

Национальный исследовательский университет
"Высшая школа экономики"

Факультет компьютерных наук
Департамент анализа данных и искусственного интеллекта

Домашнее задание

по анализу и разработке данных

Выполнили студенты БПМИ133:
Стеценко Макар
Корытова Александра
Милеев Алексей

Москва 2015

Домашнее задание №1

1. В настоящих данных приводится статистика по NEA (Near Earth Objects) и кометам, обнаруженным исследовательской миссией NEOWISE под руководством NASA. Near-Earth Objects - это кометы и астероиды, которые были притянуты гравитационным полем ближайших планет, в следствии чего они смогли сблизиться с Землей.

Каждый объект описывается следующим набором признаков:

- Discovery Date [Дата открытия]
Качественный признак в формате YYYY-MM-DD
- H (mag) [Магнитуда]
Количественный признак, абсолютная величина
С помощью абсолютной магнитуды вычисляется диаметр астероида, чем ниже значение H, тем больше размер объекта.
- MOID (AU) - Minimum Orbit Distance [Минимальная дистанция орбиты]
Количественный признак, измеряемый относительно AU (The astronomical unit).
//AU - астрономическая единица измерения длины, приблизительно показывающая расстояние между Землей и Солнцем. Равна 149597870700 метров (примерно 150 млн км).
Minimum orbit intersection distance (MOID) - мера, используемая в астрономии для оценки потенциальных сближений и рисков столкновений между астрономическими объектами.
- q (AU) perihelion distance
Количественный признак
Perihelion - точка на орбите планеты, кометы или другого объекта, расстояние от которой до Солнца минимально.
- Q (AU) aphelion distance
Количественный признак
Aphelion - точка, в которой небесное тело максимально удалено от Солнца.
- period (yr) [Период]
Количественный признак, показывающий период обращения объекта вокруг Солнца, измеряется в годах.
- PNA (Potentially Hazardous Asteroids)
Признак, показывающий принадлежит ли астероид к классу PNA. Принимает два значения (Y/N), для удобства можно считать количественным.
- Orbit Class [Класс орбиты]
Качественный признак, множество принимаемых значений: {Apollo, Aten, Amor}.

2. Предметная область

Научный интерес к таким объектам проявлен во многом из-за их происхождения. Так, например, астероиды по сути являются уцелевшими осколками после формирования нашей солнечной системы. Поскольку эти объекты могут столкнуться, они оказывали и будут оказывать влияние на биосферу Земли. Так же астероиды являются богатым источником ресурсов. Выяснилось, что минеральных запасов в астероидном поясе Марса и Юпитера хватит, чтобы каждому человеку на Земле дать 100 миллиардов долларов.

По имеющимся данным можно пробовать строить модели для определения принадлежности небесного объекта к классу РНА.

Источник: <http://neo.jpl.nasa.gov/stats/wise/>.

Домашнее задание №2

1. Был выбран количественный признак H (mag) [Магнитуда]. Поскольку этот признак позволяет определить размер исследуемого объекта, то его подробное изучение позволит лучше понять, каких размеров достигают наиболее встречаемые астероиды. В используемом наборе данных H принимает следующие значения:

15.6 16.2 17.0 17.5 18.3 18.3 18.7 18.7 18.9 19.0 19.1 19.2 19.2 19.3 19.3 19.3 19.4 19.4 19.4 19.5 19.5
19.5 19.6 19.6 19.7 19.7 19.7 19.7 19.8 19.9 19.9 19.9 20.0 20.1 20.1 20.2 20.2 20.3 20.3 20.4 20.6 20.7
20.7 20.7 20.7 20.8 20.8 20.8 20.9 20.9 20.9 21.0 21.0 21.1 21.3 21.4 21.4 21.5 21.5 21.6 21.8 21.8 22.0
22.1 22.3 22.5 22.6 22.6 23.2 24.1

Построим гистограмму:

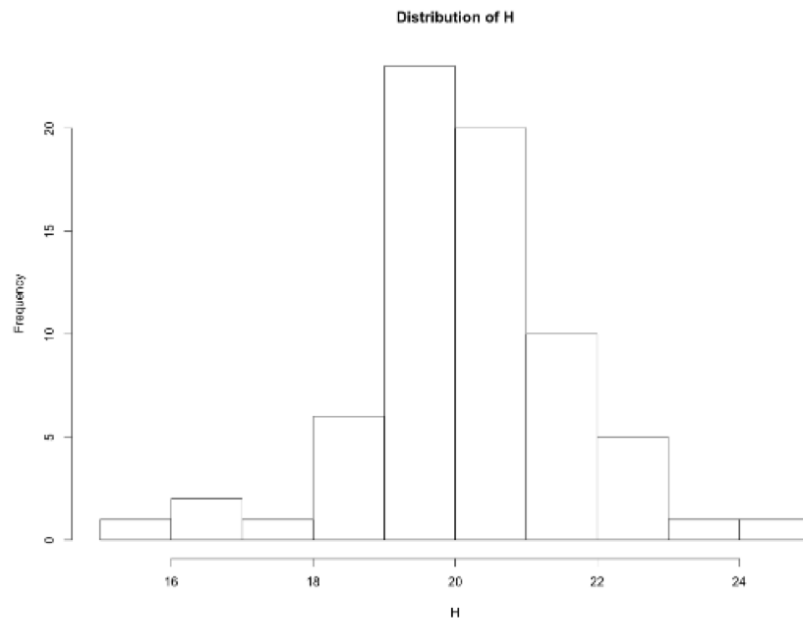


Рис. 1: Гистограмма для признака H (mag)

Полученная гистограмма позволяет нам предположить, что распределение признака H похоже на нормальное. А также понять, в каком диапазоне лежат наиболее встречаемые значения H (примерно от 19 до 21). Этот факт подтвердится, когда мы найдем моду. Построим бокс-плот:

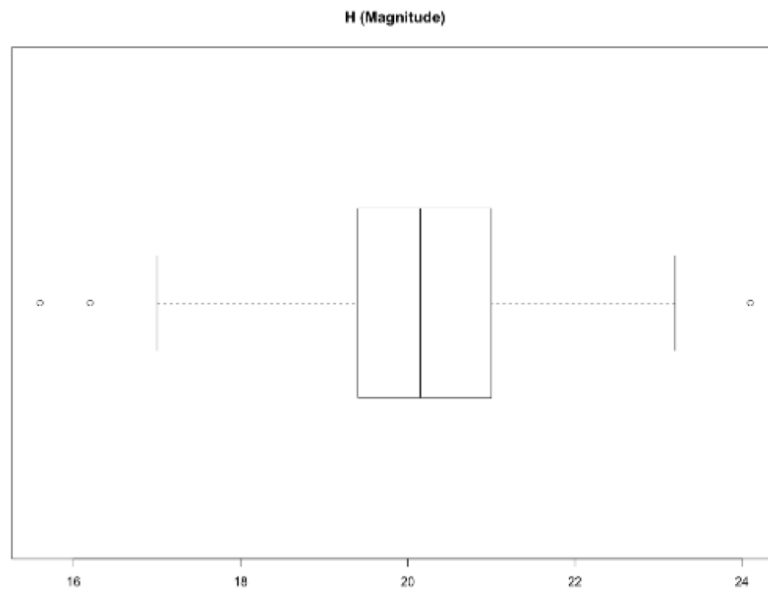


Рис. 2: Бокс-плот для H (mag)

Видно, что у нас есть 3 выброса, а именно: [16.2, 24.1, 15.6]. Так же видно значение медианы. Найдем среднее значение, моду и медиану

Среднее	Медиана	Мода
20.21	20.15	19.7 и 20.7

Как видно, найденные значения не равны, это свидетельствует о том, что величина H не подчиняется нормальному распределению, а немного отклоняется от него. Однако, если убрать из расчетов найденные выбросы [16.2, 24.1, 15.6] и пересчитать, то получим равные между собой значения:

Среднее	Медиана	Мода
20.2	20.2	$(19.7 + 20.7) / 2 = 20.2$

2. Теперь построим доверительные интервалы тремя методами:

1. Статистический
2. Опорный бутстрэп
3. Безопорный бутстрэп

Так как наше распределение похоже на нормальное, то

$$CI = \left(mean - 1.965 \frac{std}{\sqrt{n}}; mean + 1.965 \frac{std}{\sqrt{n}} \right)$$

$$CI = (19.859740; 20.560260)$$

Для 5000 средних значений от случайных выборок с повторениями построим гистограмму:

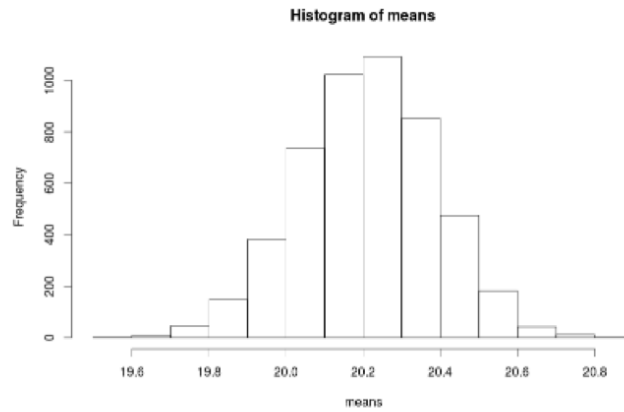


Рис. 3: Гистограмма для признака средних значений

Видим, что гистограмма похожа на нормальное распределение, а значит применяем метод опорного бутстрэпа и получаем интервал:

$$PCI = (20.207460; 20.217272)$$

И теперь безопорный бутстрэп:

$$NPCI = (19.864286; 20.542857)$$

3. Покажем, что для моды и медианы нельзя использовать технику опорного бутстрэпа. Для это построим гистограммы аналогично случаю со средним.

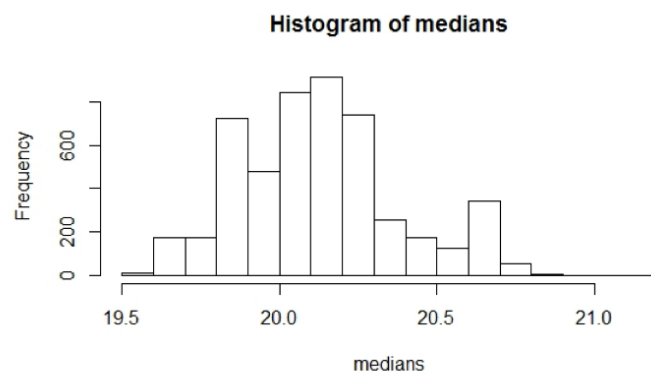


Рис. 4: Гистограмма для медиан

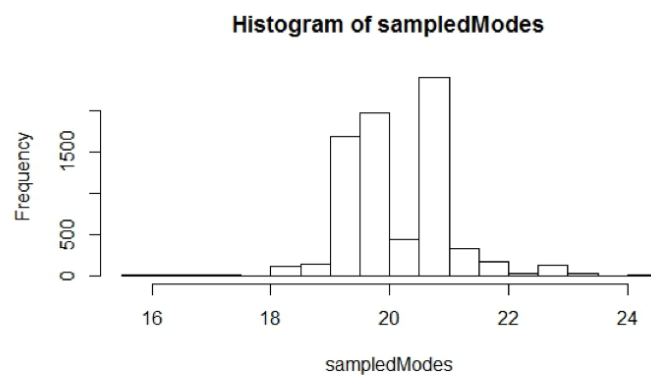


Рис. 5: Гистограмма для мод

Видим, что распределения совсем не напоминают Гауссовские. Значит, мы можем использовать только безопорный бутстрэп.

Доверительный интервал для медианы:

$$NPCI = (19.700000; 20.700000)$$

Доверительный интервал для моды:

$$NPCI = (18.300000; 20.700000)$$

Домашнее задание №3

1. Построим диаграмму разброса для имеющихся признаков:

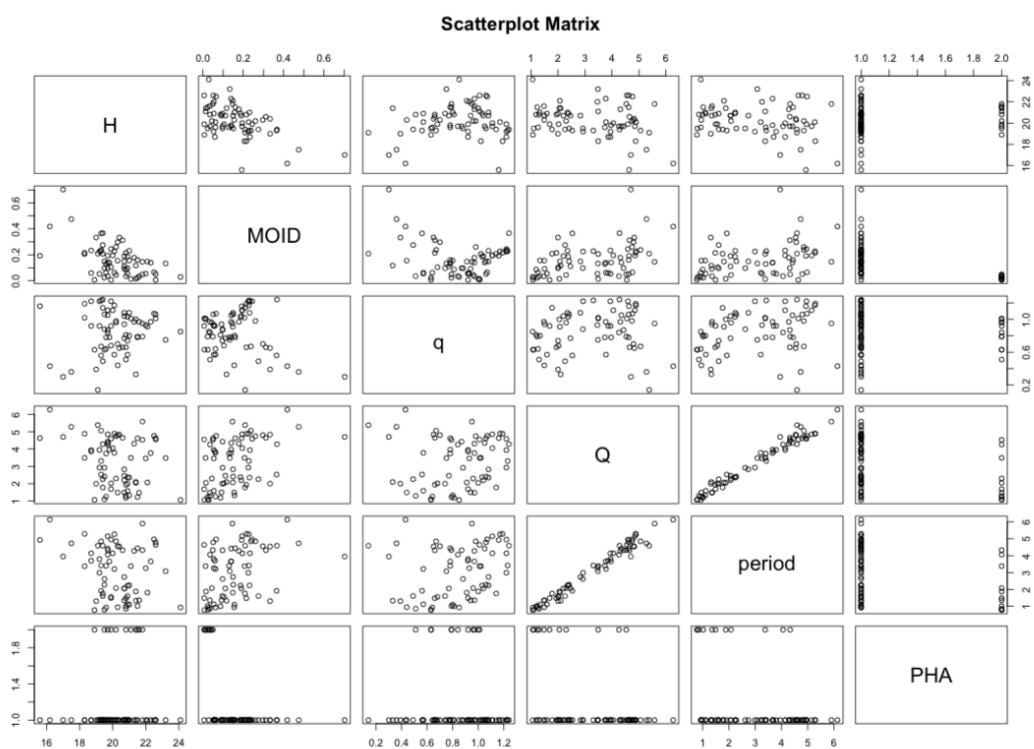


Рис. 6: Матрица разброса

Пара Q и period больше всего напоминает линейную зависимость. Рассмотрим ее отдельно:

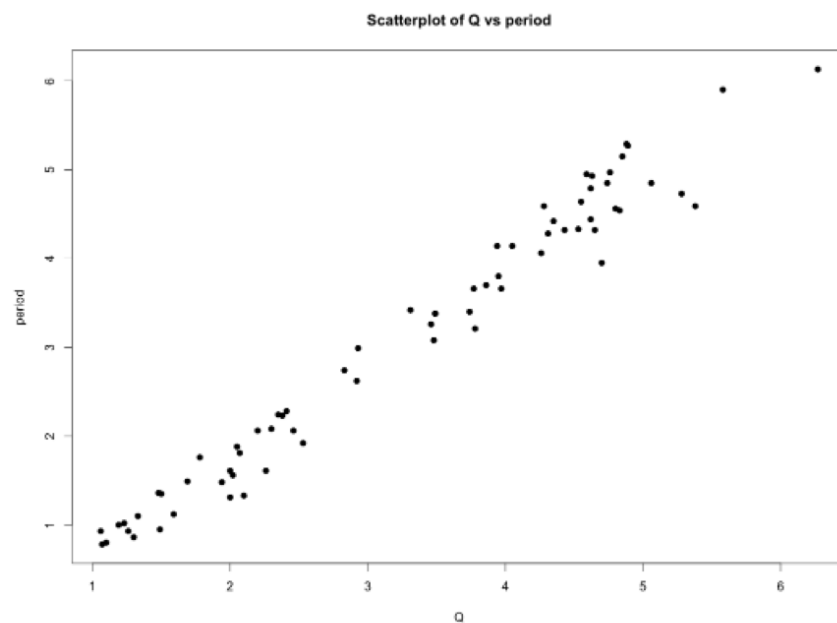


Рис. 7: Диаграмма разброса Q и period

Такую сильную зависимость можно объяснить тем, что Q - это точка, в которой небесное тело максимально удалено от Солнца, а period - это период обращения объекта вокруг Солнца. Чем дальше объект от солнца, тем больше его период. Поскольку, при обнаружении астероида, зная точку, в которой небесное тело максимально удалено от Солнца, можно высчитать период, то Q мы будем считать за X, а период за Y.

