МОСКОВСКИЙ ИНСТИТУТ ЭЛЕКТРОННОЙ ТЕХНИКИ
Институт системной и программной инженерии
и информационных технологий (Институт СПИНТех)

Лабораторный практикум по курсу "Интеллектуальные системы" (09/23 – 01/24)

Лабораторная работа 3
Статистические методы обучения. Метод опорных векторов [1].

На этом занятии компьютерного практикума Вы изучите *метод опорных векторов* (англ. SVM, *Support Vector Machine*) и примените данный метод для решения задачи классификации различных двумерных наборов данных. В последних публикациях на русском языке метод называется также *машинами поддерживающих векторов*, или, в более общем смысле, ядерными машинами (англ. *Kernel Machine*). В методах, основанных на их использовании, предусмотрен эффективный механизм обучения, а сами они позволяют представить сложные, нелинейные функции.

Ядерные машины превосходят все другие способы распознавания рукописных символов, в частности цифр; кроме того, они быстро находят применение и в других приложениях, особенно в тех, которые отличаются большим количеством входных характеристик.

Прежде чем приступить, собственно, к программированию, настоятельно рекомендуется ознакомиться с материалом лекций, а также с дополнительными материалами, имеющими отношение к задачам классификации.

Файлы, включенные в задание:

lab_intelligent_systems_SVM – ноутбук, реализующий пошаговое выполнение первой части задания по разделу «Статистические методы обучения. Метод опорных векторов»;
ex3data1.mat – 1-й набор данных;
ex3data2.mat – 2-й набор данных;
ex3data3.mat – 3-й набор данных;
svmTrain – функция обучения для метода опорных векторов;
* *plotData* – функция графического отображения данных;
visualizeBoundaryLinear – отображение линейной границы раздела данных;
visualizeBoundary – отображение нелинейной границы раздела данных;
linearKernel – ядро линейного отображения;
* *gaussianKernel* – ядро Гаусса для метода опорных векторов;
* *dataset3Params* – Параметры для 3-го набора данных

[1] Материал лабораторной работы (в Matlab варианте) составлен на основании аналогичного задания по курсу «Машинное обучение» на портале он-лайн обучения Coursera.org (профессор Эндрю Ын, Стэнфордский университет - https://ru.wikipedia.org/wiki/Ын,_Эндрю)

Функции, отмеченные знаком *, следует написать самостоятельно.

В этой Лабораторной работе следует использовать ноутбук *lab_intelligent_systems_SVM*. В ноутбуке подготовлены обращения к исходным данным. Далее производится вызов функций, написанных Вами, и отображаются результаты вычислений. Необходимо дописать функции по инструкциям упражнения.

1 Метод опорных векторов

Как было сказано выше, Вы будете использовать метод опорных векторов для классификации 2-мерных наборов данных. В ходе работы над упражнением вы изучите собственно метод, а также научитесь использовать с SVM ядро Гаусса.

Указание: Вы располагаете также написанными ранее программами, которые здесь можно, в случае необходимости, использовать.

1.1 Набор данных 1

Начнем работу с обработки первого набора данных, который может быть разделен с помощью линейной границы. Ноутбук автоматически построит график, отображающий эти данные (рис. 1). На визуализированной картине данных видно естественное разделение между «положительными» примерами (обозначенными как «+») и «отрицательными» (обозначенными как «o»). Обратите внимание на присутствие «положительного» примера (*x = 0.1; y = 4.1*), расположенного вдали от области расположения основного числа «положительных» примеров. Выполняя задание в этой части упражнения, можно понять, как такое расположение влияет на проведение прямолинейной границы раздела 2-х областей в процессе применения метода опорных векторов.

