

Лабораторная работа #6

Машинное обучение. Метод опорных векторов

1 Цель: изучить основы метода опорных векторов (support vector machine, SVM) в контексте задачи классификации, приобрести навыки работы с методом SVM в системе STATISTICA StatSoft, осуществить обработку методом SVM индивидуального набора данных и интерпретацию результатов

2 Ход работы

- 1) изучить теоретические сведения
- 2) приобрести навыки работы с методами SVM в контексте задачи классификации в системе STATISTICA StatSoft, реализуя приведенный ниже пример
- 3) на основе приобретенных практических навыков осуществить все этапы построения классификатора по методу SVM
- 4) сгенерировать собственные объекты и осуществить классификацию в режиме "Пользовательские прогнозы", подобрав признаки объектов таким образом, чтобы объекты покрыли всё множество классов
- 4) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам.

3 Содержание отчета и требования к его оформлению

- 1) отчет оформляется в печатном виде
- 2) отчет содержит титульный лист, исходные данные, результаты выполнения этапов обработки данных в виде скриншотов и обязательных комментариев по ходу выполнения работы, выводы
- 3) к отчету прилагается файл исходных данных *.sta и проекта в электронном виде с целью осуществления выборочного контроля

4 Варианты исходных данных

- исходные данные - в файле diabetes-77.xlsx (переменная-классификатор - Outcome (показатель диагноза диабета)). В соответствии с вариантом задания создать собственный файл данных и далее работать с полученными данными.

5 Краткие теоретические сведения

Метод опорных векторов (support vector machine, SVM) — один из наиболее популярных методов обучения, который применяется для решения задач классификации и регрессии. Основная идея метода заключается в построении гиперплоскости, разделяющей объекты выборки оптимальным способом. Алгоритм работает в предположении, что чем больше расстояние (зазор) между разделяющей гиперплоскостью и объектами разделяемых классов, тем меньше будет средняя ошибка классификатора. Ниже даны теоретические основы метода опорных векторов для линейно разделимого случая множества объектов на 2 класса.

5 Линейно разделимый случай

Будем решать задачу бинарной (когда класса всего два) классификации. Сначала алгоритм тренируется на объектах из обучающей выборки, для которых заранее известны метки классов. Далее уже обученный алгоритм предсказывает метку класса для каждого объекта из отложенной/тестовой выборки. Метки классов могут принимать значения $Y = \{-1, +1\}$. Объект — вектор с N признаками $x = (x_1, x_2, \dots, x_n)$ в пространстве R^n . При обучении алгоритм должен построить функцию $F(x) = y$, которая принимает в себя аргумент x — объект из пространства R^n и выдает метку класса y .

Главная цель SVM как классификатора — найти уравнение разделяющей гиперплоскости $w_1x_1 + w_2x_2 + \dots + w_nx_n + w_0 = 0$ в пространстве R^n , которая бы разделила два класса неким оптимальным образом. Общий вид преобразования F объекта x в метку класса Y : $F(x) = \text{sign}(w^T x - b)$. Будем помнить, что мы обозначили $w = (w_1, w_2, \dots, w_n)$, $b = -w_0$. После настройки весов алгоритма w и b (обучения), все объекты, попадающие по одну сторону от построенной гиперплоскости, будут предсказываться как первый класс, а объекты, попадающие по другую сторону — второй класс.

Внутри функции $\text{sign}()$ стоит линейная комбинация признаков объекта с весами алгоритма, именно поэтому SVM относится к линейным алгоритмам. Разделяющую гиперплоскость можно построить разными способами, но в SVM веса w и b настраиваются таким образом, чтобы объекты классов лежали как можно дальше от разделяющей гиперплоскости. Другими словами, алгоритм максимизирует зазор (*англ. margin*) между гиперплоскостью и объектами классов, которые расположены ближе всего к ней. Такие объекты и называют опорными векторами (см. рис.2). Отсюда и название алгоритма.

