Обобщения SVM

Дубов Дмитрий,
БПМИ141

Рассмотрим:

- One-Class SVM
- Multiclass Vector Machines

One-Class SVMs for Document Classification

One-class classification

В чем отличие?

- Input:
  - Объекты одного класса (positive information)
  - Объекты одного класса + unlabeled объекты
- Отделяет объекты одного класса от других объектов
- Примеры задач:
  - Определить статус состояния атомной электростанции как “нормальный”
  - Детекция МикроРНК генов
  - Определить интересы веб-пользователя
  - определение аномалий в данных

Schoelkopf method (1999)

Вход: $x_1, \ldots, x_l \in X, X \in R^n$

Пусть ***X*** имеет вероятностное распределение ***P*** на признаковом пространстве.

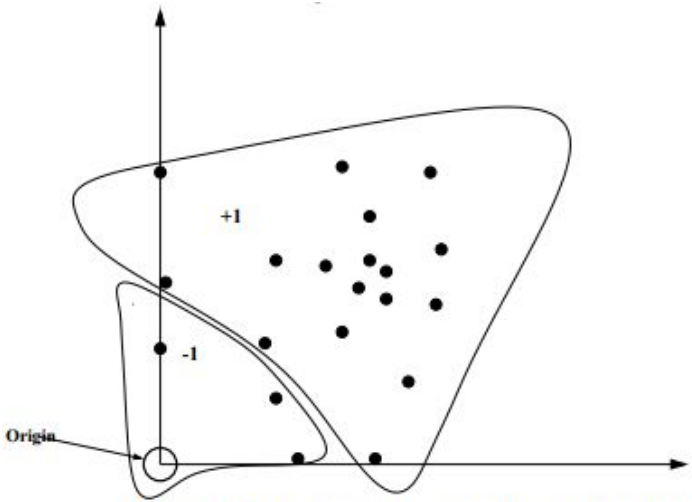
Задача: найти такое подмножество ***S*** пространства признаков, что *вероятность* для элемента из ***P*** попасть вне этого множества ограничена неким априорным значением ν .

Задача сводится к поиску такого алгоритма ***f(x)***:

$$f(x) = \begin{cases} +1 & \text{if } x \in S \\ -1 & \text{if } x \in \overline{S} \end{cases}$$

Алгоритм:

- Преобразовать данные в новое признаковое пространство ***H*** (используя ядро)
- В новом пространстве максимизировать отступ между объектами и началом координат



One-Class SVM Classifier. The origin is the only original member of the second class.

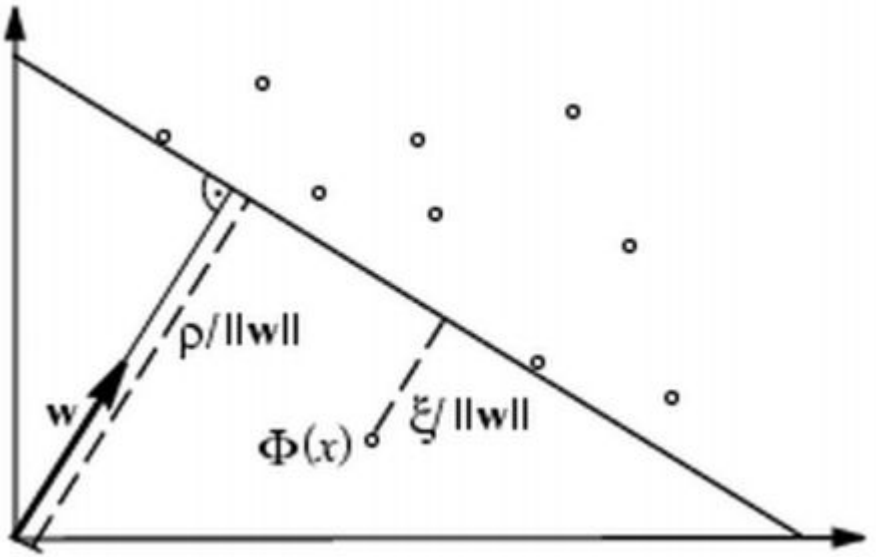
Определим ядровую функцию $\Phi : X \longrightarrow H$, и скалярное произведение $k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}}$

Решаем квадратичную задачу:

$$min\frac{1}{2}\|w\|^2 + \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho$$

$$(w \cdot \Phi(x_i)) \geq \rho - \xi_i \quad i = 1, 2, ..., l \quad \xi_i \geq 0$$

Гиперплоскость: $\langle \mathbf{x}, \mathbf{\Phi}(x) \rangle = \rho$



Параметр ν

Schoelkopf показывает, что ν :

1. Является верхней границей на долю объектов, вышедших за границу.
2. Нижней границей доли объектов, использующихся как опорные вектора.

