

Лабораторная работа #5

Машинное обучение. Метод k-ближайших соседей

Цель: изучить основы методов Machine Learning в контексте задачи классификации методом k-ближайших соседей (k-nearest neighbors, k-NN), приобрести навыки работы с методами Machine Learning в системе STATISTICA StatSoft, осуществить обработку методами Machine Learning индивидуального набора данных и интерпретацию результатов

1 Ход работы

- 1) изучить теоретические сведения
- 2) приобрести навыки работы с методами Machine Learning в контексте задачи классификации методом k-ближайших соседей, в системе STATISTICA StatSoft, реализуя приведенный ниже пример
- 3) на основе приобретенных практических навыков осуществить все этапы обработки методами Machine Learning в контексте задачи классификации методом k-ближайших соседей и интерпретацию результатов согласно варианту индивидуального задания
- 4) для полученной модели подобрать пользовательские паттерны, относящиеся ко всем классам
- 5) оформить отчет и подготовиться к защите лабораторной работы по полученным результатам и контрольным вопросам.

2 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 файл проекта в электронном виде с целью осуществления выборочного контроля.

3 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.

4 Краткие теоретические сведения

Задача классификации — задача, в которой имеется множество объектов (ситуаций), разделённых, некоторым образом на классы. Задано конечное множество объектов, для которых известно, к каким классам они относятся. Это множество называется выборкой. Классовая принадлежность остальных объектов неизвестна. Требуется построить алгоритм, способный классифицировать произвольный объект из исходного множества.

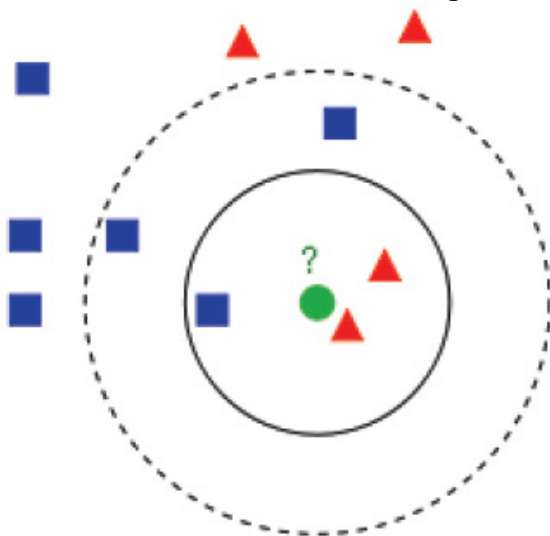
Классифицировать объект — значит, указать номер (или наименование) класса, к которому относится данный объект.

Классификация объекта — номер или наименование класса, выдаваемый алгоритмом классификации в результате его применения к данному конкретному объекту.

Пусть X — множество описаний объектов, Y — множество номеров (или наименований) классов. Существует неизвестная *целевая зависимость* — отображение $y^*: X \rightarrow Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1, y_1), \dots, (x_m, y_m)\}$. Требуется построить алгоритм $a: X \rightarrow Y$, способный классифицировать произвольный объект $x \in X$.

Метод k-ближайших соседей (англ. k-nearest neighbors algorithm, k-NN) — метрический алгоритм для автоматической классификации объектов или регрессии. В случае использования метода для классификации объект присваивается тому классу, который является наиболее распространённым среди k соседей данного элемента, классы которых уже известны. В случае использования метода для регрессии, объекту присваивается среднее значение по k ближайшим к нему объектам, значения которых уже известны.

Алгоритм может быть применим к выборкам с большим количеством атрибутов (многомерным). Для этого перед применением нужно определить функцию расстояния; классический вариант такой функции — евклидова метрика, хотя могут быть использованы иные метрики, например, матхэттенновская и др.



Пример классификации *k*-ближайших соседей. Тестовый образец (зелёный круг) должен быть классифицирован как синий квадрат (класс 1) или как красный треугольник (класс 2). Если $k = 3$, то он классифицируется как 2-й класс, потому что внутри меньшего круга 2 треугольника и только 1 квадрат. Если $k = 5$, то он будет классифицирован как 1-й класс (3 квадрата против 2 треугольников внутри большего круга)

Нормализация

Разные атрибуты могут иметь разный диапазон представленных значений в выборке (например атрибут А представлен в диапазоне от 0.1 до 0.5, а атрибут Б представлен в диапазоне от 1000 до 5000), то значения дистанции могут сильно зависеть от атрибутов с большими диапазонами. Поэтому данные обычно подлежат нормализации. При кластерном анализе есть два основных способа нормализации данных: минимакс-нормализация и Z-нормализация.

Минимакс-нормализация осуществляется следующим образом:

$$x' = (x - \min[X]) / (\max[X] - \min[X]),$$

в этом случае все значения будут лежать в диапазоне от 0 до 1; дискретные бинарные значения определяются как 0 и 1.

Z-нормализация:

$$x' = (x - M[X]) / \sigma[X]$$

где σ — среднеквадратичное отклонение; в этом случае большинство значений попадёт в диапазон $(-3\sigma; 3\sigma)$.

Выделение значимых атрибутов

Некоторые значимые атрибуты могут быть важнее остальных, поэтому для каждого атрибута может быть задан в соответствие определённый вес (например вычисленный с помощью тестовой выборки и оптимизации ошибки отклонения). Таким образом, каждому атрибуту k будет задан в соответствие вес z_k , так что значение атрибута будет попадать в диапазон $[0; z_k \max(k)]$ (для нормализованных значений по минимакс-методу). Например, если атрибуту присвоен вес 2,7, то его нормализованно-взвешенное значение будет лежать в диапазоне $[0; 2, 7]$

Взвешенный способ

При взвешенном способе во внимание принимается не только количество попавших в область определённых классов, но и их удалённость от нового значения.

Для каждого класса j определяется оценка близости:

$$Q_j = \sum_{i=1}^n \frac{1}{d(x, a_i)^2},$$

где $d(x, a_i)$ — расстояние от нового значения x до объекта a_i .

У какого класса выше значение близости, тот класс и присваивается новому объекту.

С помощью метода можно вычислять значение одного из атрибутов классифицируемого объекта на основании дистанций от попавших в область объектов и соответствующих значений этого же атрибута у объектов:

$$x_k = \frac{\sum_{i=1}^n k_i d(x, a_i)^2}{\sum_{i=1}^n d(x, a_i)^2},$$

где a_i — i -ый объект, попавший в область, k_i — значение атрибута k у заданного объекта a_i , x — новый объект, x_k — k -ый атрибут нового объекта.