Подробный вывод правил настройки весов SVM:

Чтобы разделяющая гиперплоскость как можно дальше отстояла от точек выборки, ширина полосы должна быть максимальной. Вектор w — вектор нормали к разделяющей гиперплоскости. Здесь и далее будем обозначать скалярное произведение двух векторов как $\langle a, b \rangle$ или $a^T b$. Давайте найдем проекцию вектора, концами которого являются опорные вектора разных классов, на вектор w . Эта проекция и будет показывать ширину разделяющей полосы (см. рис.3):

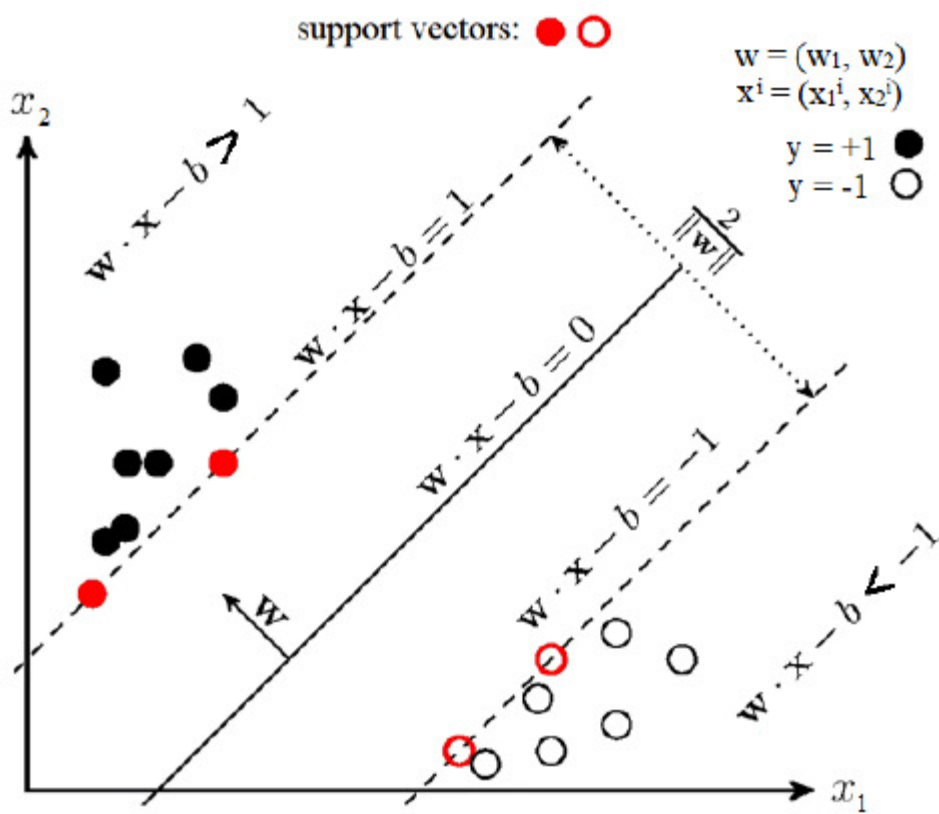


Рисунок 2 – SVM

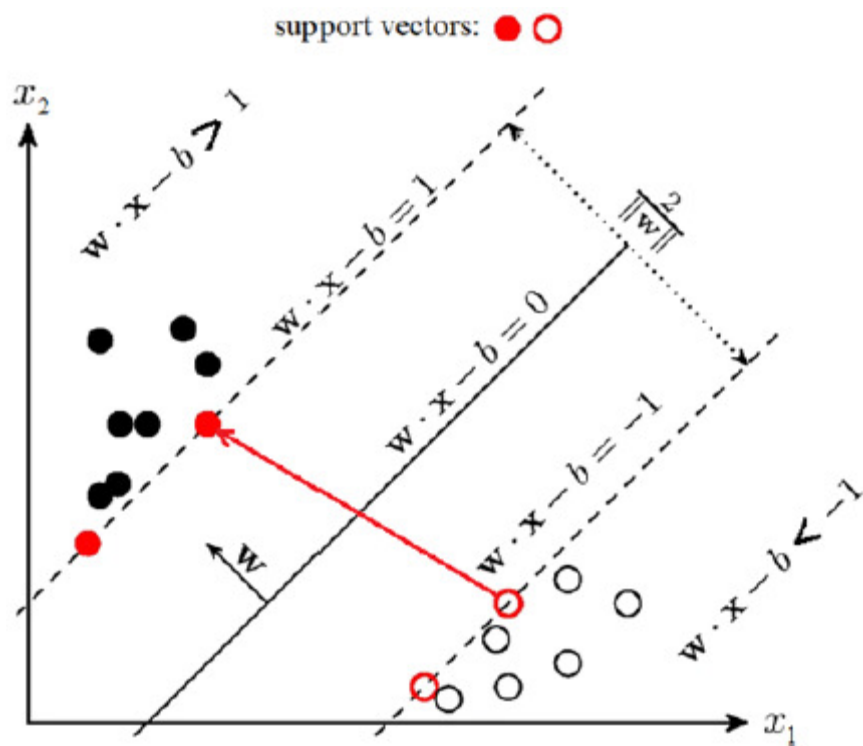


Рисунок 3 – SVM – вывод правил настройки весов

$$\langle (x_+ - x_-), w / \|w\| \rangle = (\langle x_+, w \rangle - \langle x_-, w \rangle) / \|w\| = ((b + 1) - (b - 1)) / \|w\| = 2 / \|w\|$$

