

Классификация с обучением



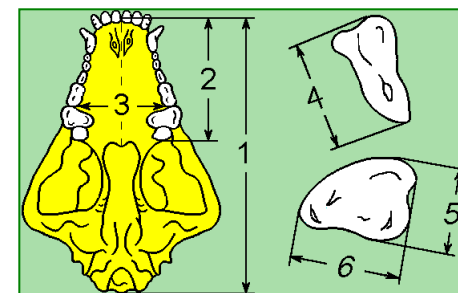
Переобучение (переподгонка, англ. overfitting) в машинном обучении и статистике — явление, когда построенная модель хорошо объясняет примеры из обучающей выборки, но плохо работает на примерах, не участвовавших в обучении (на тестовой выборке). Это связано с тем, что в обучающей выборке обнаруживаются некоторые случайные закономерности, которые отсутствуют в генеральной совокупности.

Собака или волк?



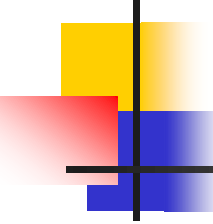
В файле [Dog_Wolf.txt](#) содержатся (стандартизированные) данные о размерах челюстей и зубов 12 волков, 30 собак и одного неизвестного животного.

- X1 — длина черепа,
- X2 — длина верхней челюсти,
- X3 — ширина верхней челюсти,
- X4 — длина верхнего карнива,
- X5 — длина первого верхнего моляра,
- X6 — ширина первого верхнего моляра.

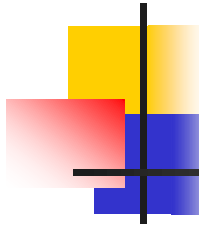


Требуется выяснить, к какому виду (собакам или волкам) вероятнее всего принадлежало неизвестное животное.

Визуальная дискриминация

- 
- Импортируйте файл `Dog_Wolf.txt` под именем `d`
 - Постройте матричную диаграмму рассеяния с раскраской:
`plot(d, col=d$TYPE)`
[`col` обозначает цвет точки на диаграммах рассеяния; в столбце `TYPE` указан номер вида животных: 1 – собака, 2 – волк, 3 – неизвестное животное; символ `$` связывает имя таблицы данных (`d`) с именем столбца этой таблицы (`TYPE`)]
 - Нажмите кнопку `Zoom` и максимизируйте окно диаграммы
 - Дайте ответы на следующие вопросы:
 - 1) На диаграммах рассеяния каких пар признаков собаки (чёрные кружки) и волки (красные кружки) разделяются лучше всего?
 - 2) На какой вид проецируется неизвестное животное (зелёный кружок), если судить только по шести крайне правым диаграммам?
 - 3) К какому виду вероятнее всего относится неизвестное животное?

Продолжение



- Установите пакет `plot3D`: на вкладке `Packages` нажмите кнопку `Install Packages`, введите имя пакета и нажмите `Install`
- Подключите пакет `plot3D`, поставив перед ним «галку» в списке пакетов на вкладке `Packages`
- Постройте трёхмерную диаграмму рассеяния (точечную диаграмму) для признаков `X5`, `X6`, `X4`:

```
scatter3D(d$X5, d$X6, d$X4, colvar=d$TYPE, type="h",  
phi=30, theta=10, d=5, bty="g", pch=20)
```

[Значения аргументов:

`colvar` — цвета точек в соответствии с видом животного,
`type` — тип трёхмерной диаграммы рассеяния
(`type="h"` означает «точки на вертикальных ножках»),
`phi` и `theta` — углы поворота кубика,
`d` — параметр перспективы,
`bty` — вид кубика (*от англ. box type*),
`pch` — тип точек на диаграмме (*от англ. point character*)]

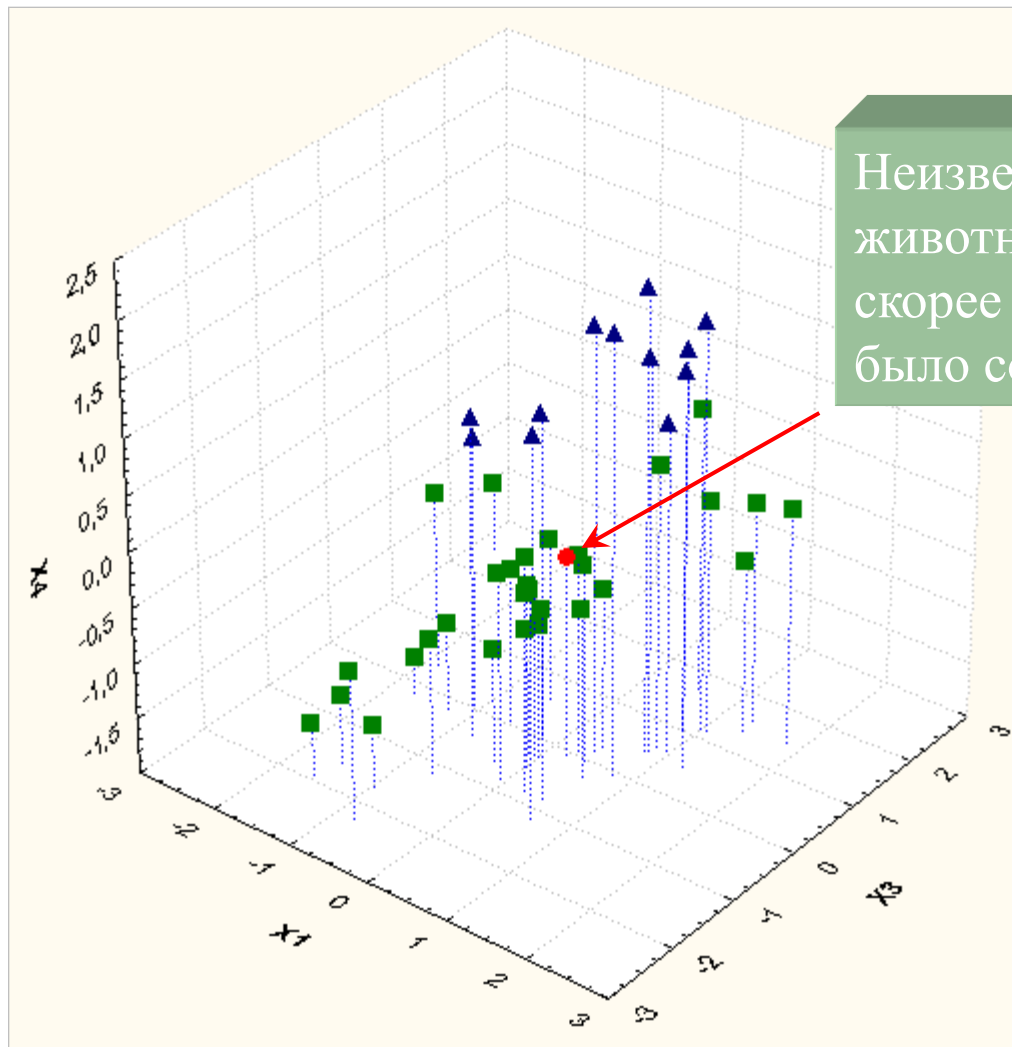
Результаты дискриминации

Справа
приведена

трёхмерная
диаграмма
рассеяния

для признаков
 X_1 , X_3 , X_4 .

Точки
раскрашены
в три цвета
в соответствии
со значениями
признака **TYPE**.



Неизвестное
животное,
скорее всего,
было собакой

■ TYPE = 1
▲ TYPE = 2
● TYPE = 3

Дискриминация по ближайшим соседям

k-Nearest Neighbour Classification

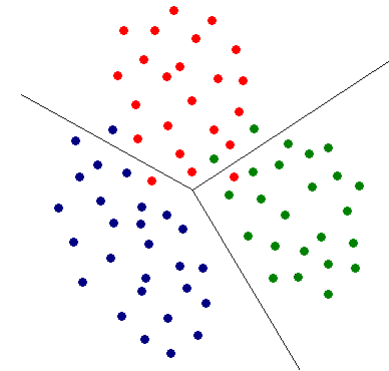
Description

k-nearest neighbour classification for test set from training set. For each row of the test set, the k nearest (in Euclidean distance) training set vectors are found, and the classification is decided by majority vote, with ties broken at random. If there are ties for the k th nearest vector, all candidates are included in the vote.