5 Пример классификации по методу К-ближайшего соседа

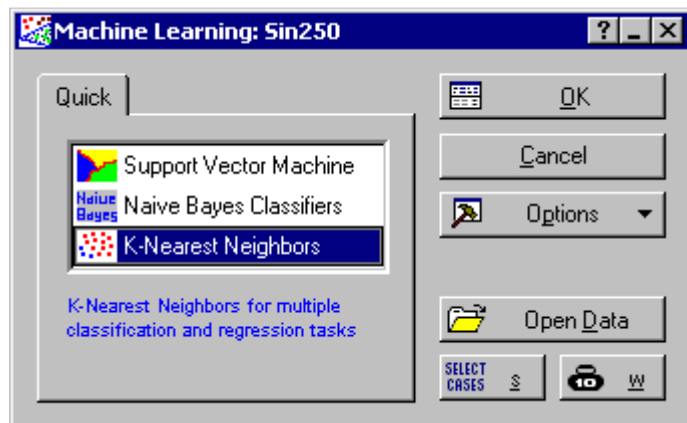
В этом примере изучим проблему классификации, то есть проблему с категориальной выходной (зависимой) переменной. Задача - построить модель классификатора K-Nearest Neighbor, которая правильно предсказывает метку класса (категорию) независимых переменных.

Для примера будем использовать классический набор данных Iris. Этот набор данных содержит информацию о трех различных типах цветов ириса: Versicol, Virginic и Setosa. Набор данных содержит измерения четырех переменных [длина и ширина чашелистика (SLENGTH и SWIDTH) и длина и ширина лепестка (PLENGTH и PWIDTH)]. Набор данных Iris имеет ряд интересных особенностей:

- Один из классов (Setosa) линейно отделим от двух других. Однако два других класса нельзя разделить линейно.
- Классы Versicol и Virginic частично пересекаются, поэтому невозможно достичь идеального уровня классификации.

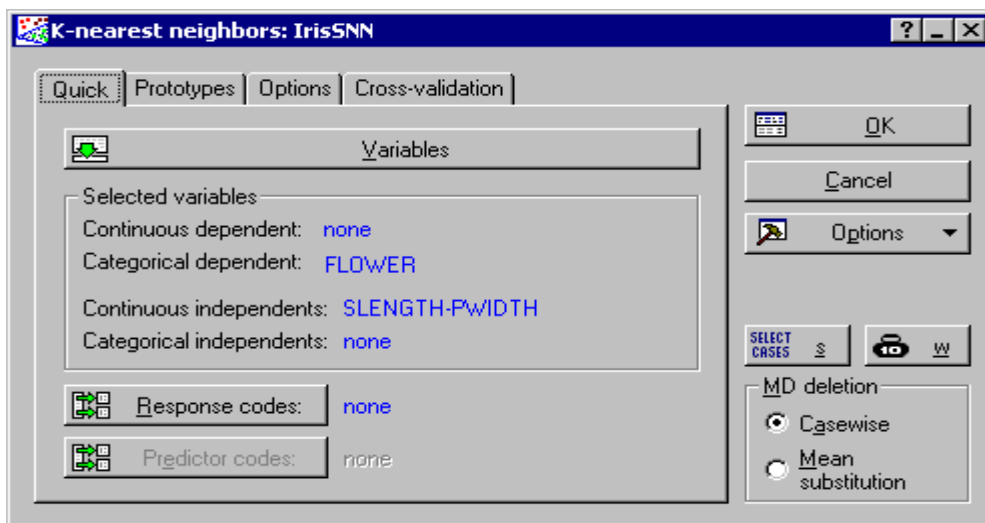
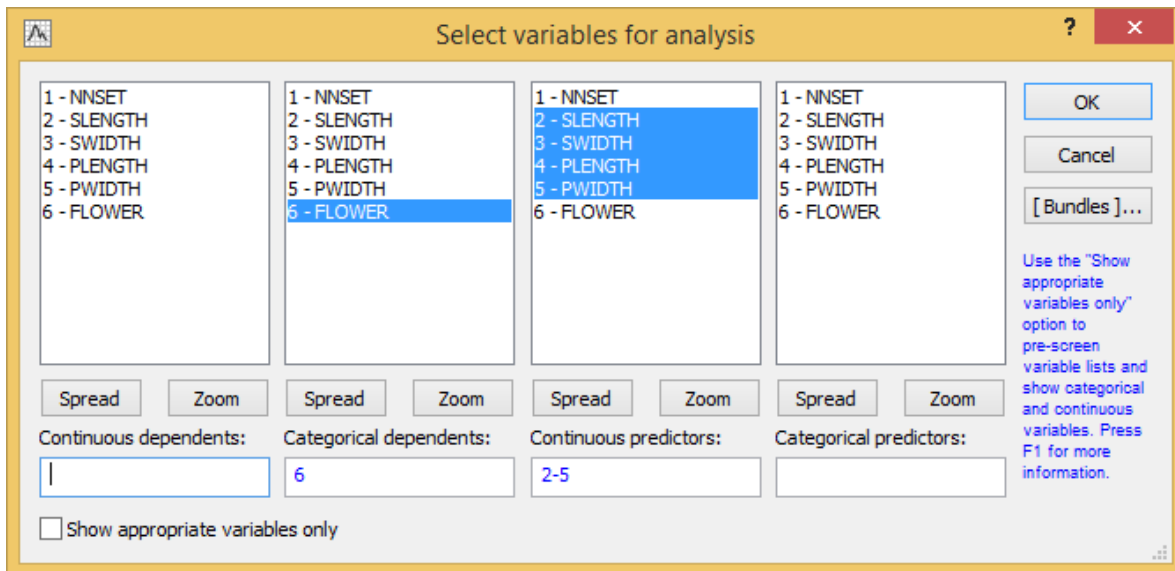
Файл данных. Откройте файл данных IrisSNN.sta.

Запускаем анализ. Выберите Машинное обучение (байесовское, опорные векторы, ближайшее соседство) (Machine Learning (Bayesian, Support Vectors, Nearest Neighbor)) в меню интеллектуального анализа данных, чтобы отобразить панель запуска машинного обучения.



Выберите К-ближайших соседей на вкладке Quick и нажмите кнопку ОК, чтобы отобразить диалоговое окно K-Nearest Neighbours. Вы также можете дважды щелкнуть K-Nearest Neighbours, чтобы открыть это диалоговое окно.

Нажмите кнопку «Переменные», чтобы отобразить стандартный диалог выбора переменных. Выберите ЦВЕТОК (*FLOWER*) в качестве категориальной зависимой переменной и переменные 2-5 из списка непрерывных предикторов (независимых) переменных и нажмите кнопку ОК.