$$2 / \|w\| \rightarrow \max$$

$$\|w\| \rightarrow \min$$

$$(w^T w) / 2 \rightarrow \min$$

Отступом (англ. *margin*) объекта x от границы классов называется величина $M = y(w^T x - b)$. Алгоритм допускает ошибку на объекте тогда и только тогда, когда отступ M отрицателен (когда y и $(w^T x - b)$ разных знаков). Если $M \in (0, 1)$, то объект попадает внутрь разделяющей полосы. Если $M > 1$, то объект x классифицируется правильно, и находится на некотором удалении от разделяющей полосы. Т.е. алгоритм будет правильно классифицировать объекты, если выполняется условие:

$$y(w^T x - b) \geq 1$$

Если объединить два выведенных выражения, то получим дефолтную настройку SVM с жестким зазором (*hard-margin SVM*), когда никакому объекту не разрешается попадать на полосу деления. Решается аналитически через теорему Куна-Таккера. Получаемая задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа.

$$\begin{cases} (w^T w) / 2 \rightarrow \min \\ y(w^T x - b) \geq 1 \end{cases}$$

6 Пример машины опорных векторов (Support Vector Machine) - Классификация

На базе данного примера будет изучена задача классификации, а именно проблема с категориальной зависимой переменной. Задача - построить модель машины опорных векторов классификации (Support Vector Machine SVM), которая правильно предсказывает метку класса нового независимого случая.

В качестве примера будем использовать классический набор данных Iris, который содержит информацию о трех различных типах цветов ириса - Versicol, Virginica и Setosa. Набор данных содержит измерения четырех переменных - длины и ширины чашелистника (*SLLENGTH* и *SWIDTH*), а также длины и ширины лепестка (*PLENGTH* и *PWIDTH*). Набор данных Iris имеет ряд интересных особенностей:

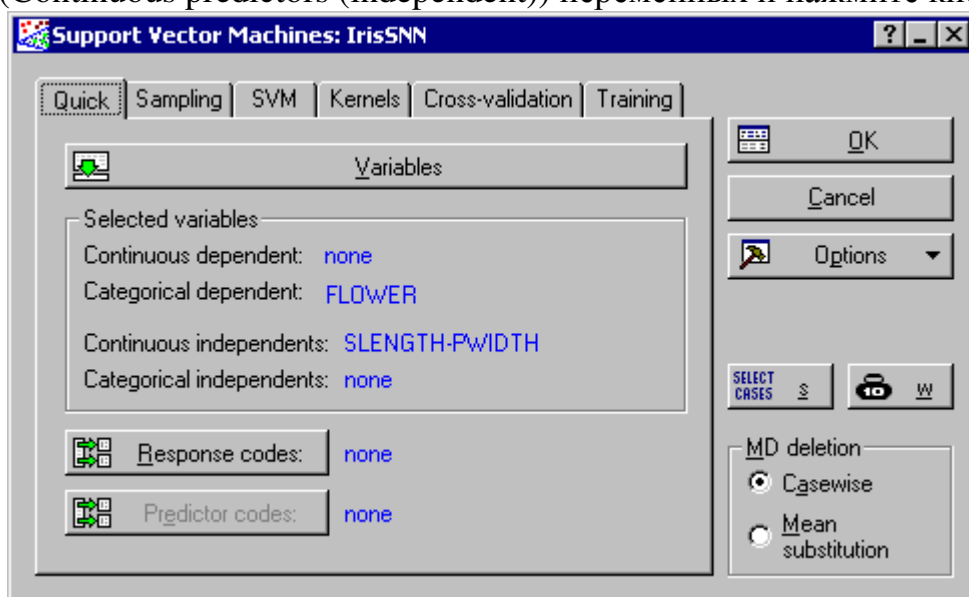
1. Один из классов (Setosa) линейно отделим от двух других. Однако два других класса нельзя разделить линейно.