2. Теперь найдем коэффициенты линейной регрессии $Y = aX + b$:

a	b
1.075	-0.424

Построим график разброса с нанесенной моделью:

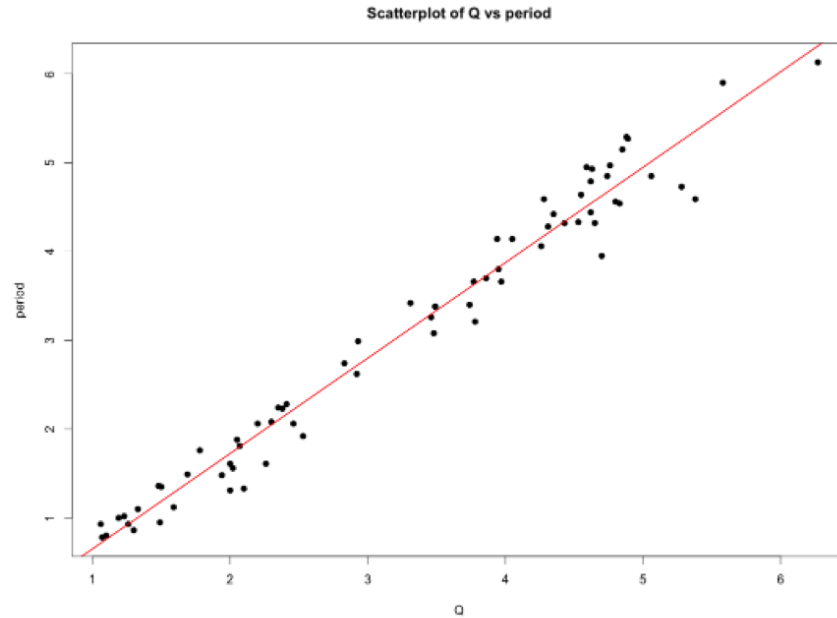


Рис. 8: Диаграмма разброса Q и period и линейная модель

Коэффициент a обозначает на сколько изменится признак `period`, если увеличить Q на малую величину. В данном случае, при увеличении Q на единицу, `period` вырастет на 1.075.

3. Найдем среднюю ошибку предсказания (Mean Absolute Error), она будет равна:

$$MAE = \frac{1}{n} \sum_{k=1}^n |period_k - (aQ_k + b)|$$

MAE
0.2095

Данная величина говорит нам, что в среднем предсказанное моделью значение отличается от имеющихся в выборке данных на 0.21. Чем меньше данная ошибка, тем больше мы можем доверять построенной модели.

4. Коэффициент детерминации равен 0.97. Коэффициент детерминации объясняет долю дисперсии Y регрессией на X . Таким образом, мы получили, что наша модель объясняет 97% дисперсии Y , это очень хороший показатель.

Корреляция признаков `period` и Q равна 0.984. Значение коэффициента корреляции указывает на то, как близко к прямой находятся точки на диаграмме рассеивания, в частности,

значение ± 1 означает точное совпадение, а значение близкое к 0, говорит об отсутствии линейной корреляции. Знак $+$ коэффициента означает, что значение `period` увеличивается с ростом `Q`.

Коэффициент детерминации равен квадрату коэффициента корреляции.

Согласно полученным значениям коэффициентов, для нашей модели мы имеем положительную связь признаков.

Безусловно, гипотеза о существовании линейной связи между признаками `Q` и `period` подтвердилась. Мы получили достаточно близкие к 1 значения коэффициентов корреляции и детерминации, это значит, что построенная регрессия довольно точно отражает реальное положение дел.