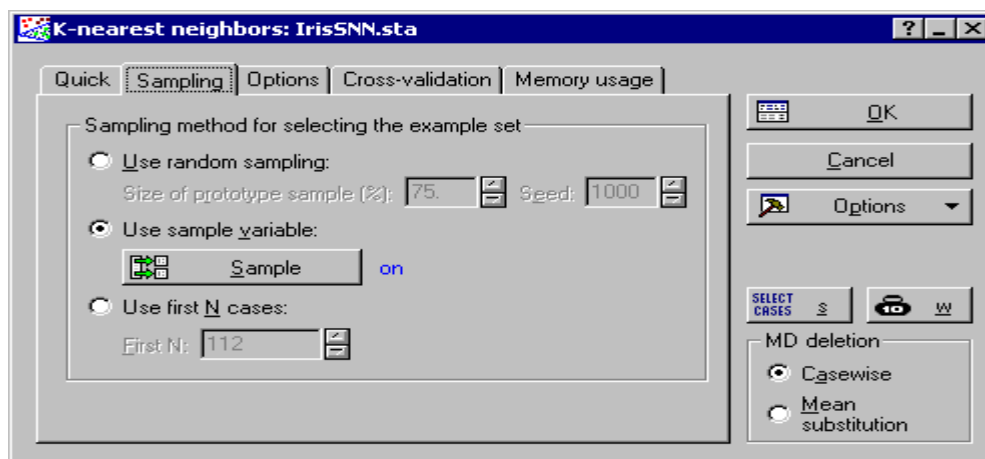
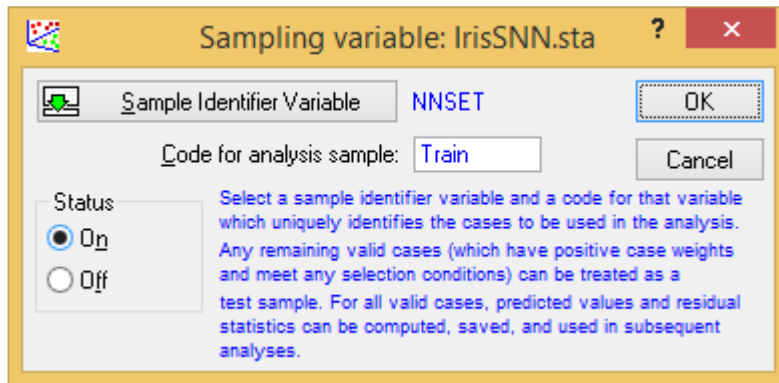


На этом этапе вы можете изменить характеристики анализа, например, метод выборки для разделения данных на примеры и тестовые выборки, а также количество ближайших соседей K, а также расстояние мера (метрика) и схема усреднения. Для анализа с более чем одной независимой переменной со значительно разными типичными значениями вы также можете стандартизировать расстояния (только для непрерывных независимых переменных).

Одним из важных параметров, который следует учитывать, является метод выборки для разделения данных на примеры и тестовые образцы (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 она детерминированно делит данные, в отличие от случайной выборки, что позволяет упростить сравнение результатов, полученные при различных экспериментальных установках, т. е. при выборе K, меры расстояния и т. д.

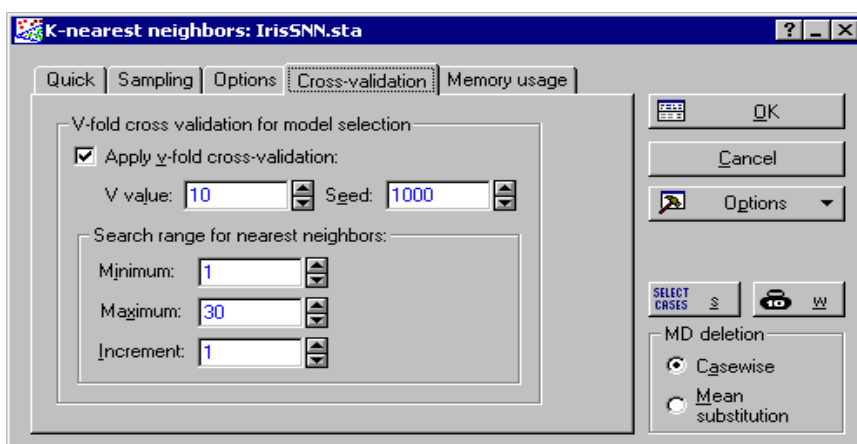
В этом примере на вкладке «Выборка» нажмите кнопку выбора «Использовать переменную образца» и нажмите кнопку «Образец», чтобы отобразить диалоговое окно «Выборка переменной». В этом диалоговом окне нажмите кнопку Sample Identifier Variable, выберите NNSET в качестве переменной выборки и нажмите кнопку ОК. Затем дважды щелкните поле «Код для образца анализа», выберите

«Train» в качестве кода для образца анализа и нажмите кнопку «ОК». В поле группы «Состояние» нажмите кнопку «Вкл.» (on) и нажмите кнопку «ОК».

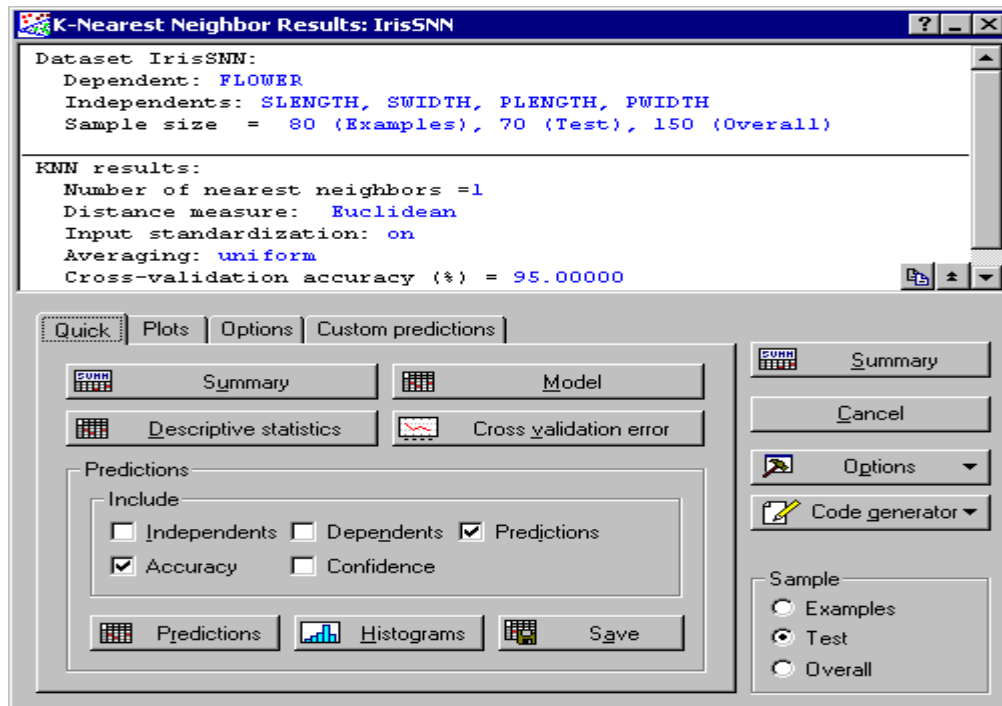


При выполнении анализа KNN рекомендуется стандартизировать независимые переменные, чтобы их типичные значения случая попадали в один и тот же диапазон. Это предотвратит искажение прогнозов независимыми переменными с обычно большими значениями. Чтобы применить это масштабирование, в диалоговом окне К-ближайших соседей - вкладка «Параметры» установите флажок «Стандартизировать расстояния» (этот параметр также доступен в диалоговом окне «Результаты»).