2. Классы Versicol и Virginic частично пересекаются, поэтому невозможно достичь идеального уровня классификации.

Файл данных. Откройте IrisSNN.sta; он находится в каталоге / Example / Datasets программы STATISTICA.

Запуск анализа. В меню Data Mining выберите Machine Learning (Bayesian, Support Vectors, Nearest Neighbor), чтобы отобразить панель запуска машинного обучения. Выберите «Машина опорных векторов» (Support Vector Machine) и нажмите кнопку «ОК», чтобы отобразить диалоговое окно «Машины опорных векторов».

Настройки анализа. На вкладке «Быстрый» (Quick) нажмите кнопку «Переменные» (Variables), чтобы отобразить стандартный диалог выбора переменных. Выберите FLOWER в качестве категориальной зависимой (Categorical dependent) переменной и переменные 2-5 в качестве списка непрерывных предикторов (независимых) (Continuous predictors (independent)) переменных и нажмите кнопку ОК.



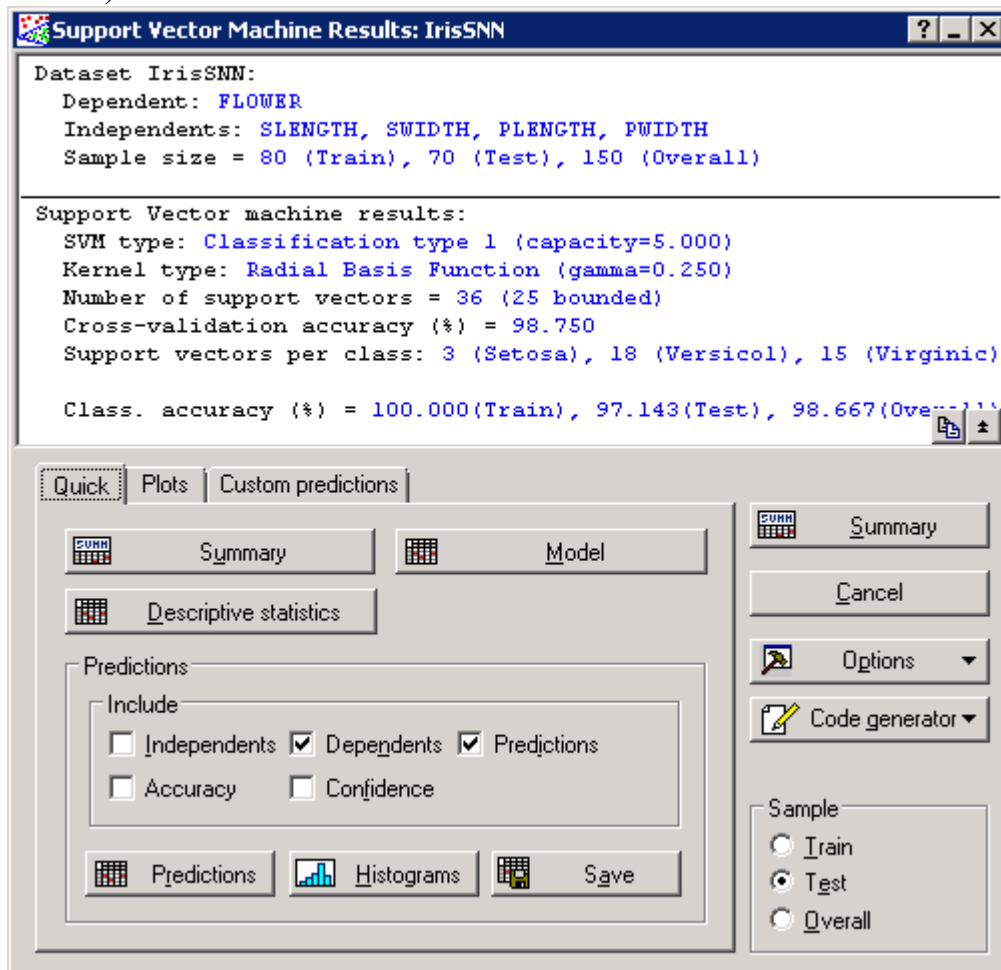
На этом этапе можно изменить спецификации анализа, например, метод выборки, который будет использоваться для разделения данных на обучающую и тестовую выборки, типы SVM и ядра и т. д. Заметим, что некоторые из этих настроек анализа недоступны до тех пор, пока переменные не выбраны.