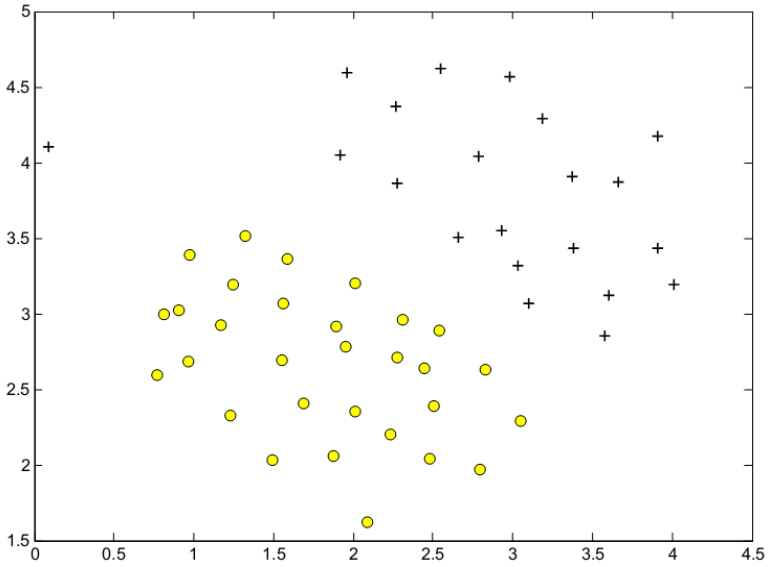

Рис. 1. Первый набор данных

Следует использовать различные значения параметра *C*, который принимает положительные значения, влияющие на значение штрафной функции при ошибочной классификации обучающего примера. Более высокие значения параметра *C* соответствуют требованию наиболее точной классификации *всех* обучающих примеров. В этом смысле этот параметр аналогичен обратным значением параметра регуляризации λ, который применялся в логистической регрессии.

Следующая часть ноутбука производит обучение алгоритма SVM (с параметром *C* = 1), при этом запускается код программы *svmTrain*. Если параметр *C* равен 1, исполнение кода метода

опорных векторов определит границу раздела 2-х областей в промежутке между ними, а также не классифицирует «+» значение, расположенное в самой левой части графика (рис. 2).

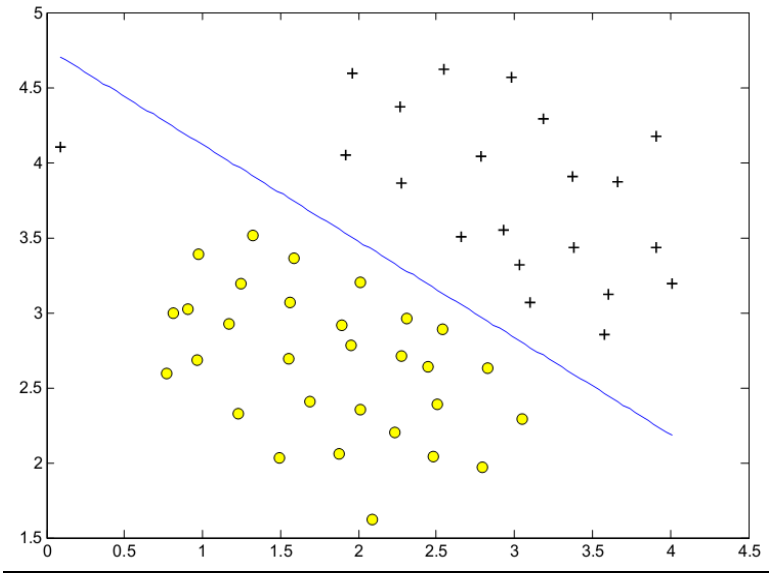

Рис. 2. Граница раздела при *C=1* (первый набор данных)

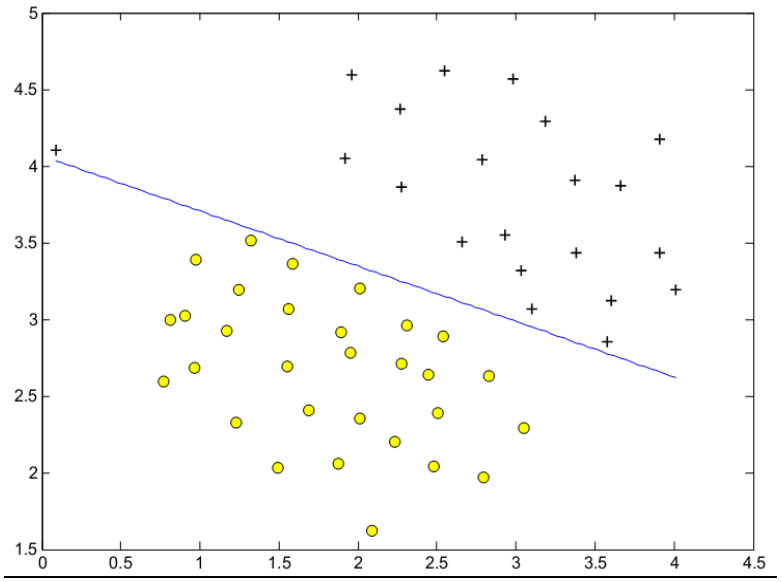

Рис. 3. Граница раздела при *C=1000* (первый набор данных)

Задание: Следует протестировать различные значения параметра *C* для представленного набора данных и оценить его влияние на качество классификации. В частности, если поменять значение параметра *C* в коде на 100 и запустить обучение SVM заново, то можно обнаружить, что метод опорных векторов классифицирует каждый пример корректно, но граница раздела не будет соответствовать «естественному» ожиданию для текущего набора данных (рис. 3).

1.2 Метод опорных векторов с ядром Гаусса

В этой части упражнения демонстрируется применение метода опорных векторов для нелинейной классификации данных. В частности, предстоит применить SVM с ядром Гаусса в ситуации, когда линейное разделение невозможно.

