Comparing Efficient Nearest Neighbor Techniques for Support Vector Machines in Large Feature Spaces

Diana Carolina Montañez, Adolfo Quiroz, Mateo Dulce, Alvaro J. Riascos Villegas

February de 2018

Contenido

Introduction

- We address the problem of efficiently finding nearest neighbors in order to train a support vector machine classifier in a large feature space.
- SVM in a nutshell.

Introduction

- We address the problem of efficiently finding nearest neighbors in order to train a support vector machine classifier in a large feature space.
- SVM in a nutshell.

Contenido

Algortihms

- We follow Camelo et.al strategy for solving for SVs. using K-NN.
- Solving the K-NN query is costly in large feture space. We follow a pivot strategy and Chavez, et.al. proximity preserving order algorithm.
- We refine the pivot strategy by choosing appropiate (base) pivots following Micó et.al.
- We experimentally explore the relative merits of the pivot strategy against the more refined strategy and the state of the art benchmark: LIBSVM.

Algortihms

- We follow Camelo et.al strategy for solving for SVs. using K-NN.
- Solving the K-NN query is costly in large feture space. We follow a pivot strategy and Chavez, et.al. proximity preserving order algorithm.
- We refine the pivot strategy by choosing appropiate (base) pivots following Micó et.al.
- We experimentally explore the relative merits of the pivot strategy against the more refined strategy and the state of the art benchmark: LIBSVM.

Algortihms

- We follow Camelo et.al strategy for solving for SVs. using K-NN.
- Solving the K-NN query is costly in large feture space. We follow a pivot strategy and Chavez, et.al. proximity preserving order algorithm.
- We refine the pivot strategy by choosing appropiate (base) pivots following Micó et.al.
- We experimentally explore the relative merits of the pivot strategy against the more refined strategy and the state of the art benchmark: LIBSVM.

Algortihms

- We follow Camelo et.al strategy for solving for SVs. using K-NN.
- Solving the K-NN query is costly in large feture space. We follow a pivot strategy and Chavez, et.al. proximity preserving order algorithm.
- We refine the pivot strategy by choosing appropiate (base) pivots following Micó et.al.
- We experimentally explore the relative merits of the pivot strategy against the more refined strategy and the state of the art benchmark: LIBSVM.

Camelo, et.al

- Finding SV is costly.
- They propose solving for SV on small subsamples and look to K-NN of SV.
- They demonstrate two theorems that rationalizes their sampling strategy.

Camelo, et.al

- Finding SV is costly.
- They propose solving for SV on small subsamples and look to K-NN of SV.
- They demonstrate two theorems that rationalizes their sampling strategy.

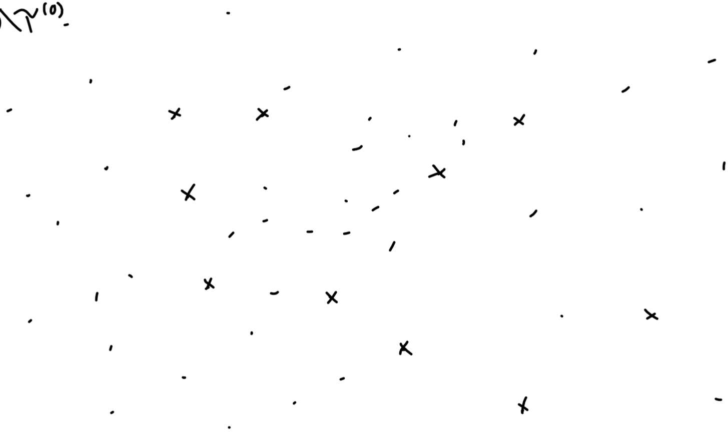
Camelo, et.al

- Finding SV is costly.
- They propose solving for SV on small subsamples and look to K-NN of SV.
- They demonstrate two theorems that rationalizes their sampling strategy.