Одна из важных настроек, которую следует учитывать, - это метод выборки для разделения данных на примеры и тестовые выборки (на вкладке «Выборка» (Sampling)). Хотя выбор переменной выборки не является настройкой по умолчанию (поскольку переменная выборки может быть недоступна в вашем наборе данных), вы можете использовать эту опцию, поскольку, в отличие от случайной выборки, она детерминированно разделяет данные, что позволяет вам сравнить результаты, полученные при различных экспериментальных установках. Рекомендуется всегда предоставлять тестовый образец как средство проверки производительности модели SVM при представлении невидимых данных.

Чтобы использовать переменную выборки для разделения данных, выберите вкладку «Выборка» (Sampling). В рамке группы Выборка (Sampling) нажмите кнопку выбора Использовать переменную выборки (Use sample variable). Затем нажмите кнопку «Образец» (Sample), чтобы отобразить диалоговое окно «Выборка переменной»

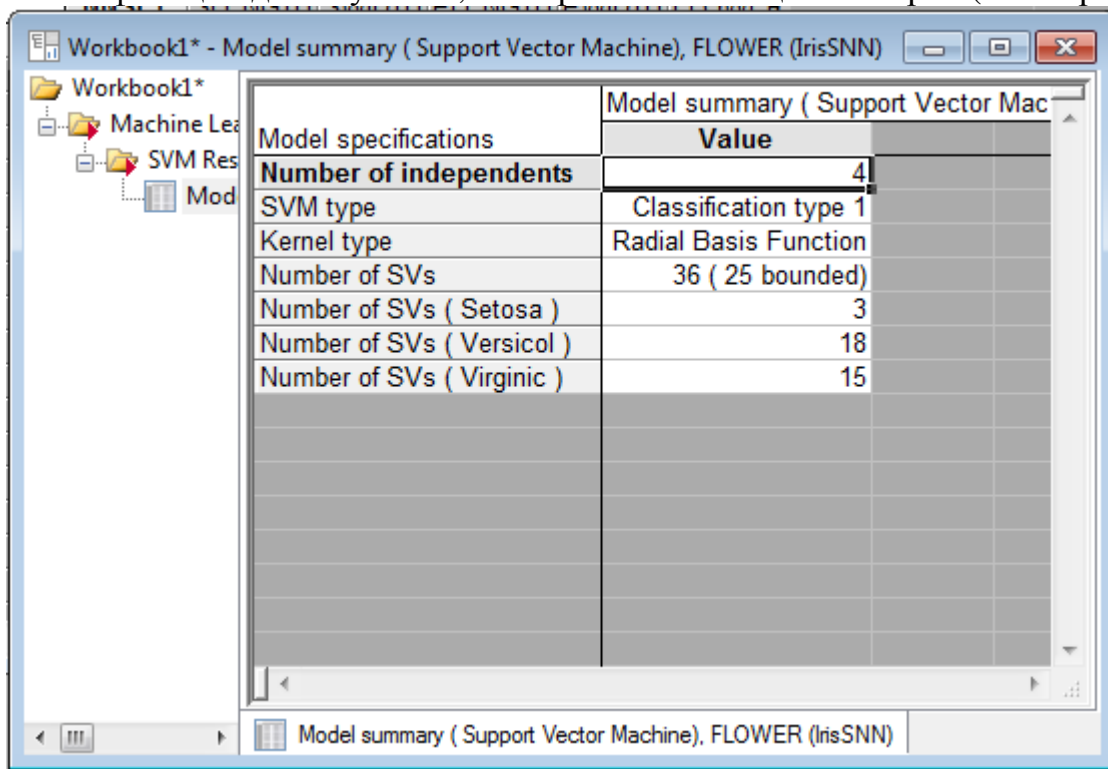
(Sampling variable). Нажмите кнопку «Выборочная переменная идентификатора» (Sample Identifier Variable), чтобы отобразить диалоговое окно «Спецификации перекрестной проверки» (Cross-Validation Specifications), выберите «NNSET» и нажмите кнопку «ОК», чтобы вернуться в диалоговое окно «Выборочная переменная». В поле группы Статус (Status) нажмите кнопку выбора Вкл. (On). Затем дважды щелкните поле Code for analysis sample, чтобы отобразить диалоговое окно Variable code, выберите Train в качестве идентификатора обучающей выборки (случаи, принадлежащие Train, будут использоваться в качестве обучающей выборки для соответствия модели SVM), нажмите кнопку ОК, чтобы вернуться в диалоговое окно «Выборка переменных» (Sampling variable) и нажмите здесь кнопку «ОК», чтобы вернуться в диалоговое окно «Машины опорных векторов» (Support Vector Machines).