По сути - данный параметр контролирует баланс между максимизацией отступа и попаданием объектов в множество ***S***

Двойственная задача

Выглядит так:

$$\min_{\boldsymbol{\alpha}} \frac{1}{2} \sum_{ij} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$

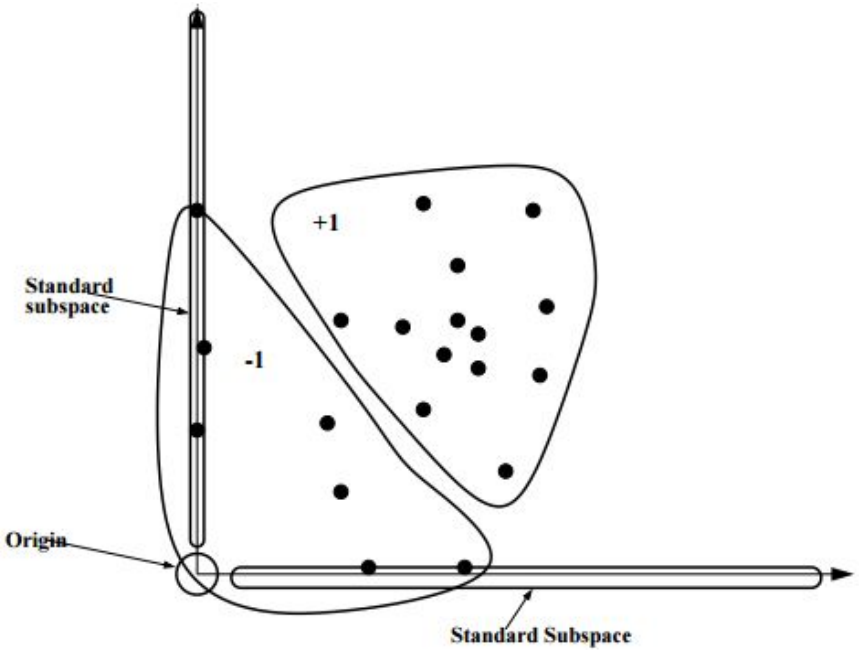
$$0 \leq \alpha_i \leq \frac{1}{\nu \ell}, \quad \sum_i \alpha_i = 1.$$

А итоговый алгоритм так:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_i \alpha_i k(\mathbf{x}_i, \mathbf{x}) - \rho \right)$$

Outlier SVM

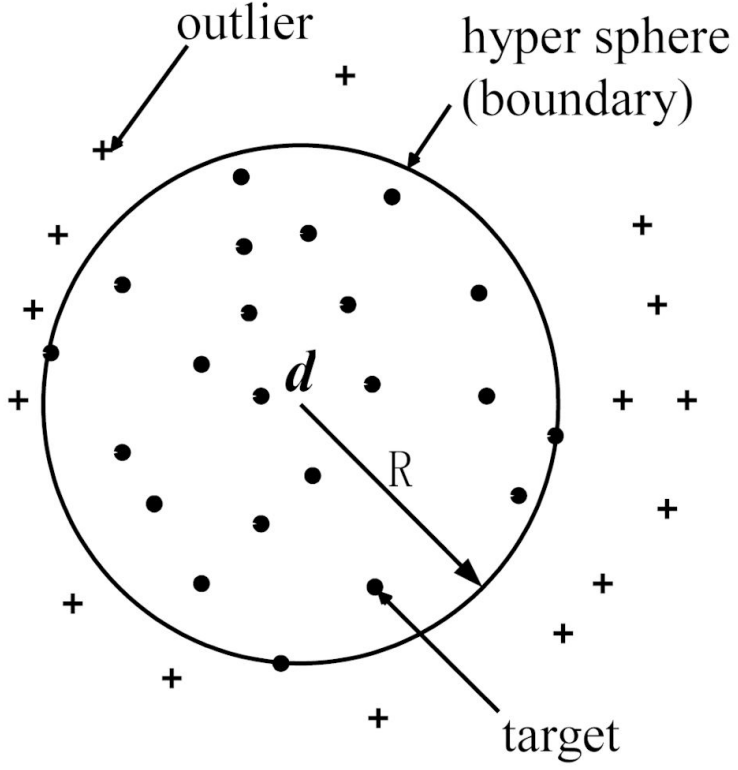
- Найти точки “достаточно близкие” к началу координат
- Отнести их классу “-1”
- Запустить обычный двухклассовый SVM



Outlier SVM Classifier. The origin and small subspaces are the original members of the second class. The diagram is conceptual only.