- Подключите пакет `class`, поставив перед ним «галку» [ties — совпадения]
- Возьмите из таблицы `d` без 7-го столбца качестве обучающей выборки первые 40 строк, а в качестве контрольной — оставшиеся 3 строки:
`train=d[1:40,-7]; test=d[41:43,-7]`
- Задайте классификацию на обучающей выборке с помощью команды `factor`, создающий вектор в номинальной шкале из названий классов:
`cl=factor(c(rep("dog", 30), rep("wolf", 10)))` [cl — class, rep — repetition]
- Классифицируйте (дискриминируйте) объекты из контрольной выборки с учётом 10 ближайших соседей:
`knn(train, test, cl, k = 10, prob=TRUE)` [prob — вероят. принадл. классу]

Базовые методы дискриминации

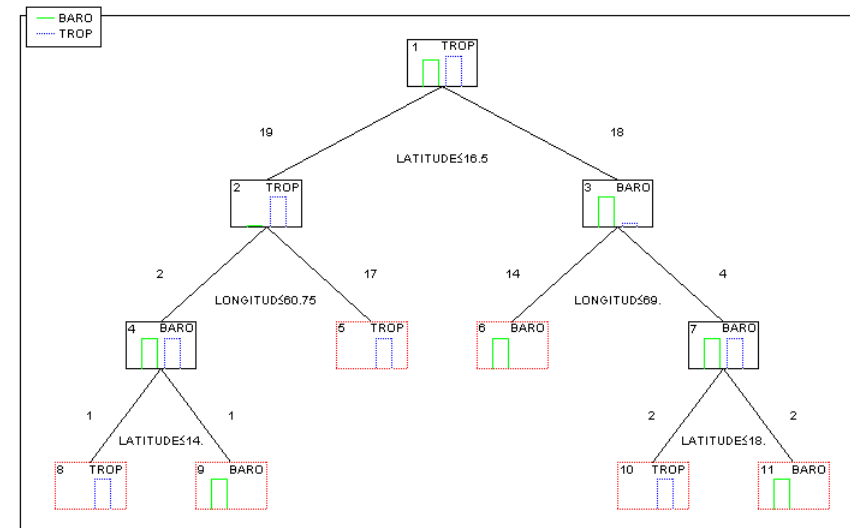
- **Дискриминантный анализ**
[пересекающиеся эллиптические «облака»; построение разделительных поверхностей (гиперплоскостей)]



- **Логистическая регрессия**
[отклик принимает только значения 0 и 1; оценивание регрессионных коэффициентов]

$$Y = 1/(1 + \exp\{-b_0 - b_1X_1 - \dots - b_mX_m\}) < 1/2$$

- **Деревья принятия решения**
[классы имеют форму объединений многомерных параллелепипедов; построение деревьев принятия решений]



Двумерная нормальная плотность

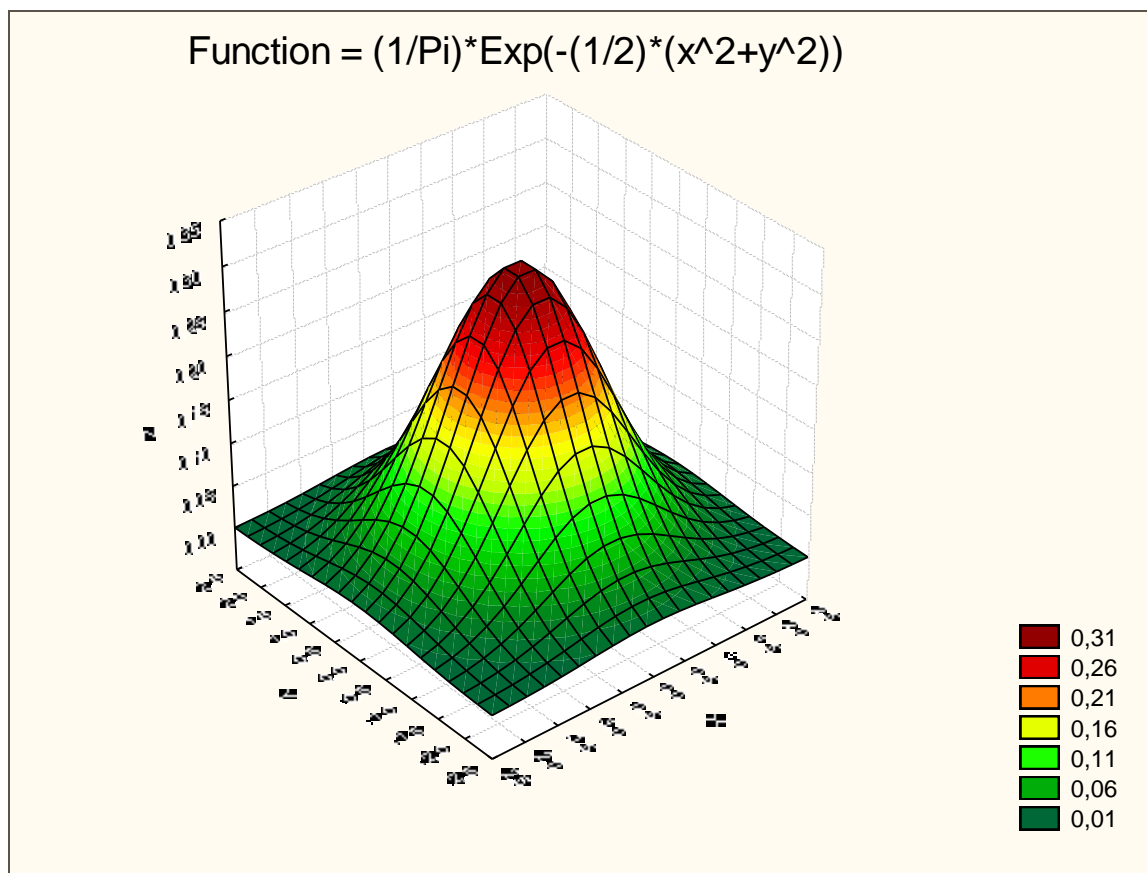
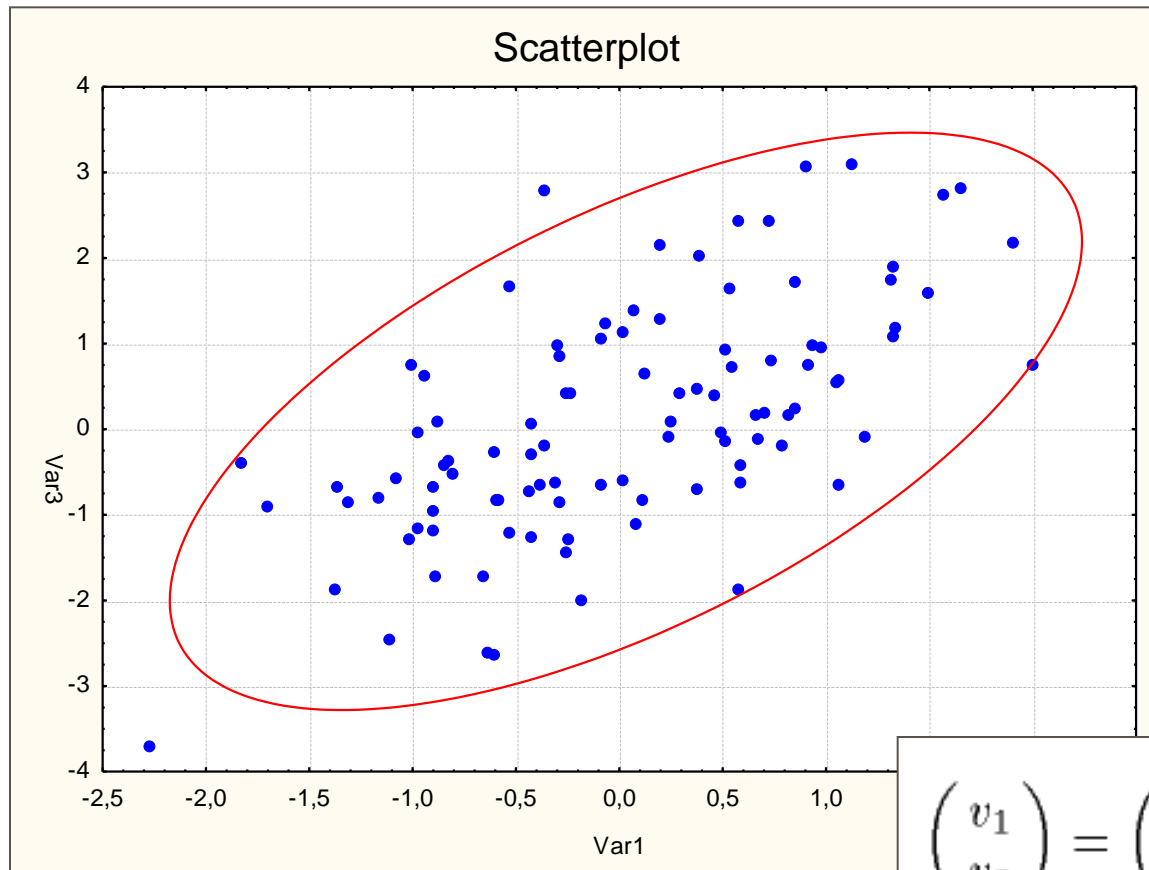