Наконец, выберите вкладку «Перекрыстная проверка» (Cross validation). Установите флажок Применить перекрыстную проверку v-образной формы; в рамке группы «Диапазон поиска ближайших соседей» введите 30 в поле «Максимум» (чтобы увеличить максимальное количество ближайших соседей K до 30).



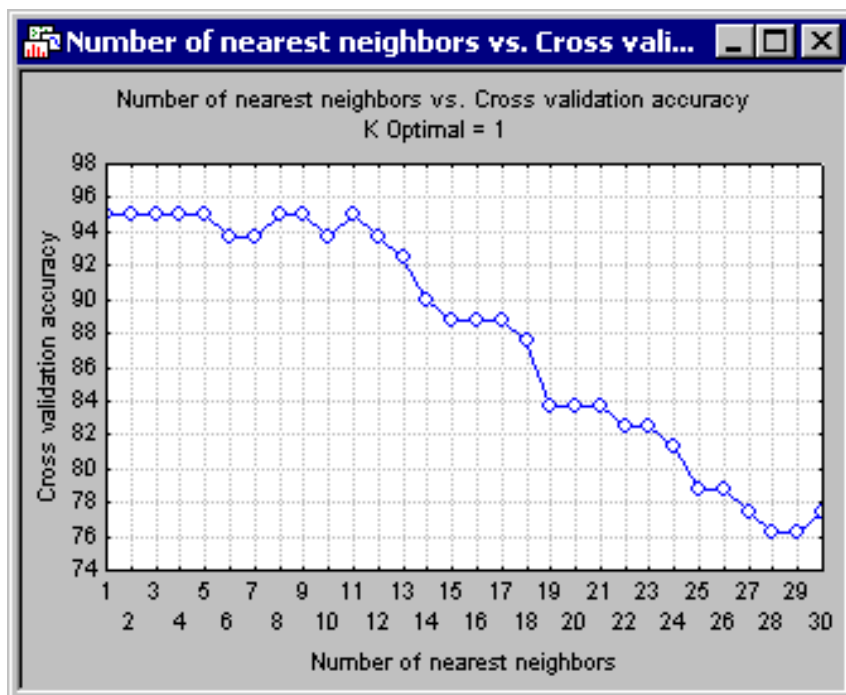
Щелкните кнопку ОК. Пока KNN ищет оценку K с использованием алгоритма перекрестной проверки, отображается индикатор выполнения, за которым следует диалоговое окно K-Nearest Neighbor Results.



Просмотр результатов. В диалоговом окне «Результаты K-ближайших соседей» вы можете выполнять прогнозы KNN и просматривать результаты в виде таблиц, отчетов и графиков.

В поле «Сводка» в верхней части диалогового окна «Результаты» можно увидеть некоторые характеристики вашего анализа KNN, включая список переменных, выбранных для анализа, и размер примеров, тестов и общих образцов (если применимо). Отображается также количество ближайших соседей, мера расстояния и используются ли стандартизация ввода и взвешивание на основе расстояния. Можно также просмотреть ошибку перекрестной проверки. Обратите внимание, что это спецификации, сделанные в диалоговом окне K-Nearest Neighbours.

На вкладке «Быстрая» диалогового окна «Результаты» нажмите кнопку «Ошибка перекрестной проверки». Это создаст график ошибки перекрестной проверки для каждого значения K, опробованного алгоритмом перекрестной проверки.



Первое, что вы должны искать на этом графике, - это наличие седловой точки, то есть значение К с максимальной точностью классификации по сравнению с соседними точками. Наличие максимума указывает на то, что диапазон поиска для К был достаточно широким, чтобы включить оптимальное (с точки зрения перекрестной проверки) значение. Если его нет, вернитесь в диалоговое окно KNN, нажав кнопку «Отмена» в диалоговом окне «Результаты», и увеличьте диапазон поиска.

Как обсуждалось ранее, STATISTICA KNN делает прогнозы на основе подмножества, известного как примеры (или экземпляры). Нажмите кнопку «Модель», чтобы создать электронную таблицу, содержащую значения наблюдений для этого конкретного образца.

Data: K-Nearest Neighbors (IrisSNN)*				
K-Nearest Neighbors (IrisSNN.sta)				
Number of nearest neighbors = 1, Distance = 1				
Averaging: uniform				
Examples	SLLENGTH	SWIDTH	PLENGTH	PWIDTH
3	4.700000	3.200000	1.300000	0.200000
5	5.000000	3.600000	1.400000	0.200000
6	5.400000	3.900000	1.700000	0.400000
7	4.600000	3.400000	1.400000	0.300000
11	5.400000	3.700000	1.500000	0.200000
12	4.800000	3.400000	1.600000	0.200000
13	4.800000	3.000000	1.400000	0.100000
14	4.300000	3.000000	1.100000	0.100000
15	5.800000	4.000000	1.200000	0.200000
18	5.100000	3.500000	1.400000	0.300000
21	5.400000	3.400000	1.700000	0.200000
23	4.600000	3.600000	1.000000	0.200000
25	4.800000	3.400000	1.900000	0.200000
27	5.000000	3.400000	1.600000	0.400000
29	5.200000	3.400000	1.400000	0.200000
31	4.800000	3.100000	1.600000	0.200000
33	5.200000	4.100000	1.500000	0.100000
35	4.900000	3.100000	1.500000	0.100000
37	5.500000	3.500000	1.300000	0.200000
38	4.900000	3.100000	1.500000	0.100000
39	4.400000	3.000000	1.300000	0.200000
40	5.100000	3.400000	1.500000	0.200000
41	5.000000	3.500000	1.300000	0.300000

Дополнительную информацию об анализе классификации можно получить, нажав кнопку «Описательная статистика» (Descriptive statistics), в результате чего отобразятся две таблицы, содержащие сводную информацию о классификации и матрицу неточностей.