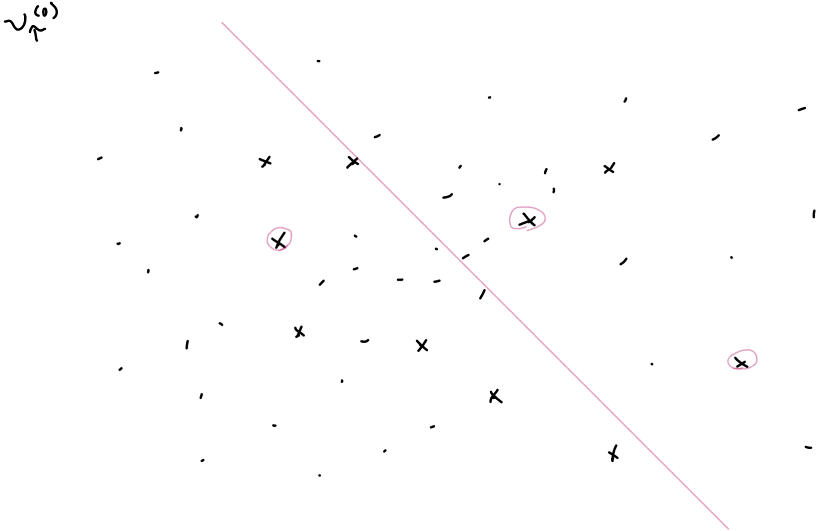
Camelo, et.al

$\gamma^{(0)} = \delta D$

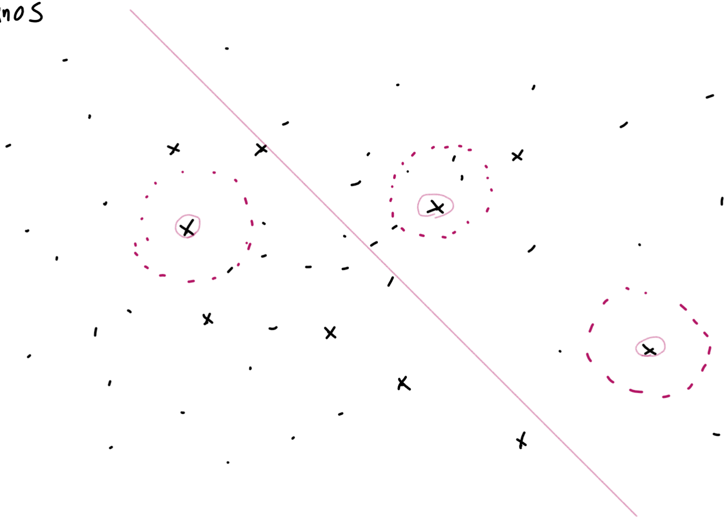
$S^{(0)} = D \setminus \gamma^{(0)}$.