Теперь откройте вкладку «Перекрестная проверка» (Cross-validation) и установите флажок «Применить перекрестную проверку v-образной формы» (Apply v-fold cross-validation). Поскольку выбранная модель SVM относится к типу классификации 1 (выбрана по умолчанию на вкладке SVM), для анализа применима только константа емкости (Capacity). Оставьте остальные параметры по умолчанию и нажмите ОК, чтобы начать обучение SVM (подгонку модели), которое выполняется в два этапа. На первом этапе выполняется поиск оценки постоянной емкости, которая обеспечивает наивысшую точность классификации. На втором этапе обучения оценочное значение емкости используется для обучения модели SVM с использованием всей обучающей выборки. По окончании обучения отображается диалоговое окно «Результаты машины опорных векторов» (Support Vector Machine Results).



Просмотр результатов. Используйте диалоговое окно «Результаты» (Results) для просмотра результатов обучения SVM, а также прогнозов в виде таблиц, отчетов и графиков.

В поле «Сводка» (Summary) в верхней части диалогового окна «Результаты» вы можете просмотреть спецификации модели SVM, включая количество опорных векторов и их типы, а также ядра и их параметры. Также перечислены спецификации, сделанные в диалоговом окне «Машины опорных векторов» (Support Vector Machines), включая список зависимых и независимых переменных, а также значения обучающих констант (capacity, epsilon и nu). Отображаются также результаты перекрестной проверки (если применимо), а также статистика классификации для обучения, тестирования и общих выборок (если применимо).



Model summary (Support Vector Mac	
Model specifications	Value
Number of independents	4
SVM type	Classification type 1
Kernel type	Radial Basis Function
Number of SVs	36 (25 bounded)
Number of SVs (Setosa)	3
Number of SVs (Versicol)	18
Number of SVs (Virginic)	15

Примечание. Первое, на что следует обратить внимание в диалоговом окне «Результаты» (Results), - это оценки перекрестной проверки обучающих констант (емкость, эпсилон и nu). Если какое-либо из этих значений равно их максимальному диапазону перекрестной проверки, это может указывать на то, что ваш диапазон поиска недостаточно велик для включения лучших значений. В этом случае нажмите кнопку «Отмена» (Cancel), чтобы вернуться в диалоговое окно «Машины опорных векторов», чтобы расширить диапазон поиска.

SVM строит классификационную функцию через набор опорных векторов и коэффициентов. Нажмите кнопку «Модель» (Model), чтобы создать таблицы с этими величинами. Это полезно для подробного обзора модели SVM или для включения в отчеты.