Data: Classification summary (K-Nearest Neighbors), F...					
Classification summary (K-Nearest Neighbors), FLOW					
Nearest neighbors = 1, Distance: Euclidean, Standardization: on					
Class Name	Total	Correct	Incorrect	Correct(%)	Incorrect(%)
Setosa	23	23	0	100.0000	0.000000
Versicol	24	22	2	91.6667	8.333333
Virginic	23	21	2	91.3043	8.695652

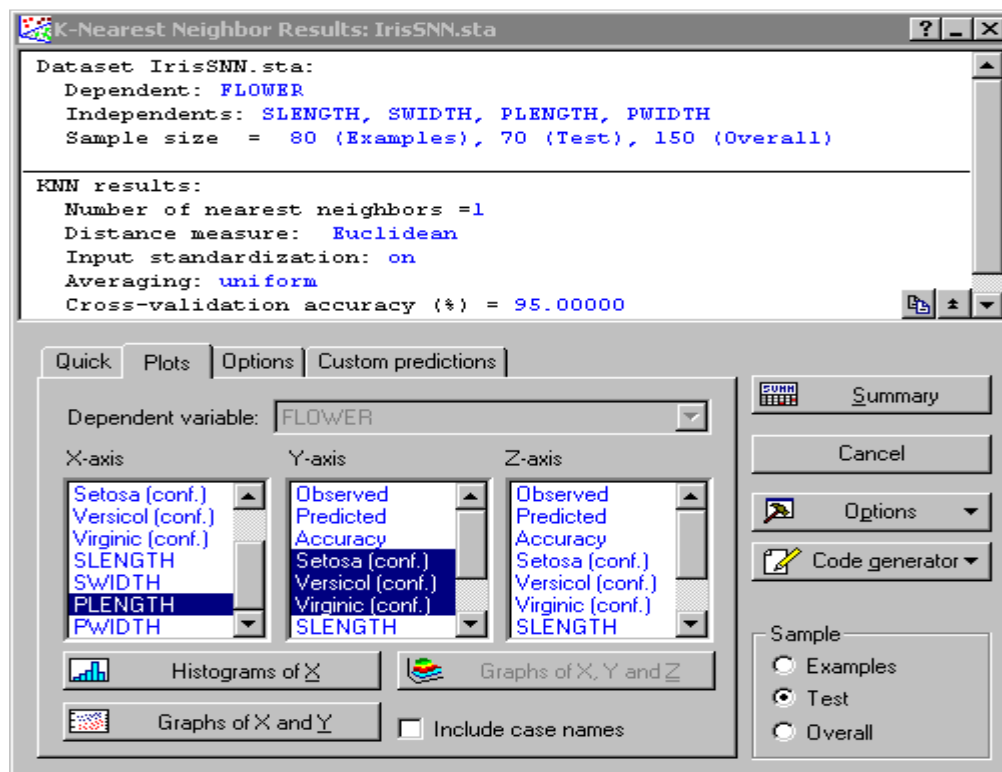
Data: Confusion matrix (K-Ne...			
Confusion matrix (K-Nearest Neighbors), FLOW			
Nearest neighbors = 1, Distance: Euclidean, Standardization: on			
Class Predicted	Setosa	Versicol	Virginic
Setosa	23	0	0
Versicol	0	22	2
Virginic	0	2	21

Для дальнейшего просмотра результатов вы можете нажать кнопку «Прогнозы», чтобы отобразить электронную таблицу прогнозов (и включить любые другие переменные, которые могут вас заинтересовать, например, независимые, зависимые и точность, установив соответствующие флажки в поле «Включить» (Include)).

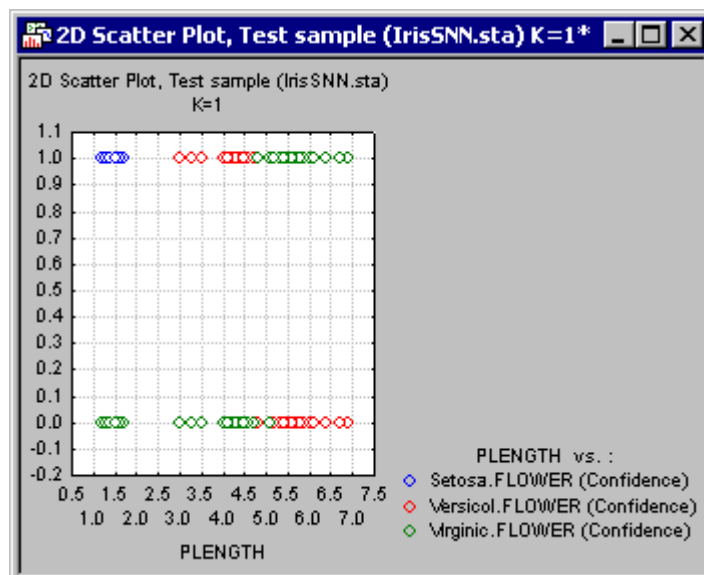
Data: Predictions (K-Nearest Neighbors), Test sample (IrisSNN)*										
Predictions (K-Nearest Neighbors), Test sample (IrisSNN.sta)										
Nearest neighbors = 1, Distance: Euclidean, Standardization: on, Averaging: uniform										
Case Name	SLLENGTH Independent	SWIDTH Independent	PLENGTH Independent	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	1.000000	0.000000	0.000000
2	4.900000	3.000000	1.400000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
4	4.600000	3.100000	1.500000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
8	5.000000	3.400000	1.500000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
9	4.400000	2.900000	1.400000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
10	4.900000	3.100000	1.500000	0.100000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
16	5.700000	4.400000	1.500000	0.400000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
17	5.400000	3.900000	1.300000	0.400000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
19	5.700000	3.800000	1.700000	0.300000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
20	5.100000	3.800000	1.500000	0.300000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
22	5.100000	3.700000	1.500000	0.400000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
24	5.100000	3.300000	1.700000	0.500000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
26	5.000000	3.000000	1.600000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
28	5.200000	3.500000	1.500000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
30	4.700000	3.200000	1.600000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
32	5.400000	3.400000	1.500000	0.400000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
34	5.500000	4.200000	1.400000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
36	5.000000	3.200000	1.200000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
42	4.500000	2.300000	1.300000	0.300000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
44	5.000000	3.500000	1.600000	0.600000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
47	5.100000	3.800000	1.600000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
48	4.600000	3.200000	1.400000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
49	5.300000	3.700000	1.500000	0.200000	Setosa	Setosa	Correct	1.000000	0.000000	0.000000
52	6.400000	3.200000	4.500000	1.500000	Versicol	Versicol	Correct	0.000000	1.000000	0.000000
54	5.500000	2.300000	4.000000	1.300000	Versicol	Versicol	Correct	0.000000	1.000000	0.000000
58	4.900000	2.400000	3.300000	1.000000	Versicol	Versicol	Correct	0.000000	1.000000	0.000000
61	5.000000	2.000000	3.500000	1.000000	Versicol	Versicol	Correct	0.000000	1.000000	0.000000
66	6.700000	3.100000	4.400000	1.400000	Versicol	Versicol	Correct	0.000000	1.000000	0.000000

Вы также можете отображать эти переменные в виде графиков гистограмм.