Диаграмма рассеяния двумерного нормального вектора



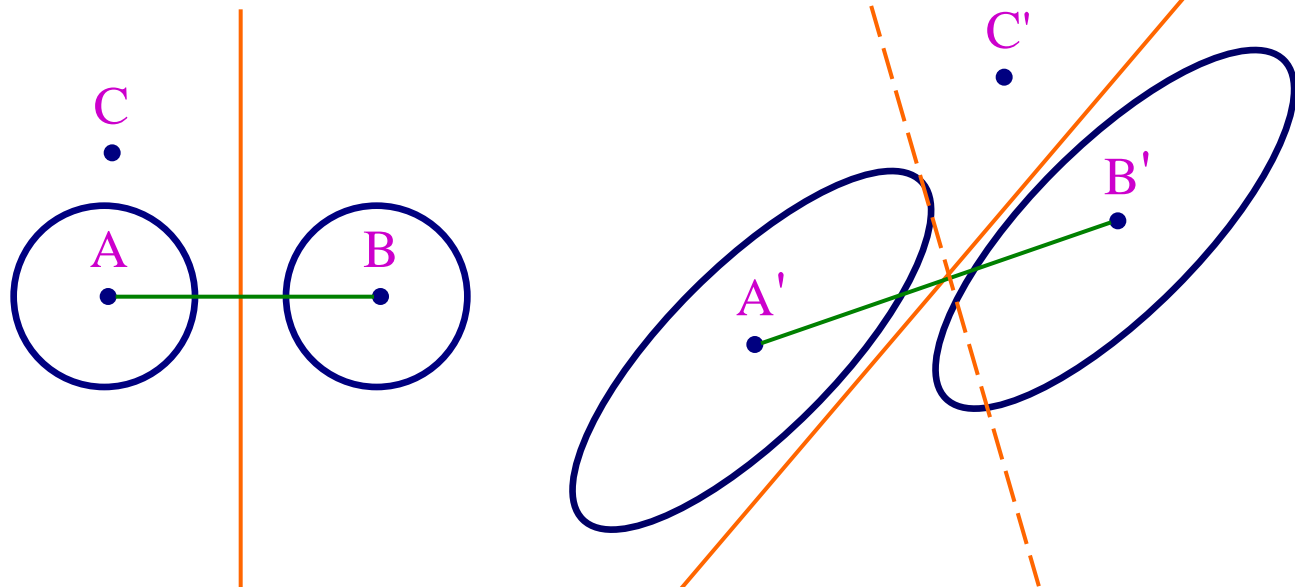
$$\begin{pmatrix} v_1 \\ v_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}$$

Оценка двумерной плотности

- 
- Подключите пакет `MASS`, поставив «галку» в окне `Packages`
 - Моделируйте выборку (x, y) размера n из двумерного нормального распределения, постройте диаграмму рассеяния:
`n=100; x=rnorm(n); y=x+rnorm(n); plot(y~x)`
 - Вычислите оценку двумерной плотности: `z=kde2d(x, y, h=2)`
[two-dimensional kernel density estimation, h — ширина «окна»]
 - Постройте цветовую карту оценки плотности: `image(z)`
 - Задайте нижний и верхний уровни для z (уровни сечений):
`image(z, zlim=c(0, 0.05))` [z limits]
 - Постройте карту линий уровня оценки плотности:
`contour(z, col="red", drawlabels=FALSE)` [не выводить уровни]
 - Постройте изображение поверхности оценки плотности:
`persp(z, phi=30, theta=20, d=5)` [нажмите кнопку `Zoom`]
 - Измените ширину «окна» сглаживания h на `1` и снова постройте изображение поверхности оценки плотности:
`z=kde2d(x, y, n=50, h=1)` [n — число «узлов» сетки]
`persp(z, phi=30, theta=20, d=5, col="yellow", shade=.5)`

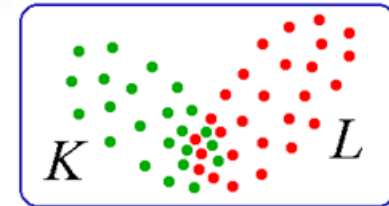
Дискриминация на основе расстояния Махаланобиса

$$d^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x} - \mathbf{y})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \mathbf{y})$$



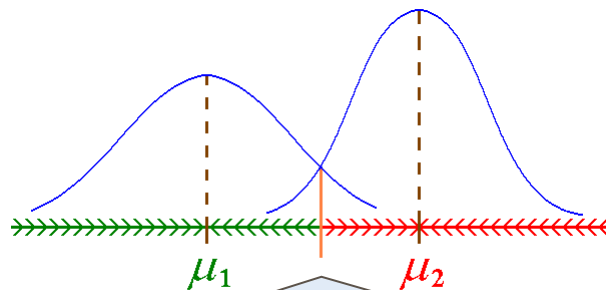
В случае классов эллиптической формы дискриминировать новый объект можно по расстояниям Махаланобиса до центров классов, а в случае классов сложной формы — по расстояниям Махаланобиса до центров «ядер» классов (областей наиболее высокой сгущённости).

Дискриминантный анализ



Приведем метод разделения *многомерных нормальных* выборок на классы, базирующийся на принципе максимального правдоподобия. В соответствии с этим принципом будем считать *областью притяжения* закона $\mathcal{N}(\mu_l, \Sigma_l)$ ($l = 1, \dots, k$) множество таких точек $\mathbf{x} = (x_1, \dots, x_m) \in \mathbf{R}^m$, где плотность распределения $\mathcal{N}(\mu_l, \Sigma_l)$ больше других. Это равносильно тому, что пропорциональная логарифму плотности величина

$$h_l(\mathbf{x}) = \ln \det \Sigma_l + (\mathbf{x} - \mu_l)^T \Sigma_l^{-1} (\mathbf{x} - \mu_l),$$

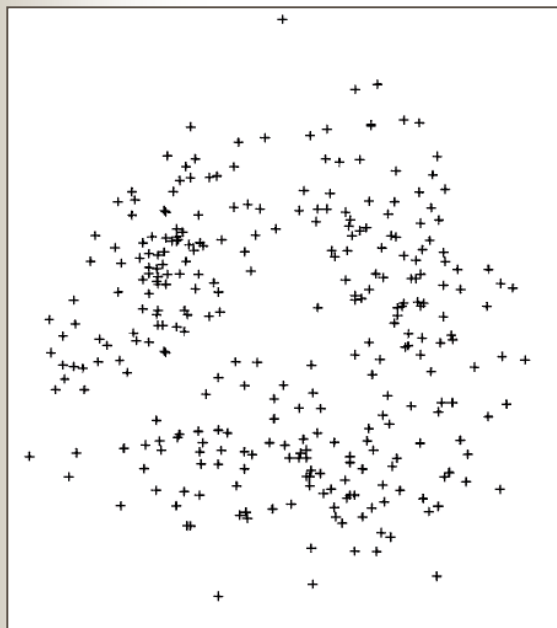


имеет наименьшее значение среди h_1, \dots, h_k . Рисунок иллюстрирует получаемое при $m = 1$ и $k = 2$ разбиение прямой \mathbf{R} на две области притяжения.

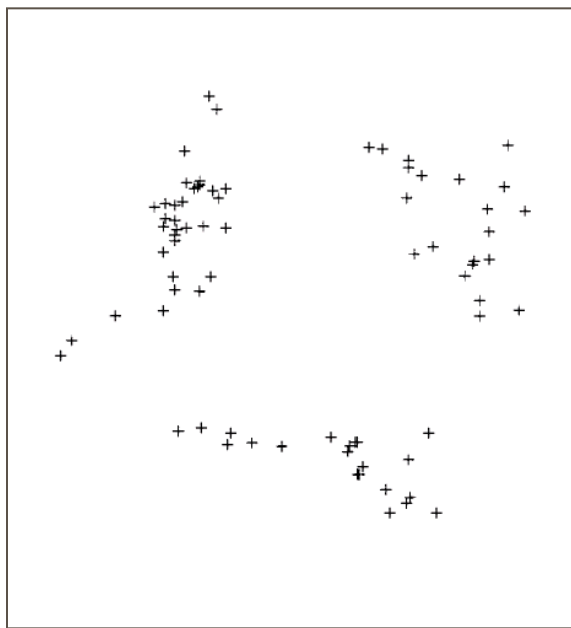
Кроме расстояния до центра класса учитывается также форма и размер класса

При таком выборе границы минимизируется сумма вероятностей ошибочной классификации

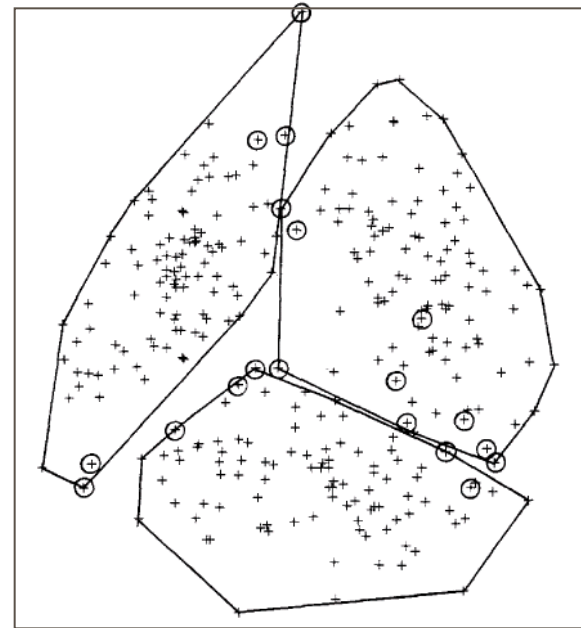
Дискриминация с предварительным выделением сгущений объектов