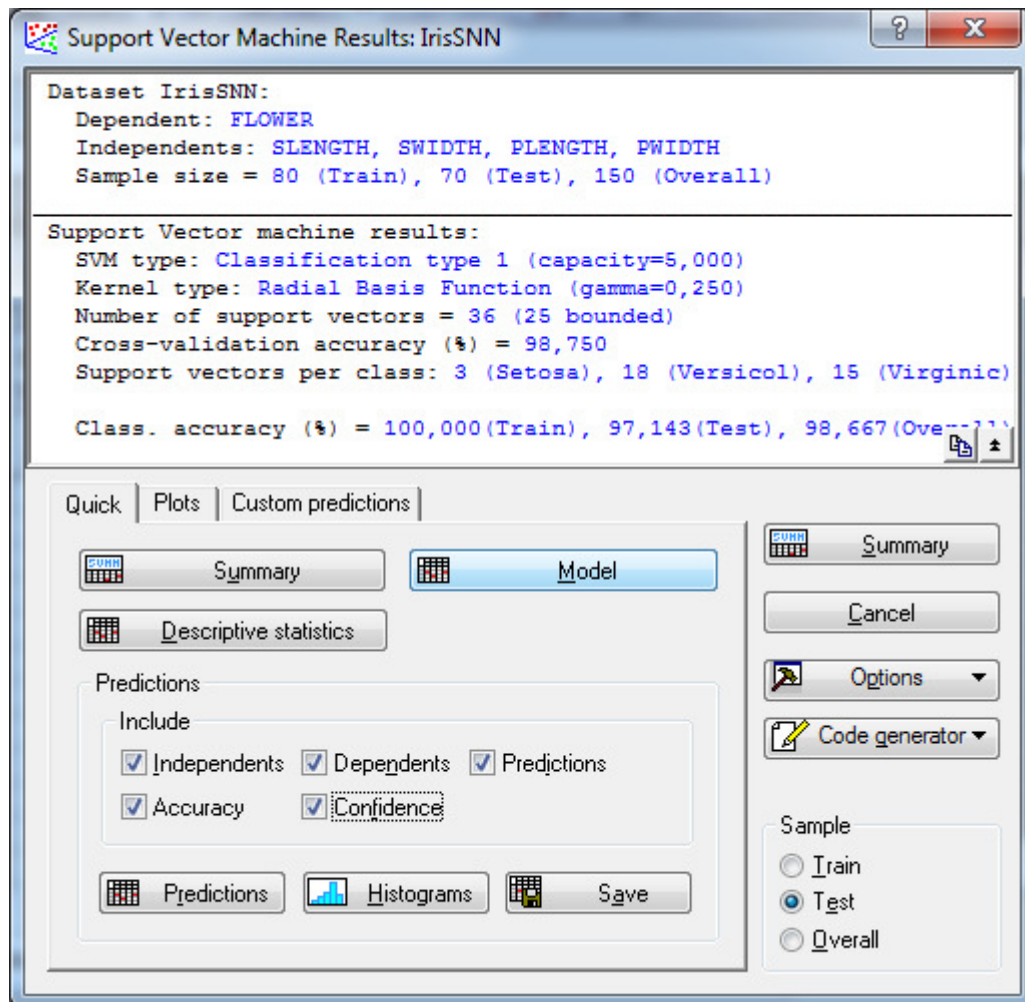
SVM model specifications (decision constants), (IrisSNN)	
SVM: Classification type 1 (C=5,000000)	
Kernel: Radial Basis Function (gamma=0,250000)	
Decision Constant	Value
1	0,014464
2	0,005240
3	-0,116939

Дополнительную информацию можно получить, нажав кнопку «Описательная статистика» (Descriptive statistics), в результате чего будут созданы две электронные таблицы, содержащие сводку классификации и матрицу неточностей.

Classification summary (Support Vector Machine), FLOWER, T...						
SVM: Classification type 1 (C=5,000), Kernel: Radial Bas						
Number of support vectors= 36 (25 bounded)						
Class Name	Total	Correct	Incorrect	Correct(%)	Incorrect(%)	
Setosa	23	23	0	100,0000	0,000000	
Versicol	24	23	1	95,8333	4,166667	
Virginic	23	22	1	95,6522	4,347826	

Confusion matrix (Support Vector Machine), FLOWER, Te...					
SVM: Classification type 1 (C=5,000), Kernel: Radial Bas					
Observed (rows) x Predicted (columns)					
Class Observed	Setosa	Versicol	Virginic		
Setosa	23	0	0		
Versicol	0	23	1		
Virginic	0	1	22		

Для дальнейшего просмотра результатов вы можете отобразить электронную таблицу прогнозов на вкладке Пользовательские прогнозы (Custom predictions) (и включить любую другую переменную, которая может вас заинтересовать, например, независимая, точность и т. д., выбрав соответствующую кнопку выбора на вкладке Быстрые (Quick)). Вы также можете отобразить эти величины в виде графиков гистограмм.



Workbook2* - Predictions (Support Vector Machine), Test sample (IrisSNN)