Дальнейший графический обзор результатов может быть выполнен на вкладке «Графики», 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.

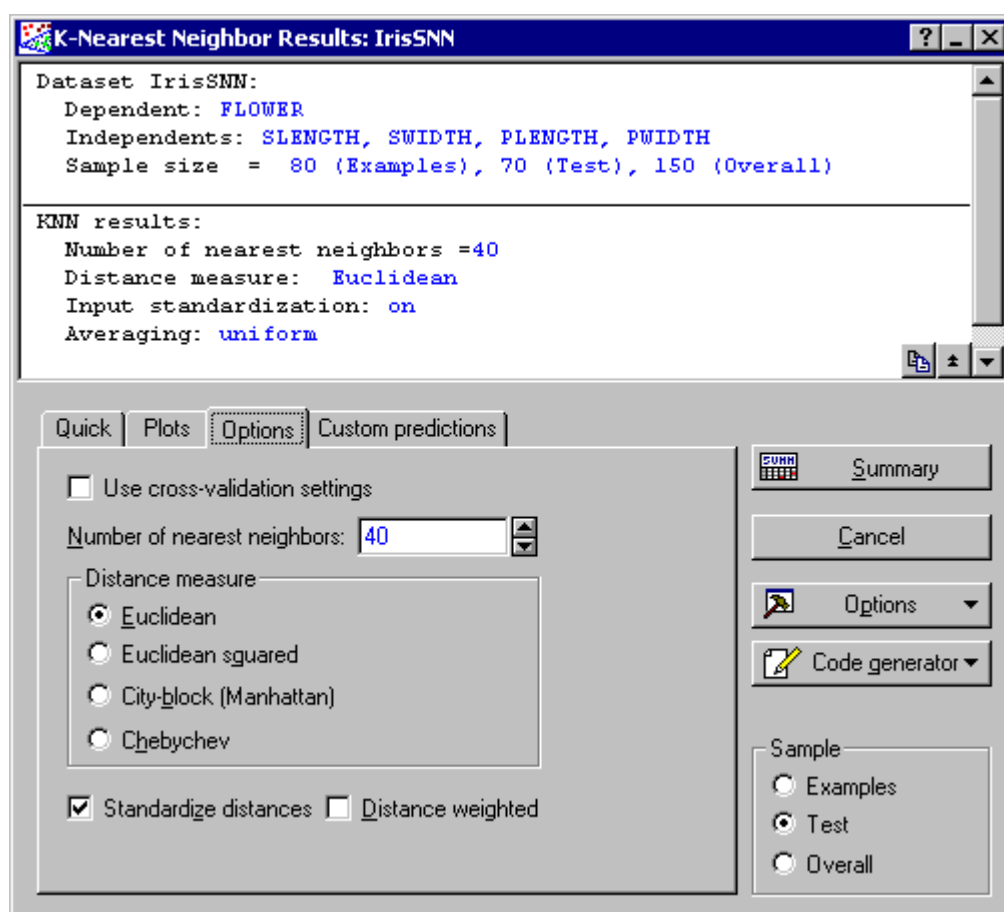


Например, выше показан график рассеяния независимой переменной PLENGTH от уровня достоверности категориальных уровней Setosa, Versicol и Virginic. Из отображаемых (относительных) значений ясно, что Setosa хорошо разделена, в то время как существует некоторая степень перекрытия между Versicol и Virginic. Это перекрытие объясняет неверно классифицированные случаи, которые вы можете просмотреть в таблице прогнозов (см. Ниже). Чтобы создать показанный график, в диалоговом окне K-Nearest Neighbours Results нажмите кнопку параметра Test в

рамке группы Sample, выберите PLENGTH из списка оси X и Setosa (conf.), Versicol (conf.) и Virginic (conf.) из списка оси Y. Затем нажмите кнопку «Графики X и Y».

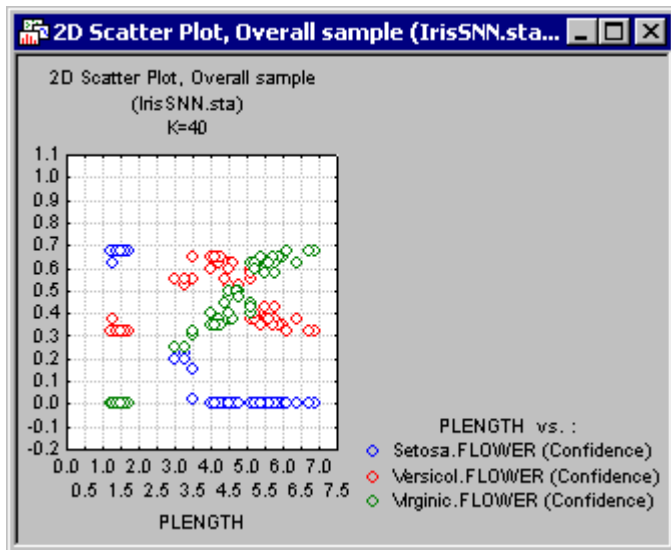
Обратите внимание, что вы можете определить образец (подмножество), для которого вы хотите отобразить результаты. Сделайте это, сделав выбор в рамке группы Sample в диалоговом окне K-Nearest Neighbours Results. Например, нажмите кнопку выбора «Общее» (Overall), чтобы включить как примеры, так и тестовые образцы в электронные таблицы и графики. Однако обратите внимание, что вы не можете делать прогнозы (и другие связанные переменные, например, точность или достоверность) для выборки примеров, поскольку она используется KNN для прогнозирования тестовой выборки.

Поскольку в анализе KNN нет подгонки модели, результаты, которые вы можете получить из диалогового окна «Результаты», ни в коем случае не ограничиваются спецификациями, сделанными в диалоговом окне K-ближайших соседей. Чтобы продемонстрировать это, выберите вкладку «Параметры» в диалоговом окне «Результаты» и снимите флажок «Использовать параметры перекрестной проверки». Это активирует остальные элементы управления на этой вкладке, которые в противном случае недоступны. (Примечание. Это действие не приведет к отмене результатов перекрестной проверки. Вы всегда можете повторно установить этот параметр для анализа, снова установив флажок). Измените количество ближайших соседей на 40.