3 выборки (по 100 точек
в каждой) из трёх
разных двумерных
нормальных законов



25% объектов («ядра»
классов) с наименьшей
суммой расстояний до
ближайших 5 соседей



Разбиение на основании
принципа максимального
правдоподобия (ошибки
обведены кружками)



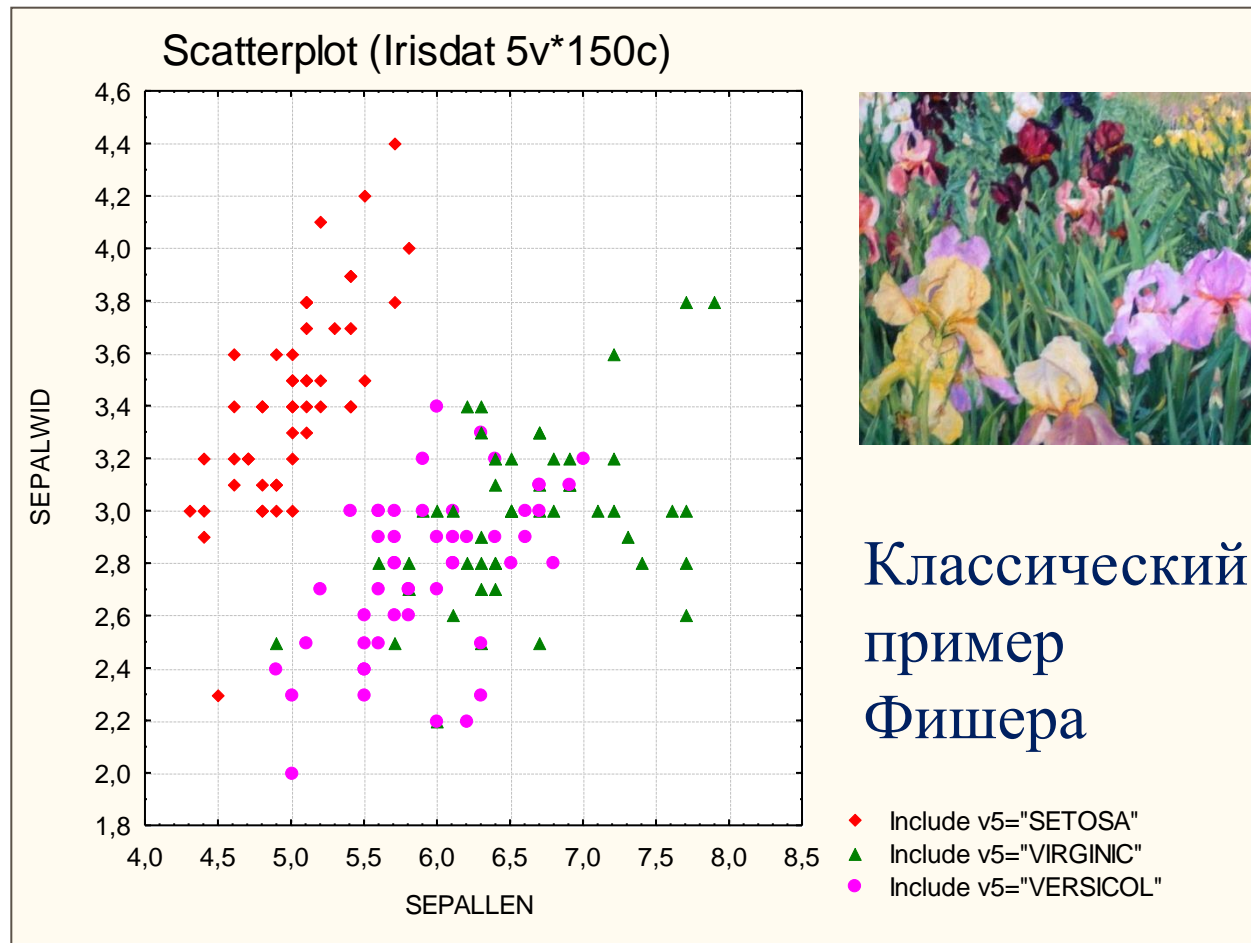
Допущения в модели Фишера

В классической дискриминантной модели Фишера предполагается, что

- 1) все признаки измерены в **непрерывной шкале**,
- 2) более того, совместное распределение признаков является **многомерным нормальным**,
- 3) **ковариационные матрицы одинаковы** во всех классах.

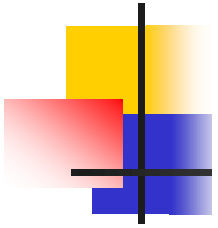
Для визуальной проверки двух последних допущений следует построить **матричную** диаграмму рассеяния с раскраской классов с помощью команды **plot**, аргументу **col** которой присвоен вектор, задающий цвета классов.

Размеры чашелистиков и лепестков ирисов



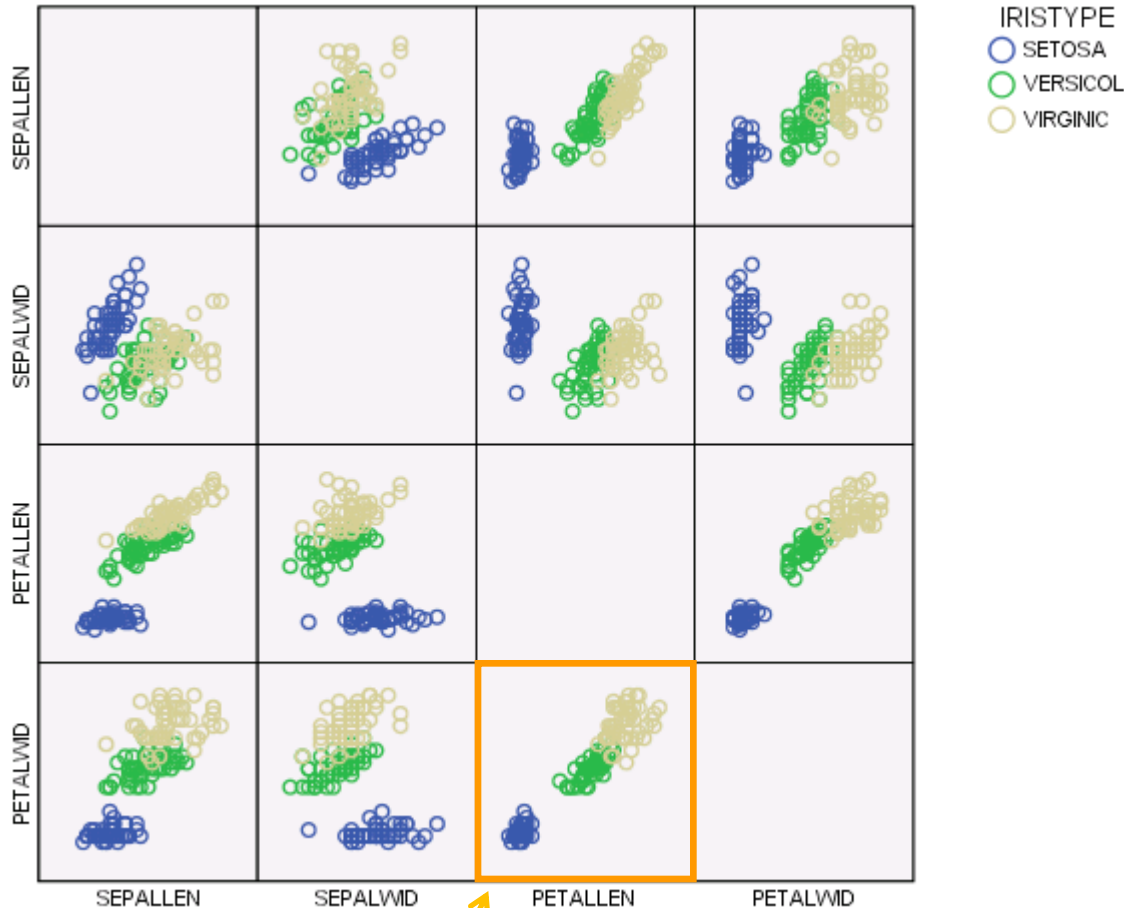
sepal — чашелистик, petal — лепесток

Визуальный анализ данных об ирисах



- Удалите признак `Species` из таблицы данных `iris`: `d = iris[,-5]`
- Классифицируйте методом *K*-средних на 3 класса:
`k=kmeans(d, centers=3, nstart=10)`
- Выведите диаграмму рассеяния с раскрашенными классами:
`plot(d, col=k$cluster, pch=16)`
[символ № 16 — закрашенные кружки]
- Нажмите кнопку `Zoom` и максимизируйте окно диаграммы
По какой паре признаков классы разделяются лучше всего?
- Подключите пакет `cluster`, поставив перед ним «галку»
- Примените алгоритм `pam` для кластеризации на 3 класса:
`p=pam(d, 3)` [`pam` является модификацией метода *K*-средних]
- Раскрасьте новые классы: `plot(d, col=p$clustering, pch=16)`
- Раскрасьте точки в соответствии с *видом ириса* (`Species`):
`plot(d, col=iris$Species, pch=16)`
- Нажимая кнопки  и  сравните 3 разбиения визуально
- Используя команду `table`, подсчитайте случаи неправильного определения вида ириса методами *K*-средних и `pam`

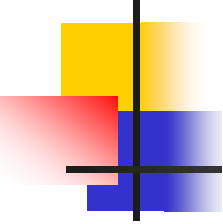
Матричная диаграмма рассеяния



Слева приведена построенная с помощью статистического пакета SPSS матричная диаграмма рассеяния для данных об ирисах с окраской точек в соответствии с признаком IRISTYPE.