Домашнее задание №4

1. Выберем качественный признак `Orbit Class`, который принимает значения `Apollo`, `Aten`, `Amor`, и количественный признак `q` - минимальное расстояние от солнца, на которое может приблизиться объект. Предположим, что эти признаки зависимы. Построим Бокс-плот:

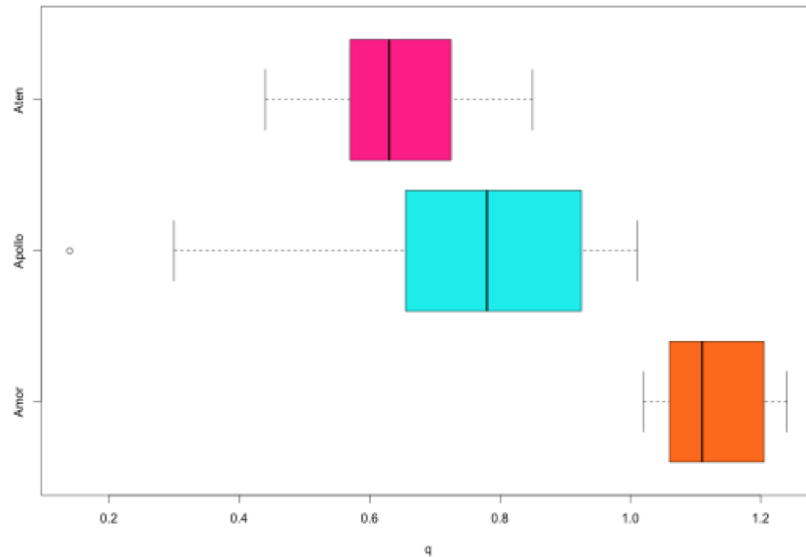


Рис. 9: Бокс-плот MOID для каждого типа объекта

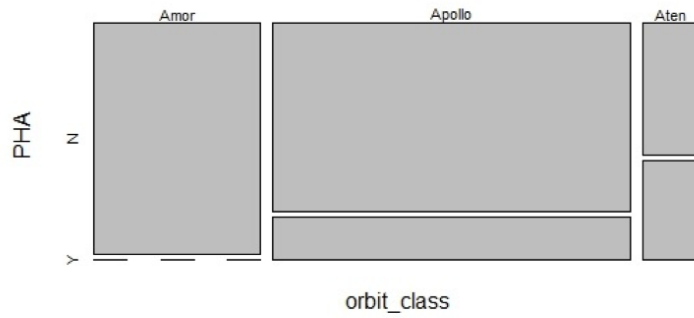
Видно, что орбиты объектов типа `Amor` находятся дальше всех от солнца. У объектов типа `Aten` `q` зачастую меньше, чем у `Apollo`. Данный график придает уверенности в том, что переменные зависимы. Построим регрессионную таблицу и посчитаем корреляционное отношение. Таблица:

y_k	p_k	\bar{x}_k	σ_k
y_1	43	0.7463	0.2210
y_2	7	0.6443	0.1472
y_3	20	1.125	0.0754
	70	0.1611	0.2570

Корреляционное отношение является аналогом коэффициента детерминации. Следовательно, чем ближе оно будет к 1, тем сильнее зависимость между q и Orbit Class. В нашем случае, корреляционное отношение равно **0.4884**, что не свидетельствует о сильной зависимости. Гипотеза не подтвердилась.

2. За пару качественных признаков возьмем Orbit Class из предыдущей части и РНА (признак, показывающий, принадлежит ли астероид к классу потенциально опасных, принимает два значения - Y и N). Выдвинем предположение, что эти признаки зависимы.

Построим для них мозаичную диаграмму:



Построим таблицу сопряженности 2-х признаков в абсолютных значениях:

Orbit Class	N	Y
Amor	20	0
Apollo	35	8
Aten	4	3

И в относительных:

Orbit Class	N	Y
Amor	0.28571429	0
Apollo	0.50000000	0.11428571
Aten	0.05714286	0.04285714

Вычислим теперь матрицу коэффициентов Кетле:

Orbit Class	N	Y
Amor	0.18644068	-1.0000000
Apollo	-0.03429247	0.1839323
Aten	-0.32203390	1.7272727

Максимальное значение коэффициента Кетле у объектов, которые принадлежат к классу Aten и являются потенциально опасными (принадлежат к классу РНА). Он равен **1.727273**.

Максимум в таблице коэффициентов Кетле показывает наличие лучшей связи между категориями.

Значение интегрального коэффициента Кетле **Q = 0.1127674**.

Это свидетельствует о том, что связь присутствует, хотя и очень слабая.

