Московский Государственный Университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики

Метод опорных векторов

Шаталов Н.А
кафедра ММП, группа 317
2017

Содержание

1 Введение

В данном отчете описаны результаты исследования метода опорных векторов. Рассматриваются следующие методы для решения задачи SVM:

1. Метод внутренней точки (МВТ) для прямой задачи.
2. Метод внутренней точки (МВТ) для двойственной задачи с поддержкой перехода к полиномиальному и RBF ядру.
3. Метод PEGASOS для решения задачи без ограничений.
4. Метод субградиентного спуска и стохастического субградиентного спуска для решения задачи без ограничения.

2 Сравнение различных методов

Посмотрим как меняется время работы методов в зависимости от характеристик обучающей выборки: числа объектов в обучающей выборке и размерности признакового пространства. Эксперимент проводится на линейно неразделимых данных (но хорошо классифицируемых). Для исследования зависимости времени работы метода от числа объектов в обучающей выборке возьмем 200 признаков, чтобы было видно разницу между разными случаями.

	Размер выборки			
Стратегия	500	1000	2000	5000
МВТ (прямая задача)	0.26	1.22	7.58	97.35
МВТ (двойственная задача, без ядра)	0.14	0.89	7.03	85.18
МВТ (двойственная задача, полиномиальное ядро)	0.10	0.50	3.55	55.43
МВТ (двойственная задача, RBF ядро)	0.48	2.48	17.14	129.98
Метод PEGASOS	1.19	1.14	1.13	1.14
Субградиентный спуск	2.10	5.08	18.96	45.10
Стохастический субградиентный спуск	0.61	0.55	0.63	0.50

Таблица 1: Время работы метода в зависимости от размер обучающей выборки (с)

Для субградиентных методов было взято одинаковое максимальное количество итераций. Количество итераций в методе PEGASOS тоже взято одинаково, поэтому для каждого случая он работает почти одинаково по времени.
Как видно из таблицы 1, дольше всех работает метод внутренней точки для прямой задачи и метод внутренней точки для двойственной задачи с RBF ядром. Для прямой задачи это объясняется тем, что в этом методе количество оптимизируемых параметров больше чем в методах для двойственной задачи. Для двойственной задачи с rbf ядром это можно связать с долгой операцией взятия поэлементной экспоненты матрицы расстояний между объектами обучающей выборки.

Для исследования зависимости времени работы метода от числа объектов в обучающей выборке возьмем 2000 объектов в обучающую выборку.

	Размерность			
Стратегия	100	200	500	1000
МВТ (прямая задача)	7.28	8.14	11.32	15.55
МВТ (двойственная задача, без ядра)	7.08	6.95	7.32	7.07
МВТ (двойственная задача, полиномиальное ядро)	5.37	3.35	5.64	4.97
МВТ (двойственная задача, RBF ядро)	20.61	10.95	19.11	34.71
Метод PEGASOS	0.76	1.26	2.34	5.68
Субградиентный спуск	7.30	0.22	42.79	74.33
Стохастический субградиентный спуск	0.44	0.55	0.85	1.23

Таблица 2: Время работы метода в зависимости от количества признаков (с)

Как видно из таблицы 2, при увеличении размерности признакового пространства увеличивается время работы метода внутренней точки прямой задачи, так как только в этом методе количество параметров пропорционально количеству признаков в отличие от метода внутренней точки для двойственной задачи.

Посмотрим на значение целевой функции в различных методах.

Стратегия	Целевая функция
МВТ (прямая задача)	0.832
МВТ (двойственная задача, без ядра)	-0.832
МВТ (двойственная задача, полиномиальное ядро)	-0.001
МВТ (двойственная задача, RBF ядро)	-0.998
Метод PEGASOS	0.833
Субградиентный спуск	0.841
Стохастический субградиентный спуск	0.947

Таблица 3: Значение целевой функции в различных методах

В таблице 3 можно заметить, что у двойственной задачи получились отрицательные значения целевой функции. Это потому что в двойственной задаче решается задача максимизации. Как видно, значение функционала линейных методов внутренней точки для прямой и двойственной задачи, метода PEGASOS и субградиентного спуска похожи, так как решается одна задача. При добавлении ядра (полиномиального, rbf) происходит оптимизация уже другого функционала.

Посмотрим на скорость сходимости субградиентных методов и метода PEGASOS. Исследовать будем на линейно неразделимой выборке из 2000 объектов.

Как видно из рисунка 1, полный субградиентный спуск практически сразу дает оптимальное решение, так как длина обучающей выборки большая. Метод PEGASOS работает заметно быстрее и устойчивее, чем стохастический субградиентный спуск, который стабилизируется очень медленно.