Отметим, что на диаграмме рассеяния для признаков PETALLEN – PETALWID почти полностью отсутствует наложение классов SETOSA, VIRGINIC и VERSICOL.

Продолжение: дискриминация ирисов



Чтобы продемонстрировать, как выполняется дискриминация объектов в модели Фишера [linear discriminant analysis], рассмотрим её на примере с ирисами.

- Выберите из 150 чисел случайное подмножество размера 75:
`train=sample(1:150, 75)`
- Узнайте, сколько в нём ирисов разных видов:
`table(iris$Species[train])`
- Постройте модель Фишера для обучающей выборки train:
`m=lda(d, iris$Species, prior=c(1,1,1)/3, subset=train)`
[prior — априорные вероятности принадлежности классам]
- Постройте диаграмму рассеяния объектов: `plot(m)`
[LD1, LD2 — две первые главные оси выб. ковар. матрицы]
- Запомните предсказанные значения для тестовой выборки:
`pred=predict(m, d[-train,])$class`
- Запомните наблюдаемые виды ирисов для тестовой выборки:
`obs=iris[-train,]$Species`
- Постройте перекрёстную таблицу для наблюдаемых и предсказанных видов ирисов: `table(obs, pred)`

Relative frequencies of tag trigrams is selected Spanish texts

Description

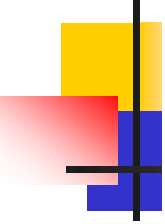
Relative frequencies of the 120 most frequent tag trigrams in 15 texts contributed by 3 authors.

```
> spanishMeta
```

	Author	YearOfBirth	TextName	PubDate	Nwords	FullName
1	C	1916	X14458g11	1983	2972	Cela
2	C	1916	X14459g11	1951	3040	Cela
3	C	1916	X14460g11	1956	3066	Cela
4	C	1916	X14461g11	1948	3044	Cela
5	C	1916	X14462g11	1942	3053	Cela
6	M	1943	X14463g11	1986	3013	Mendoza
7	M	1943	X14464g11	1992	3049	Mendoza
8	M	1943	X14465g11	1989	3042	Mendoza
9	M	1943	X14466g11	1982	3039	Mendoza
10	M	1943	X14467g11	2002	3045	Mendoza
11	V	1936	X14472g11	1965	3037	VargasLLOsa
12	V	1936	X14473g11	1963	3067	VargasLLOsa
13	V	1936	X14474g11	1977	3020	VargasLLOsa
14	V	1936	X14475g11	1987	3016	VargasLLOsa
15	V	1936	X14476g11	1981	3054	VargasLLOsa

Задача — выяснить, можно ли установить авторство текстов на основе частотной информации, представленной в таблице spanish.

Дискриминантный анализ авторства

- 
- Нажмите на вкладке `Packages` кнопку `Install`, введите имя пакета `languageR` и установите его. Затем подключите пакет `languageR`, поставив «галку» перед ним на вкладке `Packages`
 - Транспонируйте таблицу `spanish`: `s=t(spanish)`
 - Упорядочите строки таблицы `s` по их именам (для соответствия с порядком строк в `spanishMeta`): `s=s[order(rownames(s)),]`
 - Посмотрите на начальный фрагмент таблицы: `s[1:5,1:4]`
 - Вычислите главные компоненты: `p=prcomp(s, scale=TRUE)`
[На основе корреляционной, а не ковариационной матрицы]
 - Доли дисперсии, приходящиеся на компоненты: `summary(p)`
Какова доля разброса, приходящаяся на первые 8 компонент?
 - Создайте таблицу данных из проекций объектов на главные оси: `d=data.frame(p$x)`
 - Примените модель Фишера к первым 8 главным компонентам: `m=lda(d[,1:8], spanishMeta$Author)`
 - Постройте диаграмму рассеяния объектов: `plot(m)`
 - Выведите предсказанные вероятности принадлежности к классам (авторства): `round(predict(m)$posterior, 4)`

Логистическая регрессия

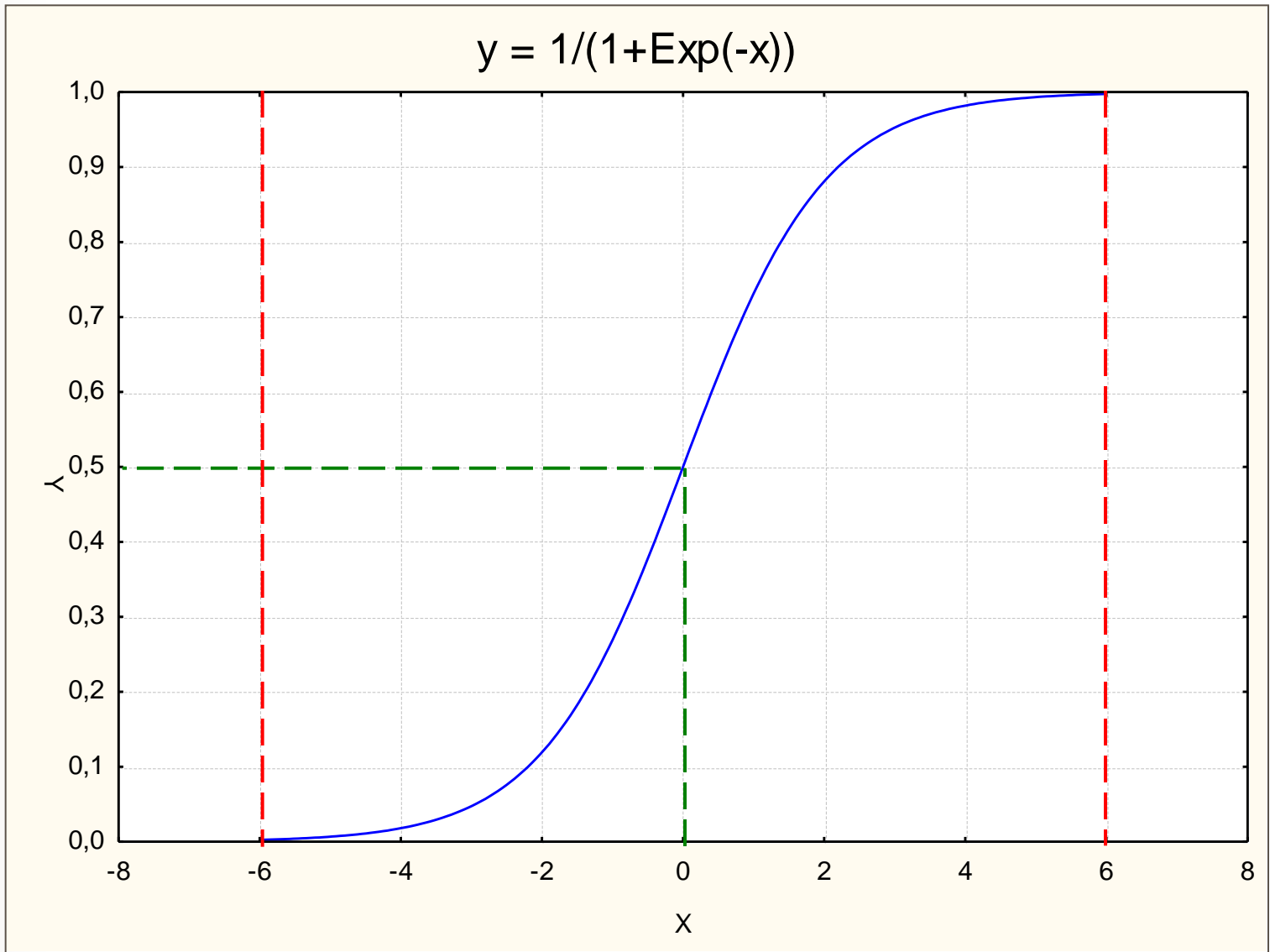
Начнем с обсуждения *бинарной логистической модели*. Она используется для оценки вероятности “успеха” (положительного ответа) в случае, когда возможны только два исхода: “успех” или “неудача” (ответы “да” или “нет” в анкете). Например, сотрудник отдела кредитования некоторого банка оценивает риск невозвращения кредита очередным заемщиком на основе информации о нем.