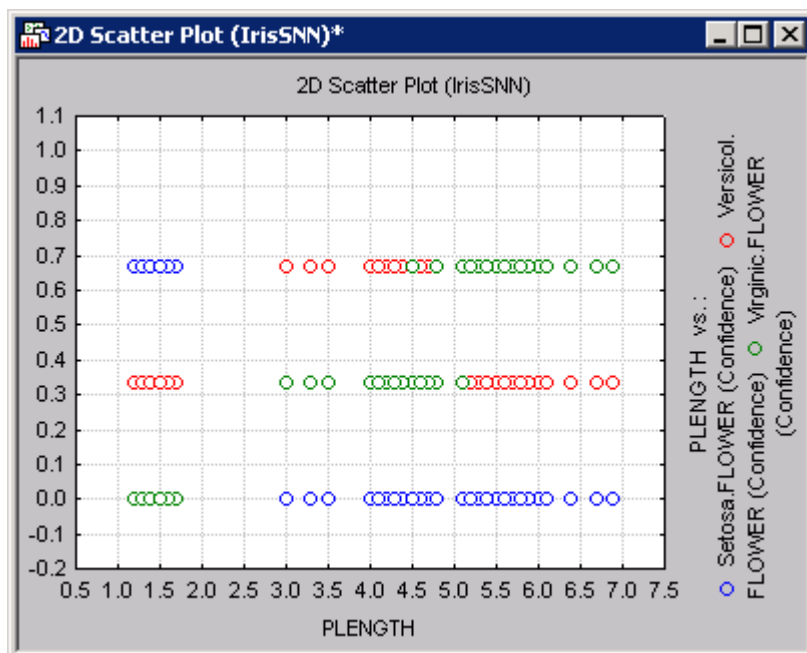
Machine Learning

SVM Results

Predictions (Support Vector Machine), Test sample (IrisSNN)
SVM: Classification type 1 (C=5,000), Kernel: Radial Basis Function (gamma=0,250)
Number of support vectors= 36 (25 bounded)

Case Name	SLENGTH Independent	SWIDTH Independent	PLENGTH Independent	PWIDTH Independent	FLOWER Dependent	FLOWER Predicted	FLOWER Accuracy	Setosa Confidence	Versicol Confidence	Virginic Confidence
1	5,100000	3,500000	1,400000	0,200000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
2	4,900000	3,000000	1,400000	0,200000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
4	4,600000	3,100000	1,500000	0,200000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
8	5,000000	3,400000	1,500000	0,200000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
9	4,400000	2,900000	1,400000	0,200000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
10	4,900000	3,100000	1,500000	0,100000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
16	5,700000	4,400000	1,500000	0,400000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
17	5,400000	3,900000	1,300000	0,400000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
19	5,700000	3,800000	1,700000	0,300000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
20	5,100000	3,800000	1,500000	0,300000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
22	5,100000	3,700000	1,500000	0,400000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000
24	5,100000	3,300000	1,700000	0,500000	Setosa	Setosa	Correct	0,666667	0,333333	0,000000

Дальнейший графический обзор результатов может быть выполнен на вкладке «Графики» (Plots), где вы можете создать двух- и трехмерные графики переменных и уровней достоверности. Обратите внимание, что вы можете отображать более одной переменной на двухмерных диаграммах рассеяния.



Например, выше показан график зависимости независимой переменной PLENGTH от достоверности классификации. Обратите внимание, что класс Setosa полностью отделен от Versicol и Virginic, в то время как невозможно полностью отличить два последних (обратите внимание на область оси X, где эти два класса значительно перекрываются, т. е. имеют аналогичные значения достоверности). Чтобы построить график, показанный выше, выберите PLENGTH из списка по оси X и Setosa (conf.), Versicol (conf.) и Virginic (conf.) Из списка по оси Y. Затем нажмите кнопку «Графики X и Y» (Graphs of X and Y).

Наконец, вы можете выполнить «а что, если?» ("what if?") анализ на вкладке "Пользовательские прогнозы" (Custom predictions). Вы можете определить новые наблюдения (которые не взяты из набора данных) и выполнить модель SVM, используя их, что позволит выполнить специальную операцию «Что, если?» ("what if?") анализа. Нажмите кнопку «Прогнозы» (Predictions), чтобы создать электронную таблицу модели.

В приведенном ниже рисунке показан результат классификации пользовательского объекта.

Independent and dependent variables		Custom predictions (IrisSNN)									
		Value									
SL	PLENGTH	7,000000									
SL	SWIDTH	5,000000									
C	PLENGTH	8,000000									
C	PWIDTH	3,000000									
H	FLOWER	Versicol									
FL											
H											
M											
SL											

6 Контрольные вопросы

1. В чем суть метода опорных векторов (support vector machine, SVM)?

2. Какова главная цель SVM как классификатора для линейно разделимого случая классификации?
3. Какую цель преследует в методе SVM соответствующая настройка весов W и b , которые являются параметрами разделяющей гиперплоскости?
4. Что представляет собой зазор (англ. margin) в методе SVM?
5. Как определяется зазор (англ. margin) в методе SVM для линейно разделимого случая классификации?
6. Пусть в результате выполнения метода SVM для линейно разделимого случая бинарной классификации двумерных образов (см. пример на рис. 2, рис. 3) получены параметры разделяющей поверхности $W=8$ и $b=2$. Чему равняется ширина разделяющей полосы?