SVDD

Пусть a — некоторая точка в пространстве образов функции φ(·), R — некоторая положительная величина. Будем считать, что точка x принадлежит нормальному классу, если она лежит $\|a-\phi(x)\|_{\ell_2}^2 \le R$:



Задача:

$$R+\frac{1}{\nu l}\sum_{i=1}^{l}\xi_i \to \min_{R,a,\xi}$$

$$s.t.\ \|\phi(x_i)-a\|_{\ell_2}^2 \le R+\xi_i,$$

$$\xi_i \ge 0.$$

Двойственная задача и решение

Задача:

$$\sum_{i=1}^{l} \alpha_i K(x_i, x_i) - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \alpha_j K(x_i, x_j) \to \max_{\alpha}$$

$$s.t.\ 0 \le \alpha_i \le \frac{1}{\nu l},$$

$$\sum_{i=1}^{l} \alpha_i = 1.$$

Решающее правило:

$$f(x) = K(x,x) - 2\sum_{i=1}^{l} \alpha_i K(x, x_i) + \|a\|^2_{\ell_2} - R.$$

Data

Данные были взяты из Reuters dataset

Document Representations:

1. Binary
2. Frequency
3. Tf-idf
4. Hadamard

Алгоритмы для сравнения

1. One-class SVM
2. Outlier-SVM
3. Prototype (Rocchio's) algorithm
4. Nearest Neighbour (modification for one-class)
5. Naive Bayes
6. Compression Neural Network

Результаты

	One-class SVM Radial Basis F_1	Outlier-SVM Linear F_1	Neural Networks F_1	Naive Bayes F_1	Nearest Neighbor F_1	Prototype F_1
Earn	0.676	0.750	0.714	0.708	0.703	0.637
Acq	0.482	0.504	0.621	0.503	0.476	0.468
Money	0.514	0.563	0.642	0.493	0.468	0.484
Grain	0.585	0.523	0.473	0.382	0.333	0.402
Crude	0.544	0.474	0.534	0.457	0.392	0.398
Trade	0.597	0.423	0.569	0.483	0.441	0.557
Int	0.485	0.465	0.487	0.394	0.295	0.454
Ship	0.539	0.402	0.361	0.288	0.389	0.370
Wheat	0.474	0.389	0.404	0.288	0.566	0.262
Corn	0.298	0.356	0.324	0.254	0.168	0.230
Avg	0.519	0.484	0.513	0.425	0.423	0.426
Macro	0.572	0.587	0.615	0.547	0.530	0.516

Multiclass Vector Machine

Multiclass classification

- OVA
- AVA
- DAGSVM
- Multi-class SVM

Direct Multi-Class SVM

1. Weston and Watkins' Multi-Class SVM
2. Crammer and Singer's Multi-Class SVM

Weston and Watkin's

Постановка задачи:

$$\min_{\mathbf{f_1},\ldots,\mathbf{f_N}\in\mathcal{H},\xi\in\mathbf{R}^{\ell(\mathbf{N}-\mathbf{1})}} \quad \sum_{i=1}^{N} ||f_i||_K^2 + C\sum_{i=1}^{\ell}\sum_{j\neq y_i}\xi_{ij}$$

$$\text{subject to :} \quad f_{y_i}(\mathbf{x_i}) + b_{y_i} \geq f_j(\mathbf{x_i}) + b_j + 2 - \xi_{ij}$$

$$\xi_{ij} \geq 0$$

Хотим:

$$f_{y_i}(\mathbf{x_i}) - f_j(\mathbf{x_i}) \geq 2$$

Иначе штраф

- На 2 из 5 датасетах был значительно лучше чем OVA, на остальных примерно так же
- На самом деле не понятно что лучше
- Гораздо дольше (N * I - переменных в задаче) чем обычный SVM

Crammer and Singer's

Постановка задачи:

$$\min_{\mathbf{f_1},\ldots,\mathbf{f_N}\in\mathcal{H},\xi\in\mathbf{R}^{\ell}} \quad \sum_{i=1}^{N} \|f_i\|_K^2 + C\sum_{i=1}^{\ell} \xi_i$$

$$\text{subject to :} \quad f_{y_i}(\mathbf{x_i}) \geq f_j(\mathbf{x_i}) + 1 - \xi_i$$

$$\xi_i \geq 0$$

Основное отличие:

Для каждого объекта - суммируется только одно отклонение (наибольшее), вместо N-1 у WW.

- Очень эффективные вычисления
- Много памяти
- Быстрее OVA
- Немного лучше в точности

Заключение

Спасибо за внимание!

References

http://www.jmlr.org/papers/volume2/manevitz01a/manevitz01a.pdf

https://papers.nips.cc/paper/1723-support-vector-method-for-novelty-detection.pdf

http://itas2016.iitp.ru/pdf/1570285426.pdf

http://www.dainf.ct.utfpr.edu.br/~kaestner/Mineracao/ArquivosExtras2016/Dan.Nick-OneClassSVM.pdf

https://www.csie.ntu.edu.tw/~cjlin/papers/multisvm.pdf