry pre každú Bregmanovú divergenciu

Univerzálna indexovacia štruktúra

Obecná metóda pre všetky Bregmanové divergencie

Máme d rozmerný priestor S v ktorom vyhľadávame

Rozšírime priestor S na d + 1 rozmerný priestor S^+, tak, že každému vektoru x v prestore S
pridáme nový rozmer ktorého hodnota bude f(x) kde f je funkcia podľa ktorej je definovaná konkrétna Bregmanová divergencia.

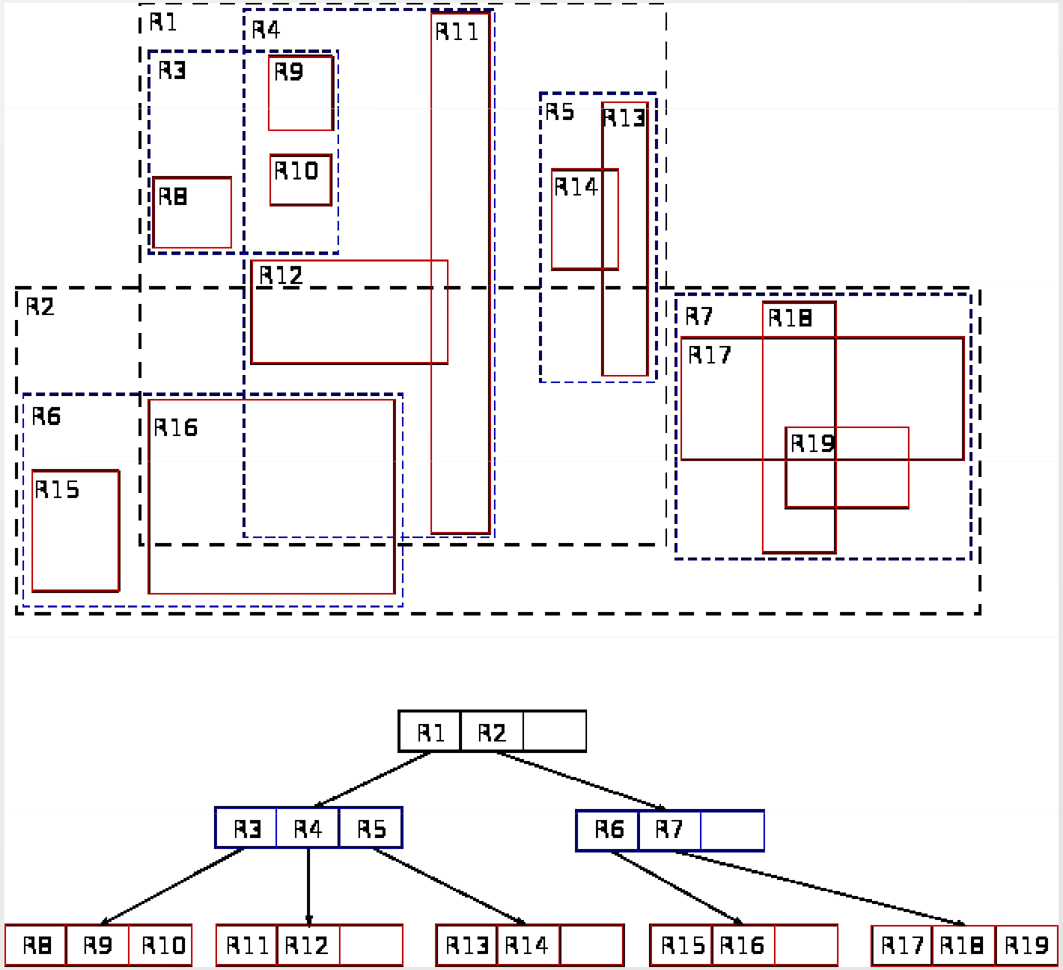
$$D_f(p,q) = f(p) - f(q) - \langle \nabla f(q), p - q \rangle$$

Pre nejaky query point budeme vždy počitat D(x,q) nie D(q,x)