Домашнее задание №5

1. Для этого задания мы будем использовать другой набор данных, а именно реальные данные, предоставленные ФНКЦ им. Рогачева по детям, больным острым лимфобластным лейкозом (ALL), в рамках протокола MB-2008.

Пациенты описываются множеством различных признаков, нам же будет интересен следующий набор: {Age, Leber, Milz, Leuc} в качестве объясняющих признаков и обозначающий соответственно возраст пациента, пальпируемый размер печени (в см), пальпируемый размер селезенки (в см) и число лейкоцитов в крови (на 1 мл крови). Все эти признаки являются количественными.

В качестве целевого признака взят признак исхода (Tod) - качественный признак, задающий 3 класса (жив, мертв, выбыл из наблюдения).

Задача классификации состоит в следующем: имеются обучающее и тестовое множества, которые в нашем случае совпадают. У объектов обучающего множества классы известны, необходимо предсказать классы объектов тестового множества.

2. Были выбраны следующие алгоритмы классификации: метод k ближайших соседей и Наивный Байес.

Метод k ближайших соседей был выбран исходя из соображений того, что все объясняющие признаки у нас количественные, и он отлично подходит для решения этой задачи.

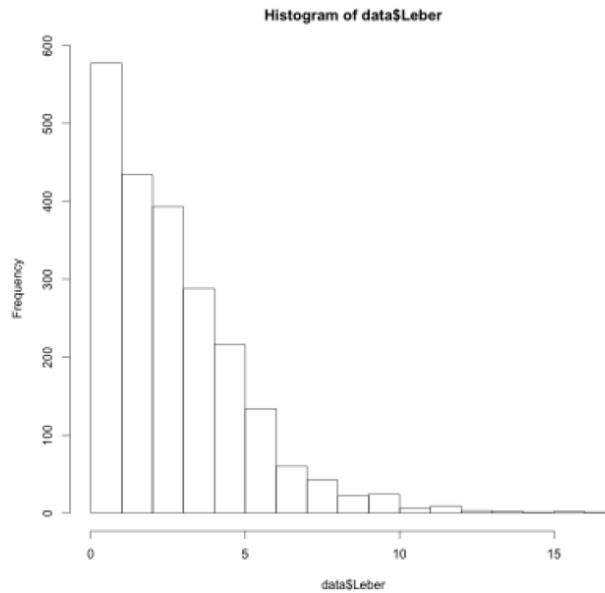
Метод же Наивного Байеса подходит потому, что все признаки независимые.

Для метода ближайших соседей лучше всего подошел параметр $k = 3$.

На выходе мы получили следующее:

target.predict \ target.test	target.test		
	0	1	2
0	1978	17	159
1	0	2	1
2	16	1	41

В предложенном наборе данных все признаки количественные, поэтому воспользуемся шкалированием и переведем их в качественные. Чтобы понять на какие промежутки разбить каждый признак, построим гистограмму. Например, для признака Leber получим:



```

buckets = hist(data$Leber)$breaks
for(i in 2:length(buckets)) {
  if (nrow(data[which(Leber > buckets[i-1] & Leber <= buckets[i]),]) > 0) {
    data[which(Leber > buckets[i-1] & Leber <= buckets[i]),]$Leber_q = i
  }
}

```

Аналогично для остальных признаков.

Результаты, которые выдал метод Наивного Байеса:

target.predict \ target.test	target.test		
	0	1	2
0	1992	18	195
1	0	2	0
2	2	0	5

3. Посчитаем по полученным матрицам ошибок точность, полноту и F-меру классификаторов:

Для метода ближайших соседей:

$0.9182915506 = P_1, 0.6666666667 = P_2, 0.701754386 = P_3$

$0.9919759278 = R_1, 0.1 = R_2, 0.2 = R_3$

$P = 0.7622375344$

$$R = 0.4306586426$$

$$F = 0.5503650498$$

Для метода Наивного Байеса:

$$0.9034013605 = P_1, 1 = P_2, 0.7142857143 = P_3$$

$$0.998996991 = R_1, 0.1 = R_2, 0.025 = R_3$$

$$P = 0.8725623583$$

$$R = 0.3746656637$$

$$F = 0.5242331784$$

В целом, можно сказать, что метод Наивного Байеса справился лучше (параметр точности выше), но по параметру полноты метод ближайших соседей выигрывает.