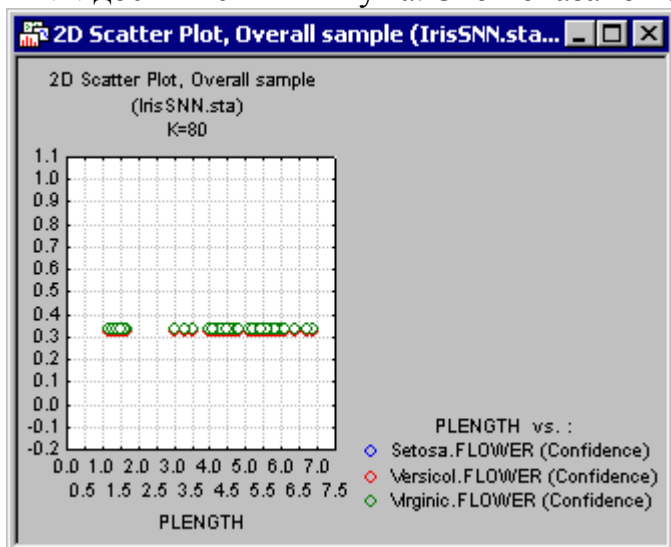
На вкладке «Графики» (Plots) выберите независимую переменную PLENGTH в качестве оси X и уровни достоверности для уровней классов (как и раньше).

Нажмите кнопку «График X и Y». Обратите внимание, что из-за большего значения K (по сравнению с оценкой перекрестной проверки 3) прогнозы уже начали ухудшаться. Вы легко можете увидеть это, заметив большую область перекрытия между уровнями классов.

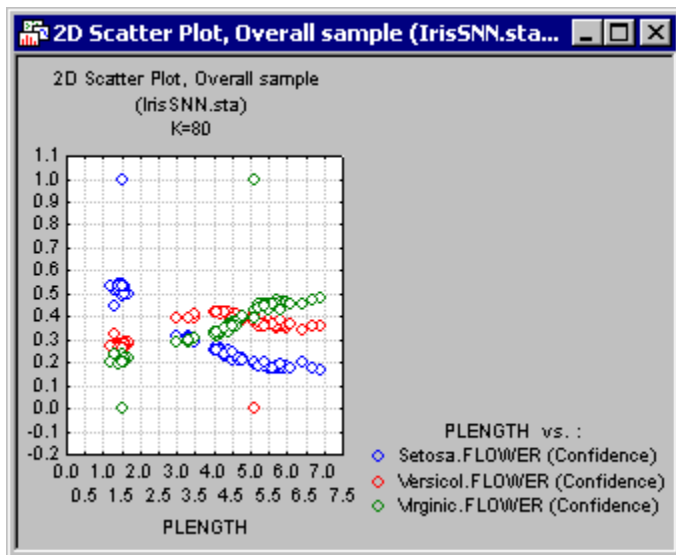


На этом этапе вы можете отобразить новую таблицу описательной статистики (вкладка «Быстрая») для сравнения с предыдущей.

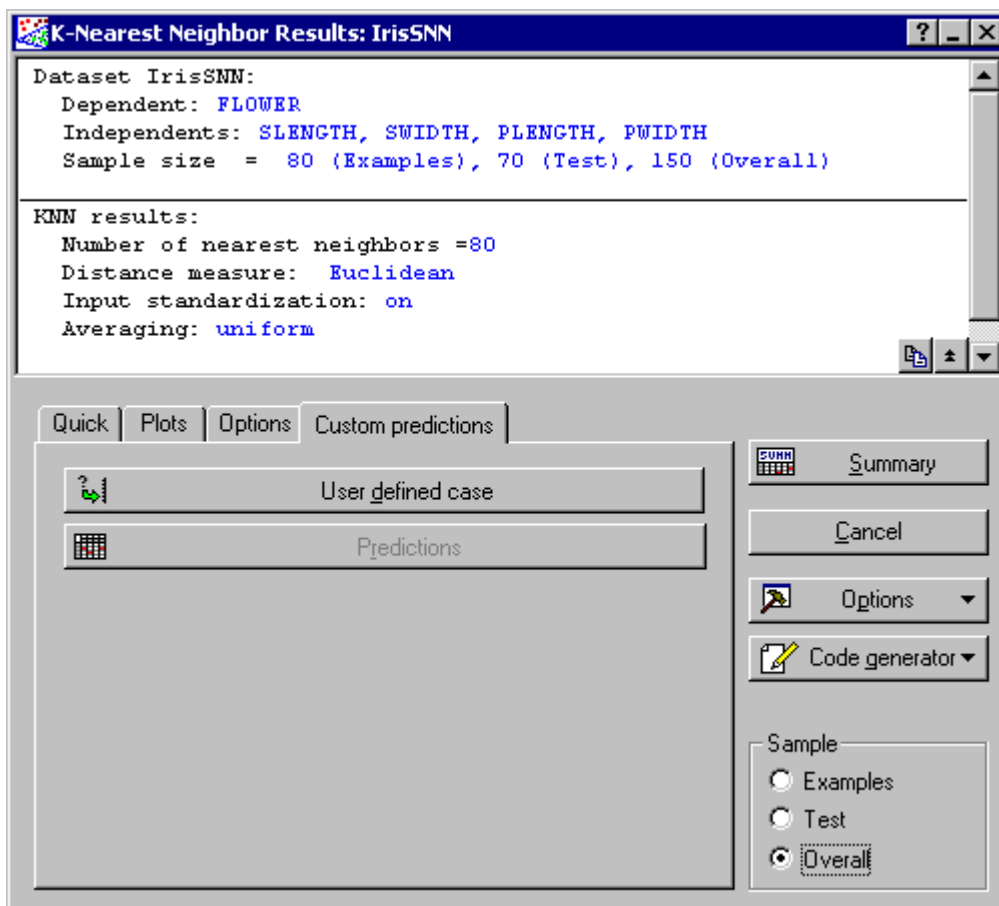
Действительно, если вы установите K = 80 (размер выборки примеров), точность KNN достигнет минимума. Это показано на рисунке ниже.



Далее изучим влияние взвешивания расстояний на прогнозы KNN. Еще раз выберите вкладку «Параметры» и установите флажок «Взвешенное расстояние» (Distance weighted). Оставьте K на 80 и еще раз постройте график PLENGTH относительно уровней достоверности. Несмотря на то, что включены все кейсы из выборки примеров, перекрытия между уровнями классов значительно меньше. Это означает, что более правильная классификация возможна, несмотря на включение всех случаев, принадлежащих к выборке примеров.