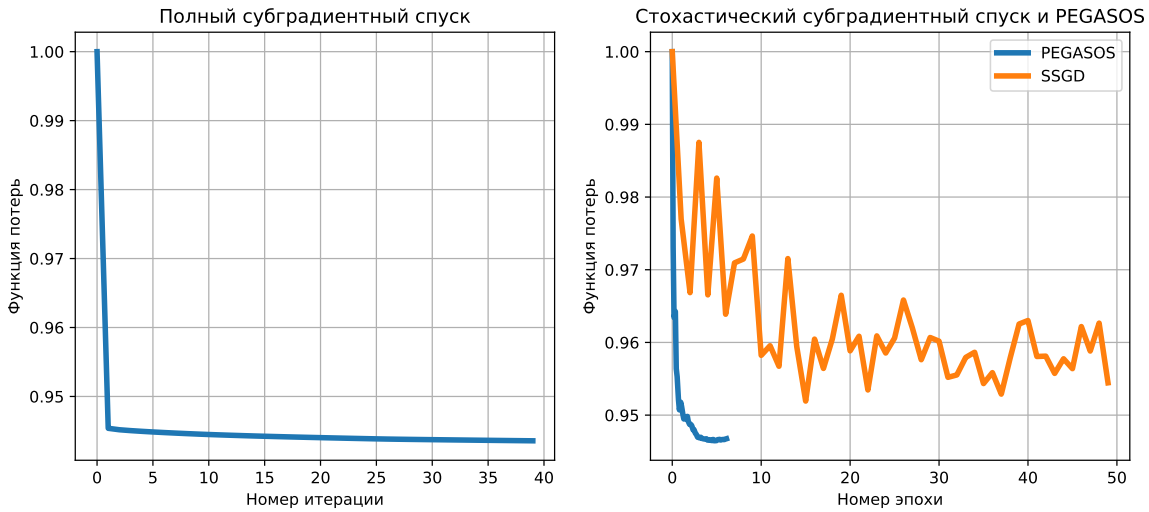

Рис. 1: Значения функции потерь в зависимости от номера итерции или эпохи

3 Подбор параметров

Параметры для методов будем подбирать скользящим контролем на трех фолдах. Подбор будем производить на линейно неразделимой, но хорошо классифицируемой выборке.

Подберем параметр C, который является общим в моделях и обозначает вес ошибок объектов при обучении модели. Подбор осуществляем на линейном методе внутренней точки для двойственной задачи.

C	1	**5**	10	25	100
Точность	0.7695	**0.8895**	0.8797	0.7701	0.7602

Таблица 4: Точность метода на скользящем контроле в зависимости от параметра C

Подберем параметр d на методе внутренней точки для двойственной задачи с переходом к полиномиальному ядру. Параметр d обозначает соответствующую степень ядра.

d	2	3	**4**	5	6	7
Точность	0.8595	0.7704	**0.8705**	0.8304	0.8304	0.8503

Таблица 5: Точность метода на скользящем контроле в зависимости от параметра d

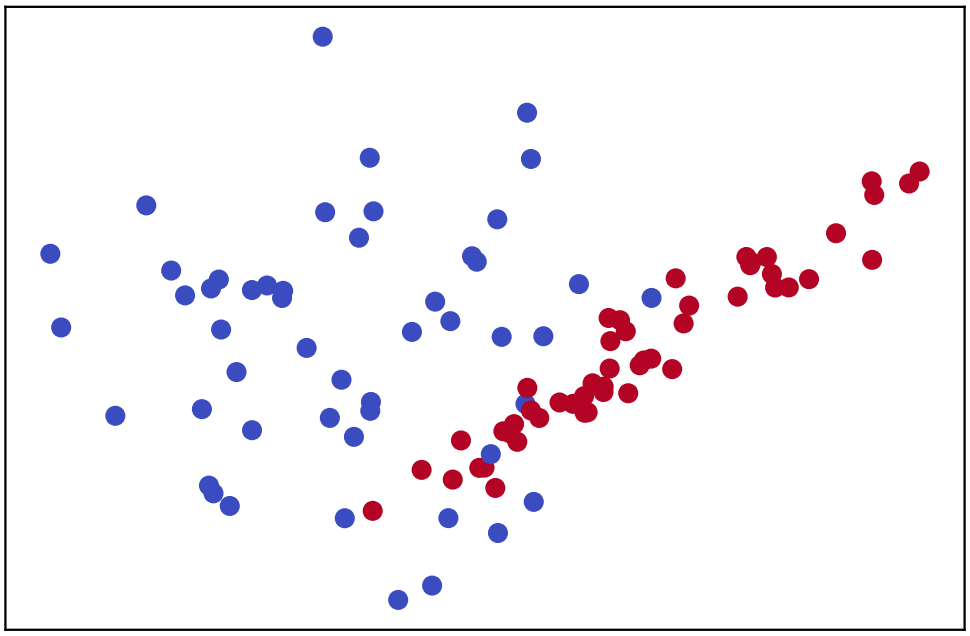

Рис. 2: Визуализация выборки, на которой производился подбор параметров

Подберем параметр γ на методе внутренней точки для двойственной задачи с переходом к RBF ядру. Параметр γ обозначает соответствующую ширину RBF окна ядра.