Поскольку вероятность события — число между 0 и 1, не удастся оценивать ее с помощью обычной линейной регрессионной модели, так как в ней зависимая переменная (“отклик”) может принимать значения, не принадлежащие отрезку $[0, 1]$. Логистическая регрессия является частным случаем так называемой *обобщенной регрессионной модели* (generalized linear model или GLZ):

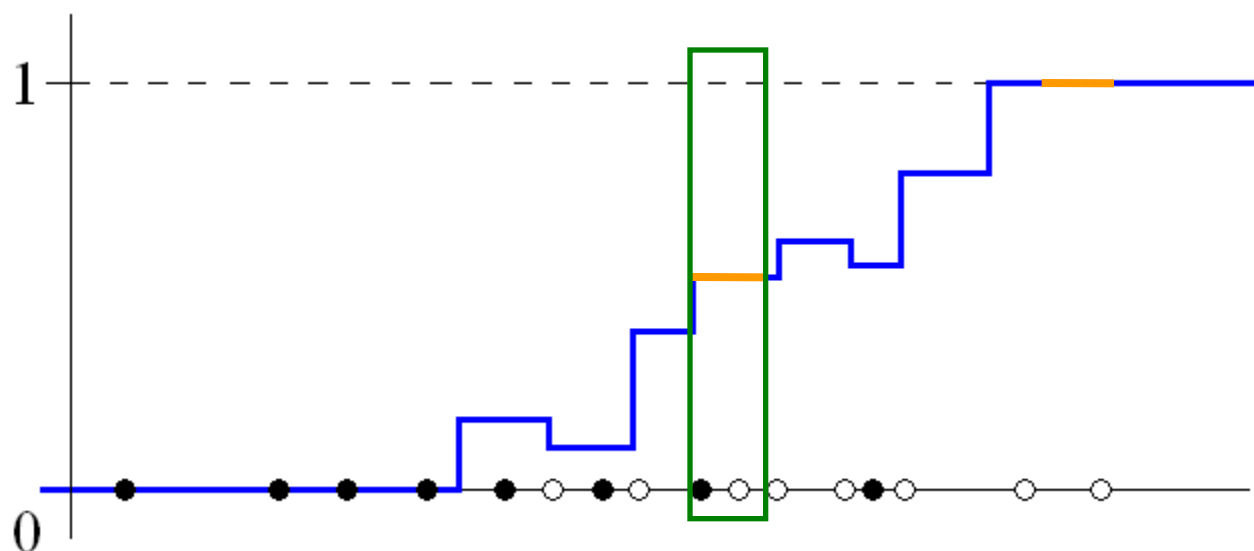
$$\eta = g(\mathbf{X}\boldsymbol{\theta}) + \epsilon,$$

где g — некоторая функция (link function). В случае логистической регрессии используется функция $g(y) = 1/(1 + e^{-y})$, которая переводит действительную прямую в $(0, 1)$.

Логистическая функция



Условная вероятность



$$m(x) = \mathbf{M}(Y|X = x).$$

Здесь $\mathbf{M}(Y|X = x)$ — условное математическое ожидание случайной величины Y при условии события $\{X = x\}$.

В данном случае имеем $m(x) = \mathbf{P}(Y = 1 | X = x)$.

Предполагаем, что при некоторых a и b $m(x) \approx \frac{1}{1 + e^{-a-bx}}$.

Вероятность возвращения кредита

В файле `BankLoan.txt` содержатся некоторые сведения о лицах, обращающихся за кредитом, в частности:

- 1 – возраст (`age`),
- 2 – уровень образования (`ed`),
- 3 – непрерывный стаж на последней работе (`employ`),
- 4 – время проживания по текущему адресу (`address`),
- 5 – годовой доход (`income`),
- 6 – отношение размера имеющегося долга к доходу (`debinc`),
- 7 – долги на кредитных карточках (`creddebt`),
- 8 – другие долги (`othdebt`),
- 9 – информация о возвращении кредита (вернул – 1, нет – 0).

Выясните, какие из указанных факторов значимо на уровне 0,05 влияют на вероятность возвращения кредита с помощью функции `glm` (generalised linear model) с аргументом `family = "binomial"`



Мультиномиальная регрессия

Модель применяется для классификации объектов на основе значений нескольких предикторных переменных и обобщает бинарную логистическую модель на случай, когда зависимая переменная принимает более двух значений.

Например, может проводиться исследование, участникам которого предлагается выбрать один из нескольких конкурирующих товаров. Полиномиальная логистическая регрессия позволяет построить профили (уровни характеристик) людей, наиболее заинтересованных в товаре определенного вида. На основе такого анализа можно разрабатывать рекламную стратегию.

Если зависимая переменная имеет k категорий, то рассматриваются K ненаблюдаемых случайных величин Y_1, \dots, Y_k , каждая из которых понимается как “предрасположенность” к j -й категории, $j = 1, \dots, k$. В случае продажи товаров Y_j выражает предпочтительность для покупателя товара j -го вида: чем больше Y_j , тем больше вероятность покупки именно такого товара. Математически зависимость вероятности j -го исхода от случайных величин Y_1, \dots, Y_k выражается формулой

$$p_j = \frac{e^{Y_j}}{e^{Y_1} + \dots + e^{Y_k}},$$

при этом предполагается, что каждая Y_j линейно связана с предикторами X_1, \dots, X_m : $Y_j = \theta_{0j} + \theta_{1j}X_1 + \dots + \theta_{mj}X_m$.

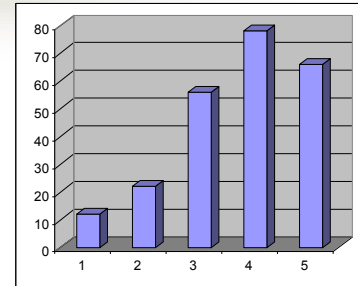
Для построения модели в R следует подключить пакет **nnet** и вызвать функцию **multinom**



Порядковая (ordinal) регрессия

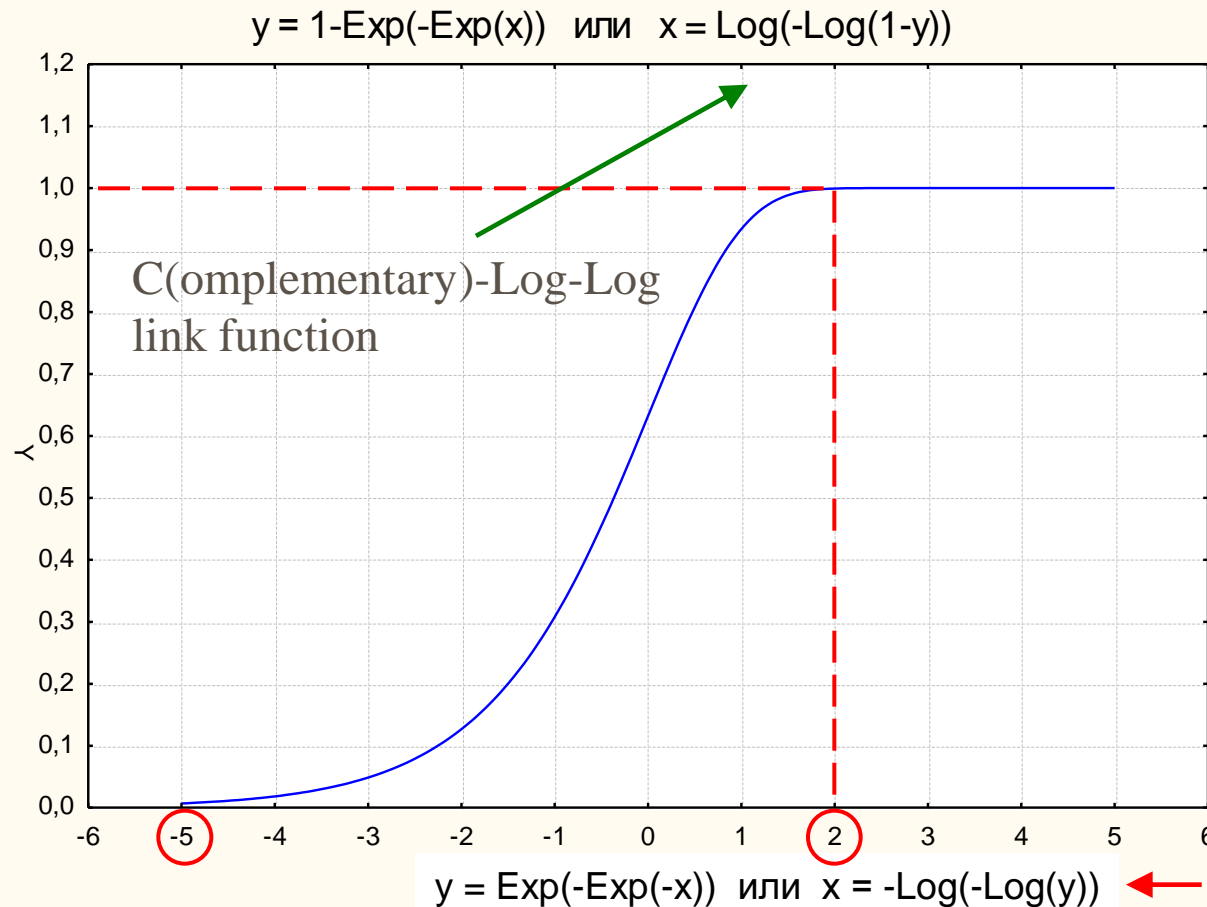
$$Y_j = -\theta_{0j} + \theta_1 X_1 + \dots + \theta_m X_m, \quad j = 1, \dots, k-1;$$