1.2.1 Ядро Гаусса

Для нахождения нелинейных границ с помощью метода опорных векторов необходимо запрограммировать функцию, реализующую применение ядра Гаусса. Под ядром Гаусса подразумевается функция, определяющая сходство пары образцов на основании оценки расстояния между ними *($x^{(i)}$, $x^{(j)}$)*. Ядро Гаусса регулируется параметром σ, который определяет, насколько быстро уменьшается «схожесть» двух примеров при увеличении расстояния между ними.

Теперь Вы можете завершить код программы *gaussianKerne*, требующийся для расчета ядра Гаусса (расстояния) между 2-мя примерами *($x^{(i)}$, $x^{(j)}$)*. Функция ядра Гаусса определена ниже:

$$K_{gaussian}(x^{(i)}, x^{(j)}) = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|^2}{2\sigma^2}\right) = \exp\left(-\frac{\sum_{k=1}^{n}(x_k^{(i)} - x_k^{(j)})^2}{2\sigma^2}\right)$$

Как только вы закончите написание программы *gaussianKernel*, ноутбук проверит Вашу функцию нахождения ядра на 2-х представленных примерах, в ответе Вы должны будете увидеть следующее значение: 0.324652.

1.2.2 Набор данных 2

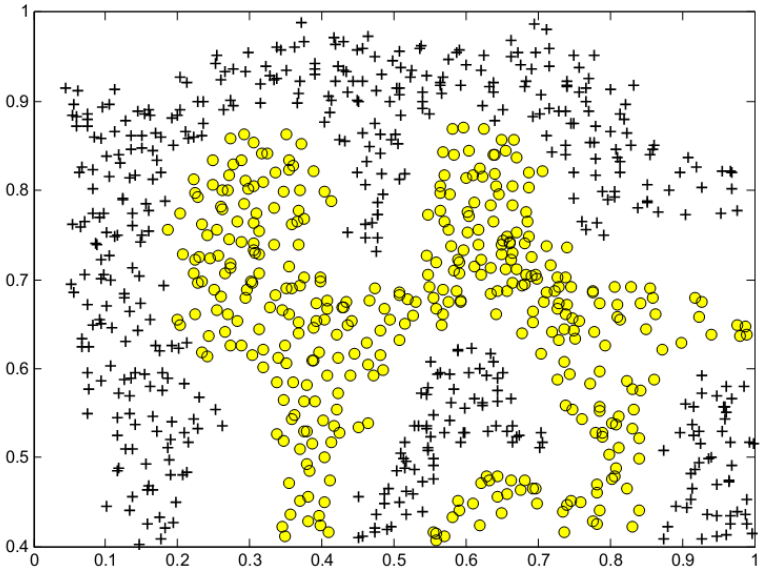

Рис. 4. Второй набор данных

Следующая часть ноутбука загрузит и отобразит набор данных для 2-й части упражнения. Нетрудно видеть, что невозможно провести линейную границу раздела, которая бы отделила «положительные» примеры от «отрицательные».

Задание: Используя SVM с ядром Гаусса, построить нелинейную границу раздела, которая наиболее точно подойдет для классификации предоставленного набора данных.

Если Вы правильно написали программу расчета ядра Гаусса, ноутбук продолжит обучение алгоритма, используя 2-й набор данных. На графике (рис. 5) изображена граница раздела 2-х областей, найденная с помощью метода опорных векторов с ядром Гаусса.

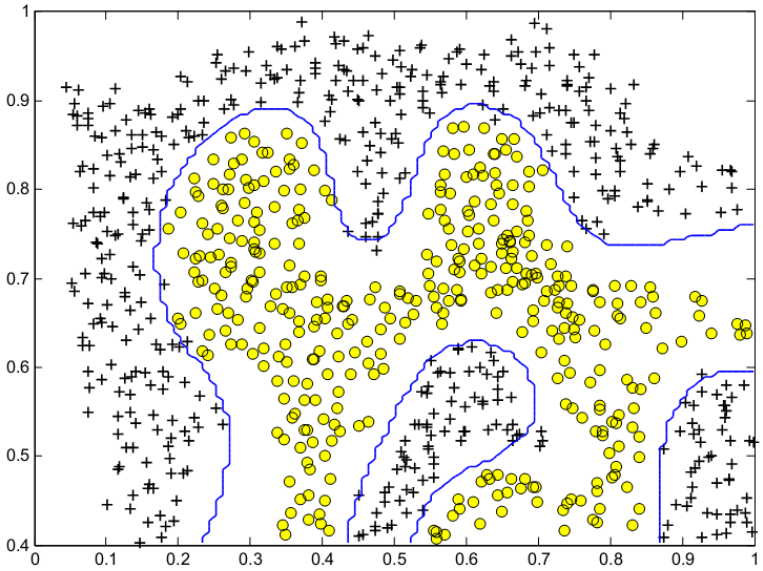

Рис. 5. Метод опорных векторов с ядром Гаусса,
граница раздела для второй выборки

1.2.3 Набор данных 3

В этой части упражнения Вы усовершенствуете свои навыки по использованию метода опорных векторов с ядром Гаусса для проведения нелинейной классификации данных. Следующая часть ноутбука загрузит и отобразит график с набором данных для этой части упражнения (рис. 6).