γ	1	10	**25**	50	100
Точность	0.8996	0.9100	**0.9299**	0.8996	0.8898

Таблица 6: Точность метода на скользящем контроле в зависимости от параметра γ

Теперь опробуем подобрать параметр C на линейно разделимой выборке.

γ	1	10	25	50	100
Точность	1	1	1	1	1

Таблица 7: Точность метода на скользящем контроле на линейно разделимой выборке

Как видно в таблице 7 такая выборка не подходит для подбора параметров, так как алгоритм во всех случаях строит идеально разделяющую поверхность.

4 Визуализация линейного SVM и SVM с RBF ядром

Построим разделяющую поверхность линейным SVM и SVM с RBF ядром в двумерном случае. Опорные векторы на рисунках отмечены крестиками.

На рисунке 3 видно, как влияет параметр γ на работу метода. При увеличении параметра количество опорных векторов вблизи границы становится больше, соответственно свой вклад

вносят большее количество векторов, что дает более грубую разделяющую поверхность. Таким образом при большом γ обобщающая способность метода уменьшается.

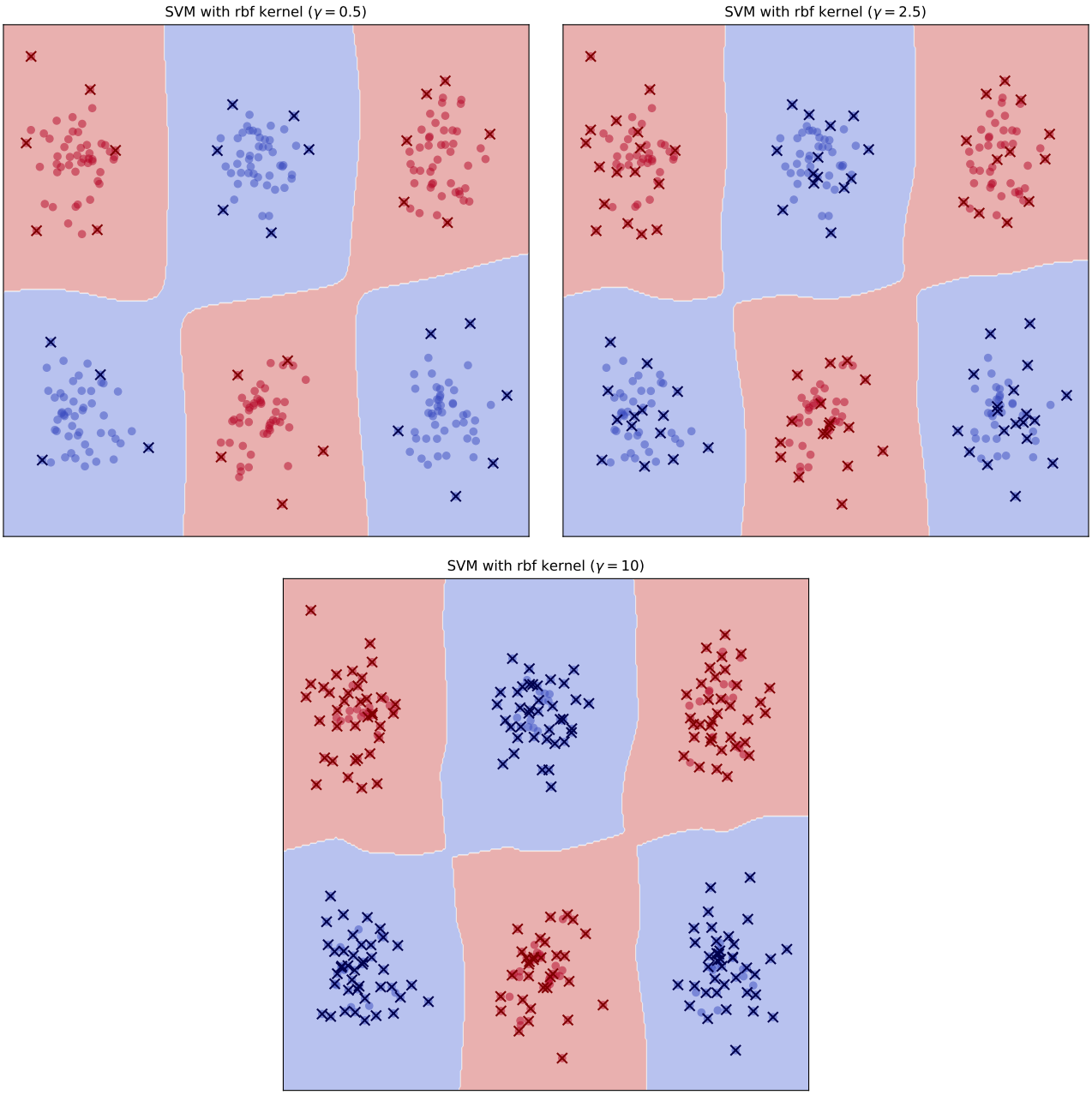

Рис. 3: Разделяющие поверхности для SVM с RBF ядром с различным γ

На рисунке 4 можно увидеть как на этой же выборке будет работать линейная модель SVM.

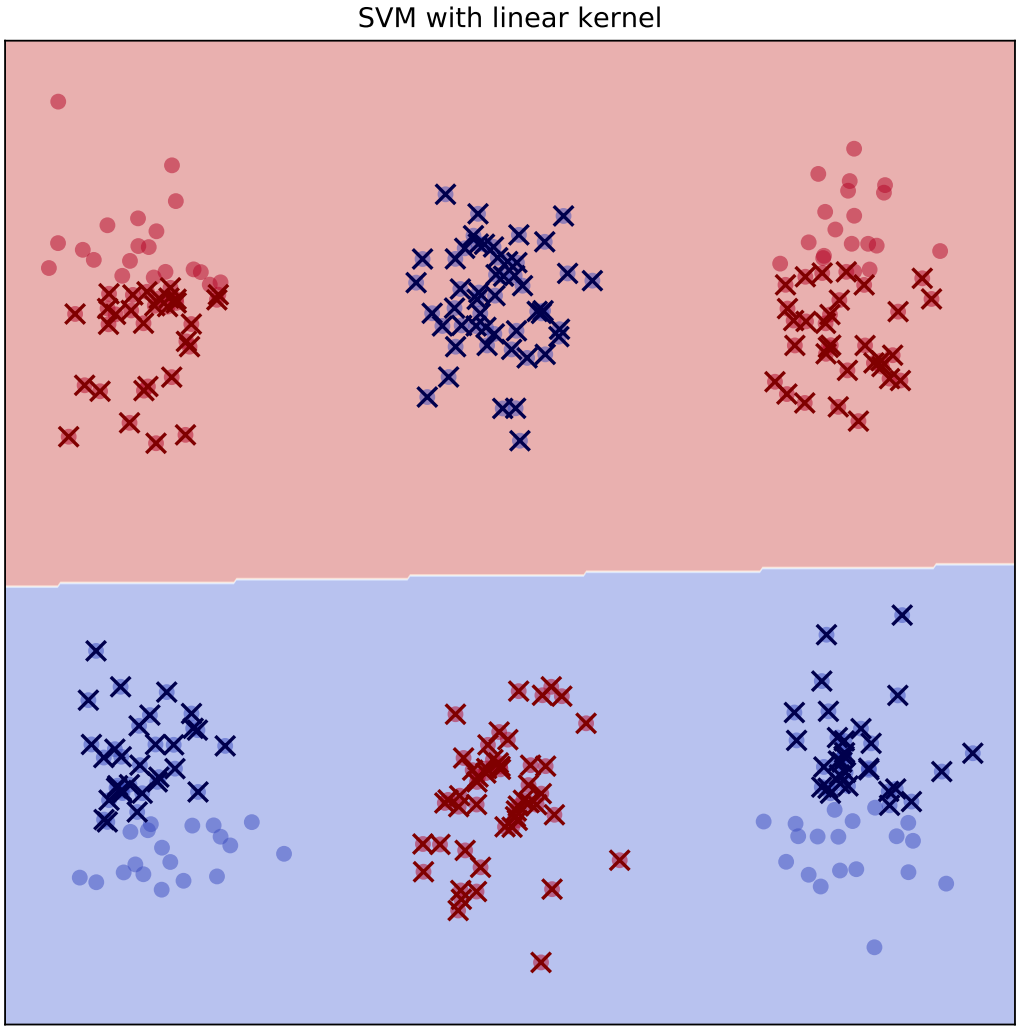

Рис. 4: Разделяющие поверхности для линейного SVM