$$\gamma_j = 1 / (1 + e^{-Y_j}), \quad \gamma_j = p_1 + \dots + p_j.$$



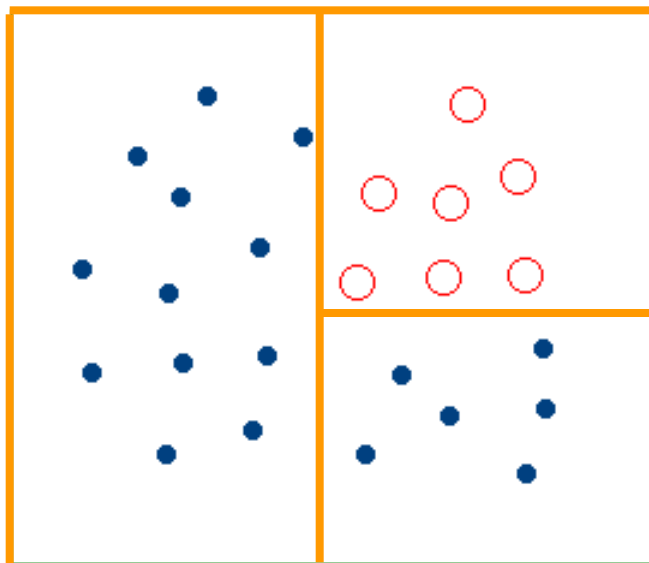
Применяется
в случае
упорядочен-
ности
категорий
отклика

Для
построения
модели в R
следует
подключить
пакет **rms** и
вызвать
функцию **lrm**



Если старшие
категории
отклика более
вероятны, т. е.
левый «хвост»
гистограммы
«тяжелее», то
рекомендуется
применять
C-Log-Log,
в противном
случае —
Log-Log.

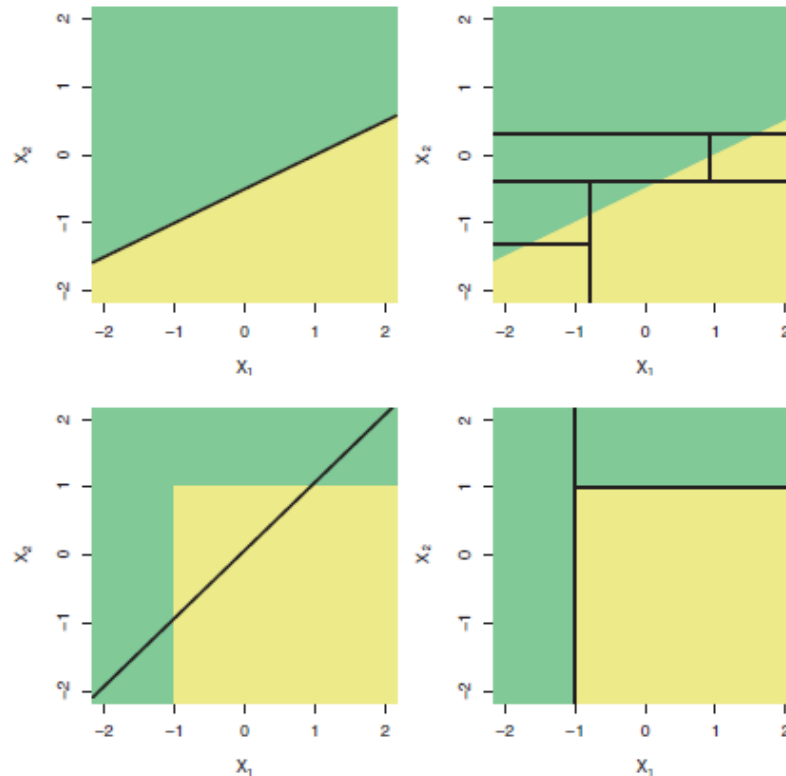
Деревья принятия решения



Один из возможных критериев оптимальности классификации — общее число правильно классифицированных объектов.

	●	○
●	18	6
○	0	7

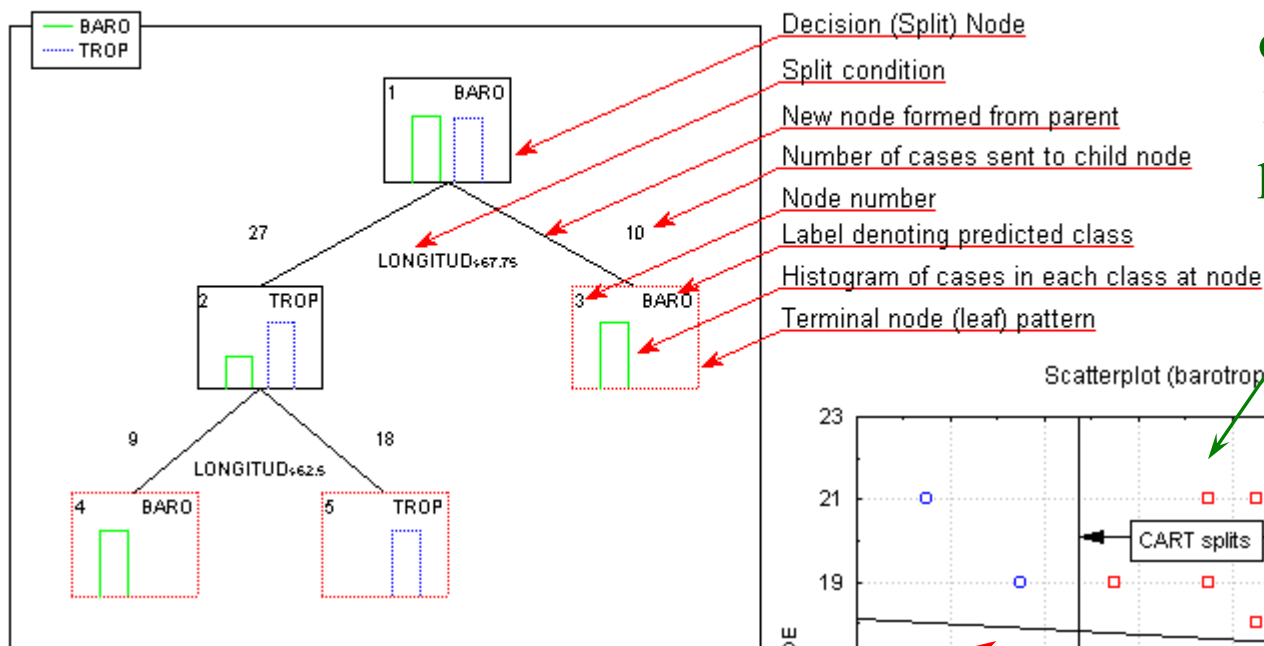
Уместность метода



Top Row: A two-dimensional classification example in which the true decision boundary is linear, and is indicated by the shaded regions. A classical approach that assumes a linear boundary (left) will outperform a decision tree that performs splits parallel to the axes (right). **Bottom Row:** Here the true decision boundary is non-linear. Here a linear model is unable to capture the true decision boundary (left), whereas a decision tree is successful (right).

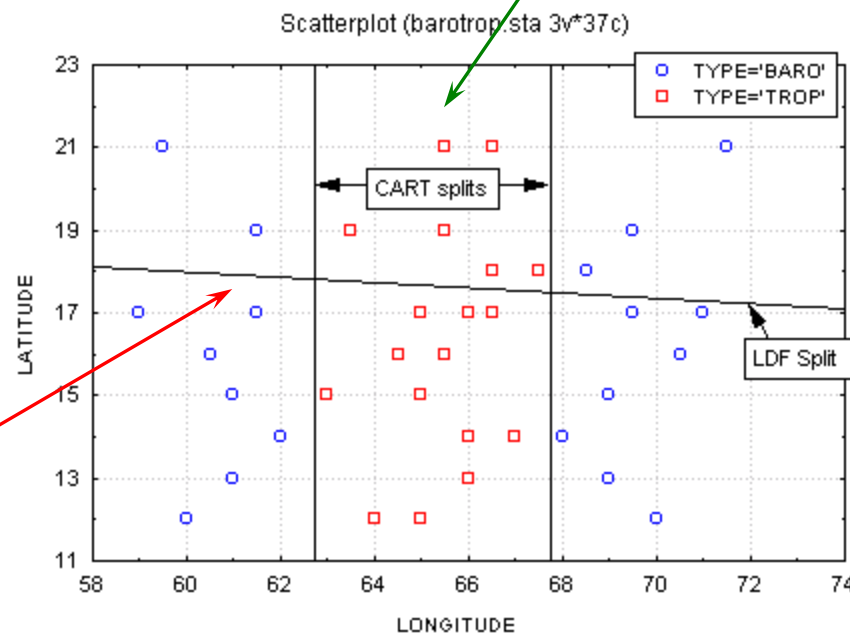
Классификация ураганов

Classification Tree for CLASS
Number of splits = 2; Number of terminal nodes = 3

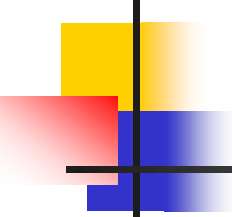


Дерево классификации обеспечивает 100%-ный результат!