Исходные данные в файле *ex3data3.mat* описываются переменными *X, y, Xval, yval*. Код в ноутбуке обучает классификатор *SVM*, используя обучающий набор данных (X, y), а также параметры, подгружаемые из программы *dataset3Params*.

Задание: Определить оптимальные параметры C и σ, используя метод перекрестной проверки с помощью множества *Xval, yval*.

Как для C, так и для σ, рекомендуется брать значения с увеличивающимся шагом (например, 0.01; 0.03; 0.1; 0.3; 1; 3; 10; 30). Заметьте, что следует перебрать все возможные пары значений C и σ (например, C = *0.3* и σ = *0.1*). Если, например, Вы попробуете перебрать все перечисленные выше значения для параметра C и σ^2, то при обучении и пересчете программы классификации данных, Вы получите 64 разных модели границы раздела 2 областей. После того, как Вы определите наилучшие параметры C и σ, Вам будет необходимо изменить их в начальном коде программы *dataset3Params*. Используя новые параметры C и σ, программа расчета *SVM* возвращает оптимальную границу раздела 2-х областей, представленную на рис. 7.

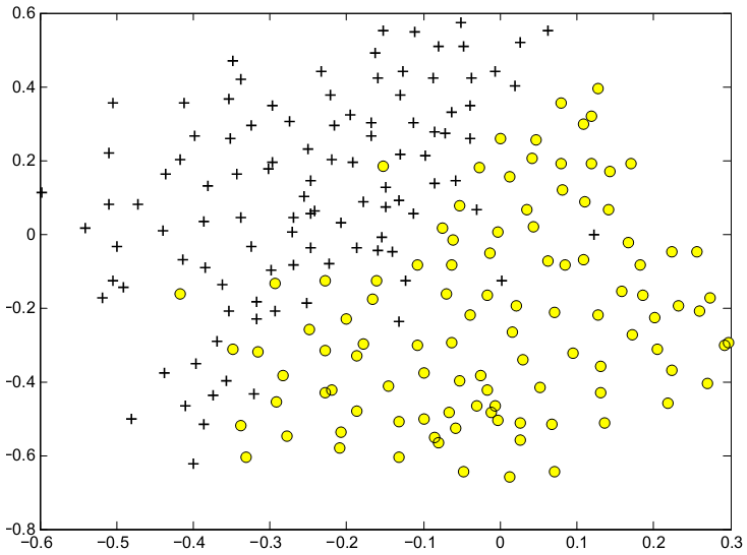

Рис. 6. Третий набор данных

Указание: При выполнении перекрестной проверки данных, для определения оптимальных параметров C и σ Вам необходимо рассчитать ошибку для набора данных, выбранных для проверки. Ошибка определяет долю примеров для перекрестной проверки, классифицированных неправильно.

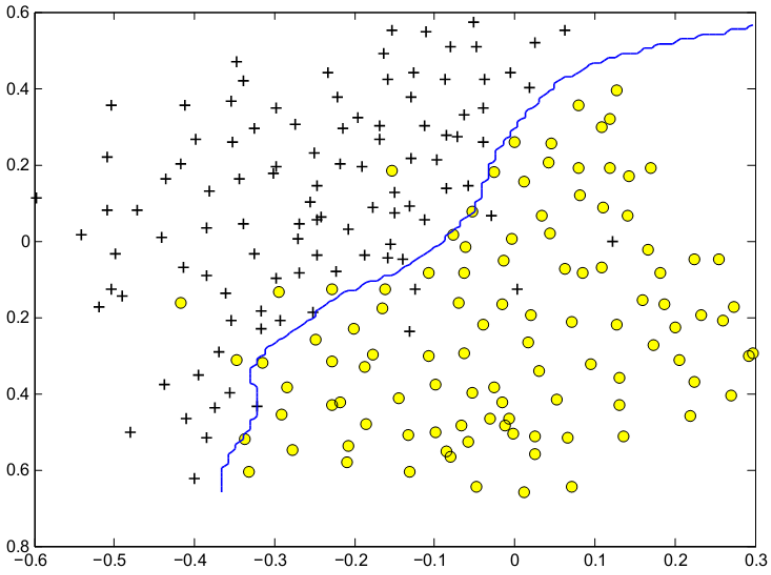

Рис. 7. Метод опорных векторов с ядром Гаусса,
граница раздела для третьего набора данных