Použijeme nejakú štruktúru, ktorá využíva bounding rectangle
Napr. R-Strom, VA File

R-Strom

Vychádza z B-Stromu
Vhodný i pre plošné data



Algoritmus 1KNN:
Vstup: R-tree T, Query q
Výstup: Set S
1: Set the k-nearest neighbor set S as empty
2: Set threshold distance θ = ∞
3: Clear a priority queue Q
4: Enqueue the root of T into Q
5: while Q is not empty do
6: Dequeue the head node N from Q
7: if N is a leaf node then
8: for each point p stored in N do
9: if Df (p,q) <θ then
10: Insert p into S and update θ
11: else
12: for each child node M of N do
13: Retrieve the MBR R of M
14: if LB(R,q) < θ then
15: Enqueue M into Q with LB(R,q)
16: Return points in S

http://gis.umb.no/gis/applets/rtree2/jdk1.1/

Ako zistit LB(R, q)

Každé x z množiny S^+ je pokryté MBR R len ak $LR_i \leq x_i \leq Ur_i$ a zároveň $LR_{i+1} \leq x_{i+1} \leq Ur_{i+1}$

$$D_f(x,q) = f(x) - f(q) - \langle \nabla f(q), x - q \rangle$$

$$= f(x) - f(q) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot (x_i - q_i) \right)$$

$$= x_{d+1}^{+} - f(q) - \sum_{i=1}^{d} \left(\nabla f'(q_i) \cdot (x_i^{+} - q_i) \right)$$

$$= x_{d+1}^{+} - f(q) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot x_i^{+} \right) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot q_i \right)$$

$$\leq Rl_{d+1} - \sum_{i=1}^{d} \left(\max\left(\left(\nabla f(q_i) \cdot Rl_i \right), \left(\nabla f(q_i) \cdot Ru_i \right) \right) \right) - \left(f(q) + \sum_{i=1}^{d} \nabla f(q_i) \cdot q_i \right)$$

Ako zistit UB(R, q)

Každé x z množiny S^+ je pokryté MBR R len ak $LR_i \leq x_i \leq Ur_i$ a zároveň $LR_{i+1} \leq x_{i+1} \leq Ur_{i+1}$

$$
\begin{aligned}
D_f(x,q) &= f(x) - f(q) - \langle \nabla f(q), x - q \rangle \\
&= f(x) - f(q) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot (x_i - q_i) \right) \\
&= x_{d+1}^{+} - f(q) - \sum_{i=1}^{d} \left(\nabla f'(q_i) \cdot (x_i^{+} - q_i) \right) \\
&= x_{d+1}^{+} - f(q) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot x_i^{+} \right) - \sum_{i=1}^{d} \left(\nabla f(q_i) \cdot q_i \right)
\end{aligned}
$$

$$
\leq Rl_{d+1} - \sum_{i=1}^{d} \left(\min\left((\nabla f(q_i) \cdot Rl_i), (\nabla f(q_i) \cdot Ru_i) \right) \right) - \left(f(q) + \sum_{i=1}^{d} \nabla f(q_i) \cdot q_i \right)
$$

$$D_f(x,q) \leq Rl_{d+1} - \sum_{i=1}^{d} (\max((\nabla f(q_i) \cdot Rl_i), (\nabla f(q_i) \cdot Ru_i))) - \left(f(q) + \sum_{i=1}^{d} \nabla f(q_i) \cdot q_i \right)$$

$f(q) + \sum_{i=1}^{d} \nabla f(q_i) \cdot q_i$ - Vypočítame raz na začiatku vyhľadávania

Na to, aby sme mohli vyradit z vyhladavania cely MBR staci, ak LB je väčšie ako θ. To či je väčšie dokážeme vypočítať v O(d) krokoch, pretoze stačí pre každú dimenziu nájsť menšiu hodnotu medzi $\nabla f(q_i) \cdot Rl_i$ a $\nabla f(q_i) \cdot Ru_i$.

Děkuji za pozornost.

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

EVROPSKÁ UNIE
MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY
OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