Дискриминантный анализ позволяет правильно классифицировать только 54% наблюдений!



Синтаксис дательного падежа

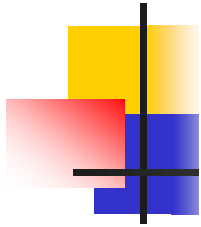


The dependent variable in this study is a factor with levels **NP** (the dative is realized as an NP, as in *John gave Mary the book*) and **PP** (the dative is realized as a PP, as in *John gave the book to Mary*). For 3263 verb tokens in corpora of written and spoken English, the values of a total of 12 variables were determined, in addition to the realization of the dative, coded as `RealizationOfRecipient` in the data set `dative`:

```
> colnames(dative)
[1] "Speaker"           "Modality"
[3] "Verb"              "SemanticClass"
[5] "LengthOfRecipient" "AnimacyOfRec"
[7] "DefinOfRec"        "PronomOfRec"
[9] "LengthOfTheme"     "AnimacyOfTheme"
[11] "DefinOfTheme"      "PronomOfTheme"
[13] "RealizationOfRecipient" "AccessOfRec"
[15] "AccessOfTheme"
```

Short descriptions of these variables are available with `?dative`. The question that we address here is whether the realization of the recipient as NP or PP can be predicted from the other variables. The technique that we introduce here is CART analysis, an acronym for Classification And Regression Trees.

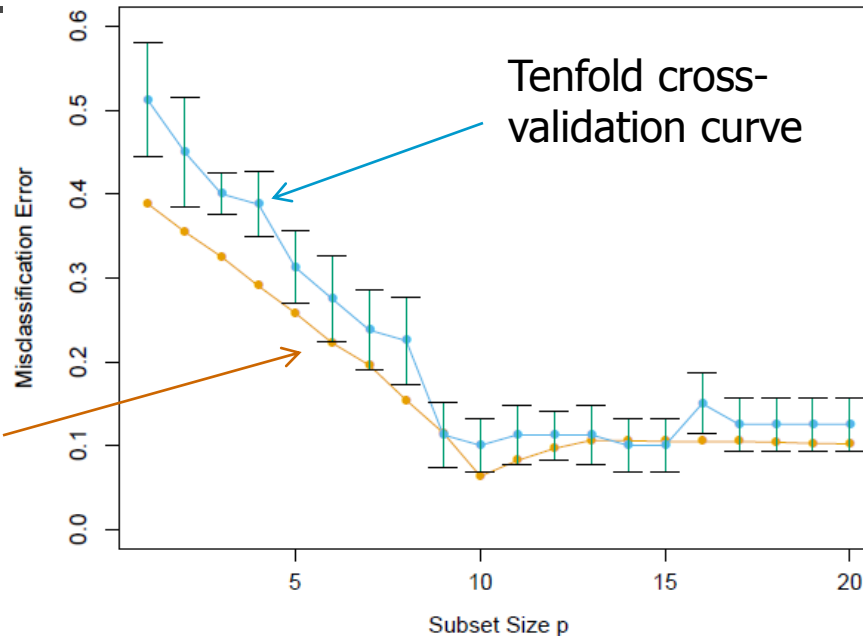
Пример дерева принятия решения



- Уберите лишние номинальные столбцы 1 и 3:
`d=dative[,-c(1,3)]`
- Подключите пакет `rpart`, поставив «галку» в окне `Packages`
- Постройте дерево принятия решения и выведите его на экран:
`t=rpart(RealizationOfRecipient ~., data=d)`
`plot(t, compress=T, branch=1, margin=0.1)`
`text(t, use.n=T, pretty=0)`
- График зависимости относительной ошибки прогноза от уровня обрезки дерева `cp`: `plotcp(t)`
- Обрезка для повышения устойчивости прогноза:
`t1=prune(t, cp=0.041)`
- Вывод на экран нового дерева:
`plot(t1, compress=T, branch=1, margin=0.1)`
`text(t1, use.n=T, pretty=0)`
- Перекрёстная таблица, показывающая точность прогноза:
`pred=ifelse(predict(t1)[,1]>0.5, "NP", "PP")`
`table(dative$RealizationOfRecipient, pred)`

Полезный метод выбора модели

Figure 7.9 shows the prediction error and tenfold cross-validation curve estimated from a single training set, from the scenario in the bottom right panel of Figure 7.3. This is a two-class classification problem, using a lin-



Пример из книги
Hastie T., Tibshirani R.,
Friedman J.
«The Elements of
Statistical Learning»
(2009)

ear model with best subsets regression of subset size p . Standard error bars are shown, which are the standard errors of the individual misclassification error rates for each of the ten parts. Both curves have minima at $p = 10$, although the CV curve is rather flat beyond 10. Often a “one-standard error” rule is used with cross-validation, in which we choose the most parsimonious model whose error is no more than one standard error above the error of the best model. Here it looks like a model with about $p = 9$ predictors would be chosen, while the true model uses $p = 10$.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$\varepsilon = \frac{S}{\sqrt{n}}$$


$$(\bar{X} - \varepsilon, \bar{X} + \varepsilon)$$

(лучше использовать
не S , а MAD)



Главное в теме

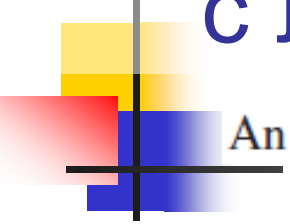
- В классической модели Фишера **дискриминантного анализа** предполагается, что наблюдения в каждом из классов представляют собой реализации многомерных выборок из **многомерных нормальных** законов с разными средними, но **одинаковыми** ковариационными матрицами. Необходимо построить **матричную диаграмму** рассеяния и визуально проверить, выполняется ли хотя бы приблизительно данные условия для всевозможных пар признаков.
- В модели с **разными** ковариационными матрицами в случае размерности больше 5 необходимо иметь много наблюдений для оценки параметров модели (всех элементов ковариационных матриц и векторов средних).
- Важным показателем качества **логистической модели** служит процент верно классифицированных случаев. Рекомендуется случайно отобрать **80%** строк для создания **обучающей выборки** и проверить модель на **20%** оставшихся. Отметим, что логистическая модель является вероятностной, а не алгоритмической, что позволяет проверять гипотезу адекватности модели и значимость отличия коэффициентов модели от 0.
- **Деревья принятия решения** с большим числом ветвей неустойчивы к добавлению (изменению) данных. Для уменьшения неустойчивости применяется метод голосования по бутстрэп-выборкам — *процедура bagging*.

The background of the image shows a vast ocean with several large, jagged icebergs floating on its surface. The sky is filled with heavy, grey clouds, creating a somber and atmospheric setting. The icebergs vary in shape and size, with some having sharp peaks and others being more rounded. The water is a deep blue-grey color, reflecting the light from the sky.

**Всякая вещь
есть форма проявления
беспредельного
разнообразия.**

Козьма Прутков

Связь этимологического возраста слов с лексическими характеристиками



An increase in **valency** (here, the number of different subcategorization frames in which a verb can be used) is closely related to an increase in the verb's number of meanings.

Irregular verbs are generally described as the older verbs of the language. Hence, it could be that they have more meanings and a greater valency because they have had a longer period of time in which they could spawn new meanings and uses. Irregular verbs also tend to be more frequent than regular verbs, and it is reasonable to assume that this high frequency protects irregular verbs through time against regularization.

In order to test these lines of reasoning, we need some measure of the age of a verb. A rough indication of this age is the kind of cognates a Dutch verb has in other Indo-European languages. On the basis of an etymological dictionary, Tabak *et al.* (2005) established whether a verb appears only in **Dutch**, in **Dutch and German**, in **Dutch, German and other West-Germanic languages**, in **any Germanic language**, or in **Indo-European**.

Here, we study whether etymological age itself can be predicted from frequency, regularity, family size, etc.

Домашнее задание

- Для применения модели **порядковой регрессии** установите пакет `rms` (и все связанные с ним пакеты) и подключите его на вкладке `Packages`
- Подключите пакет `languageR` на вкладке `Packages`
- Получите описания признаков из таблицы `etymology`: `?etymology`
- Упорядочите категории (порядковая шкала) этимологического возраста:
`etymology$EtymAge = ordered(etymology$EtymAge,
levels = c("Dutch", "DutchGerman", "WestGermanic",
"Germanic", "IndoEuropean"))`
- Сохраните сводную информацию о признаках из таблицы `etymology` в структуре `etymology.dd`:
`etymology.dd=datadist(etymology); options(datadist="etymology.dd")`
- Постройте модель порядковой регрессии:
`e=lrn(EtymAge ~ WrittenFrequency + NcountStem +
Regularity + LengthInLetters + Denominative + FamilySize +
Valency, data=etymology, x=T, y=T)`
- Выведите отчёты и графики: `e; p=Predict(e); plot(p)`
- Выясните, какой фактический уровень значимости имеет модель порядковой регрессии, какие предикторы значимы на уровне **0,05**