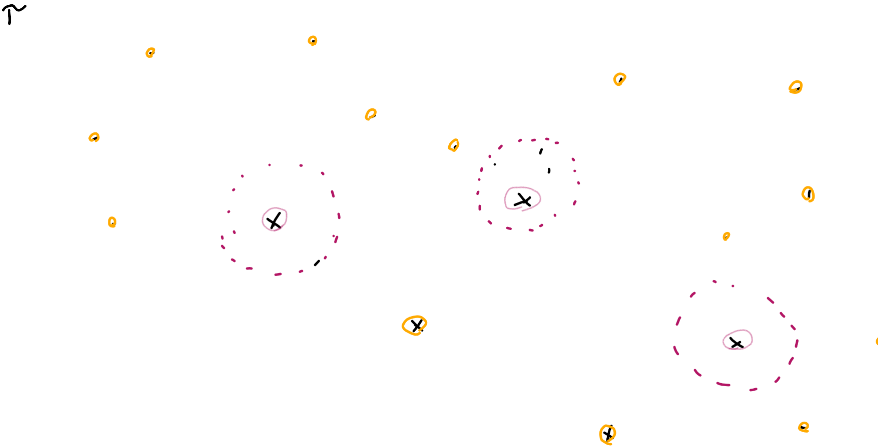


Camelo, et.al



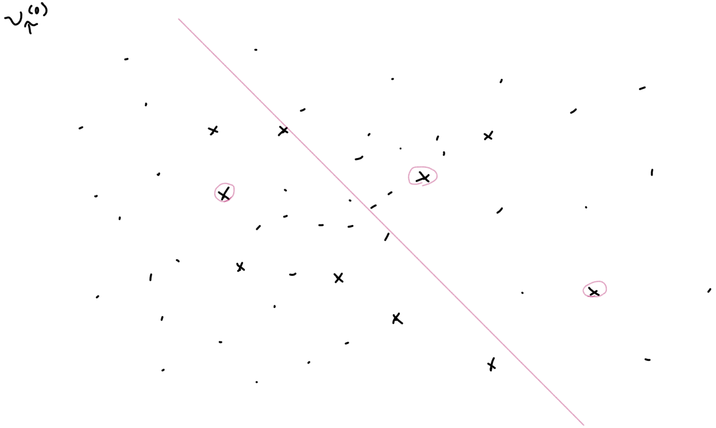
K-Vecinos




τ

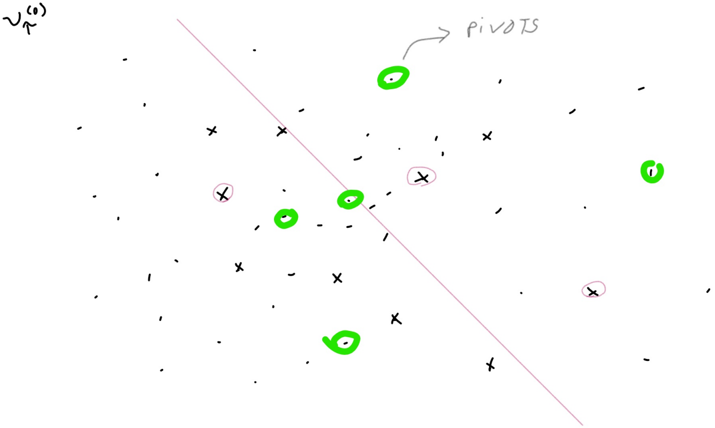
Chavez, et.al

- K-NN queries are expensive specially in high dimensional feature spaces.



Chavez, et.al

- Pivot strategy:

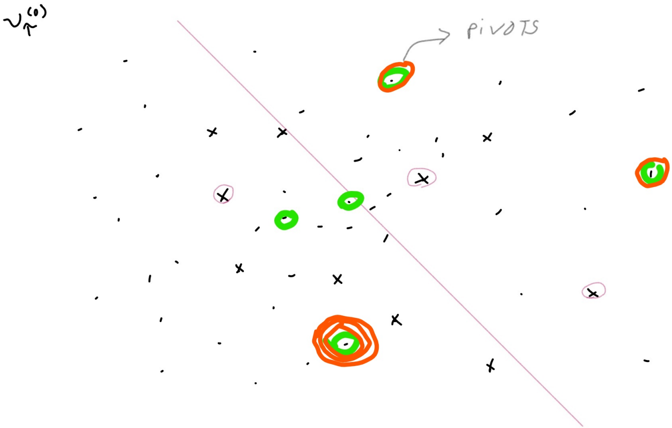


pivots.jpg

- Proximity preserving order:
  1. Then, for each $d \in \mathcal{D}$, the distance to $p_i \in \mathbb{P}$ is calculated. For each d, the p_i are ordered according to their distance to d. Therefore we have for each d a permutation π_d of the elements of $\mathbb{P}$.
  2. *Spearman's Footrule*. Let $\pi_d(p_i)$ be the position of p_i in π_d, then
  $$F(\pi_d, \pi_{d'}) = \sum_{1 \leq i \leq k} |\pi_d(p_i) - \pi_{d'}(p_i)|$$
  3. Example: $\pi_d = (p_2, p_1, p_3), \pi_{d'} = (p_1, p_3, p_2)$ then $F(\pi_d, \pi_{d'}) = |(2-1)| + |(1-3)| + |(3-2)|$

Micó, et.al

- Random vrs. well separated pivots



base pivots.jpg

Contenido

Data

Table 1. Dataset description.

Daraset	Train size	Test size	Features
real-sim	16,525	4,131	14,462
COM	8,205	2,172	9,892
CC	8,224	2,153	9,906

Results