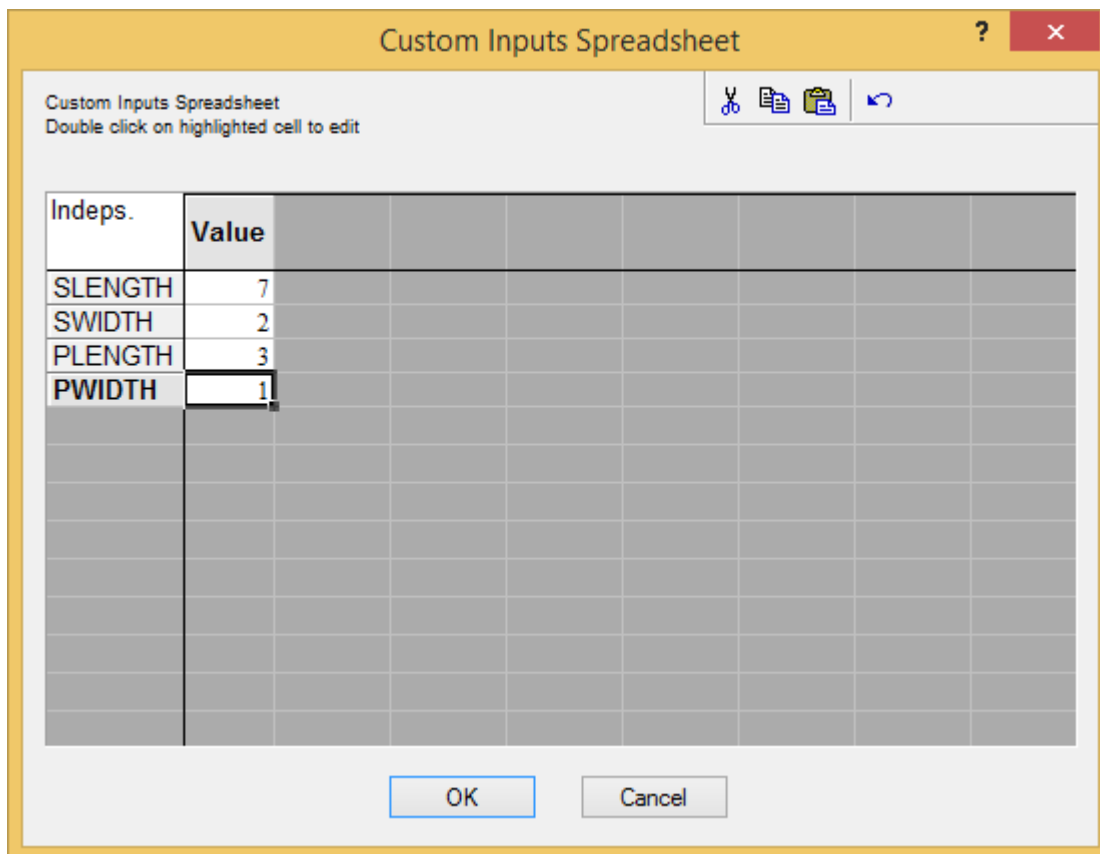


Наконец, выполним «а что, если?» анализ. Выберите вкладку Пользовательские прогнозы (Custom predictions).



Здесь вы можете определить новые наблюдения (которые не взяты из набора данных) и выполнить модель KNN. Это позволяет нам задать вопрос «А что, если?».

Для демонстрации этого, в окне Options вернем K=1, выбираем вкладку Пользовательские прогнозы (Custom predictions), задаем пользовательский паттерн (User defined case), например:



Нажмите кнопку Прогнозы (Predictions), чтобы отобразить прогноз модели.

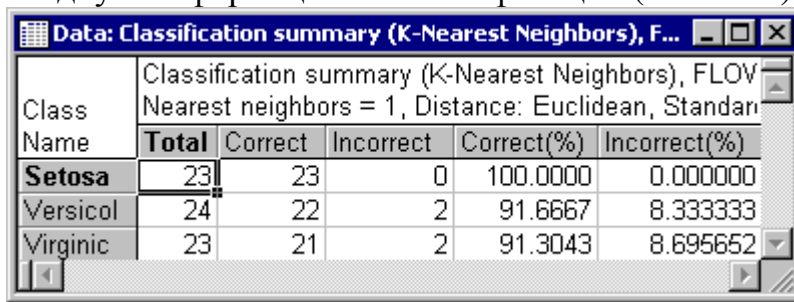
Independent and dependent variables	K-Nearest Neighbors Cust
SLENGTH	7,000000
SWIDTH	2,000000
PLENGTH	3,000000
PWIDTH	1,000000
FLOWER	Versicol

Видим, что введенный паттерн отнесен к классу Versicol.

6 Контрольные вопросы

1. Математическая постановка задачи классификации
2. Метод k-ближайших соседей (KNN)
3. Понятие и необходимость нормализация выборки. Минимакс-нормализация, Z-нормализация (привести выражения)
4. Выделение значимых атрибутов, взвешенный способ
5. «Перекрестная проверка (Cross validation) v-образной формы», ее цель в задаче KNN

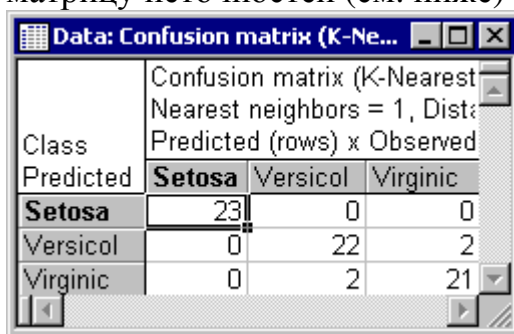
6. На основании чего определяется параметр K в задаче KNN, обеспечивающий максимальную точность классификации (показать на примере своего отчета)?
7. Дать объяснения и интерпретацию данных таблицы (из примера), содержащей сводную информацию о классификации (см. ниже)



Class Name	Total	Correct	Incorrect	Correct(%)	Incorrect(%)
Setosa	23	23	0	100.0000	0.000000
Versicol	24	22	2	91.6667	8.333333
Virginic	23	21	2	91.3043	8.695652

Дать объяснения и интерпретацию данных аналогичной таблицы, полученной в лабораторной работе

8. Дать объяснения и интерпретацию данных таблицы (из примера), содержащей матрицу неточностей (см. ниже)



Class Predicted	Setosa	Versicol	Virginic
Setosa	23	0	0
Versicol	0	22	2
Virginic	0	2	21

Дать объяснения и интерпретацию данных аналогичной таблицы, полученной в лабораторной работе

9. Как вы можете определить новые наблюдения (которые не взяты из набора данных), выполнить модель KNN и осуществить пользовательский прогноз (показать на своей модели).