4. Мы будем использовать метод кросс-валидации leave-one-out.

Суть данного метода состоит в том, что в качестве тестового множества будет выступать каждый отдельный объект по очереди, а в качестве обучающего, соответственно, все остальные.

Получили следующие результаты:

Метод ближайших соседей:

target.predict \ target.test	0	1	2
0	1947	20	195
1	1	0	1
2	46	0	4

$$0.9005550416 = P_1, 0 = P_2, 0.08 = P_3$$

$$0.9764292879 = R_1, 0 = R_2, 0.02 = R_3$$

$$P = 0.3268516805$$

$$R = 0.332143096$$

$$F = 0.3294761445$$

Метод Наивного Байеса:

target.predict \ target.test	0	1	2
0	1990	14	187
1	0	3	0
2	4	3	13

$$P_1 = 0.90826106, P_2 = 1, P_3 = 0.65$$

$$R_1 = 0.9979939819, R_2 = 0.15, R_3 = 0.065$$

$$P = 0.8527536893$$

$$R = 0.4043313273$$

$$F = 0.5485627885$$

Как видим, результаты ухудшились. В случае метода ближайших соседей очень сильно.

Это связано с тем, что мощность обучающего множества становится меньше, итераций соответственно больше, и каждый раз перестраивается модель.

По-прежнему, метод Наивного Байеса справляется лучше.

Домашнее задание №6

1. Это домашнее задание будем делать на новом наборе данных, взятых с сайта archive.ics.uci.edu.

Датасет содержит информацию о семенах, принадлежащих 3 различным сортам пшеницы, по 70 штук каждого, взятых случайным образом для некоего эксперимента. Производились рентгеновские снимки семян, после чего выполнялись измерения различных параметров.

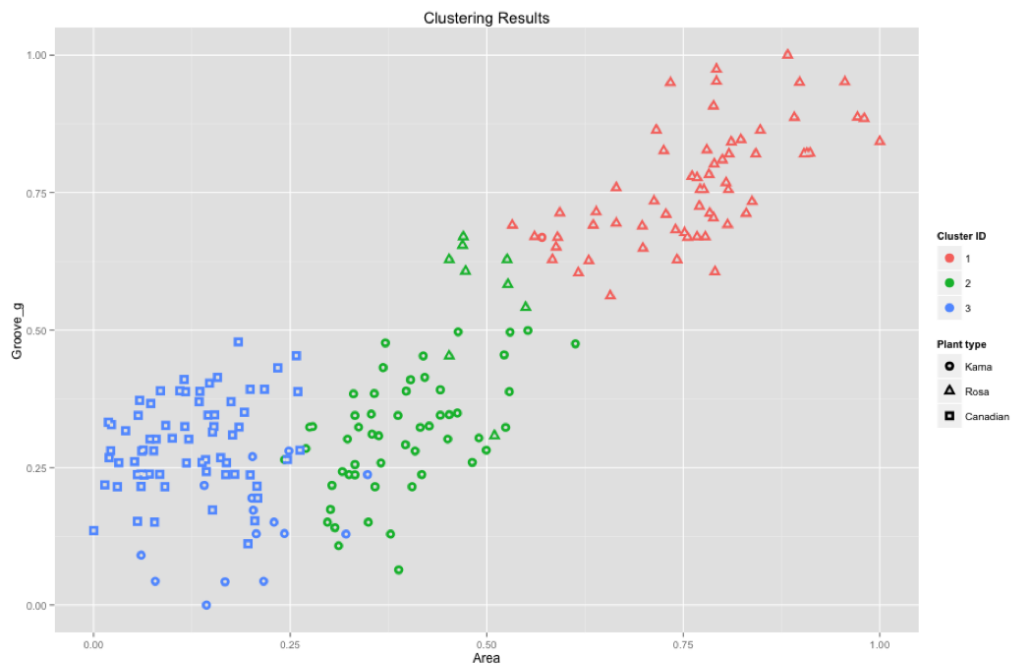
Объекты описываются следующим набором признаков:

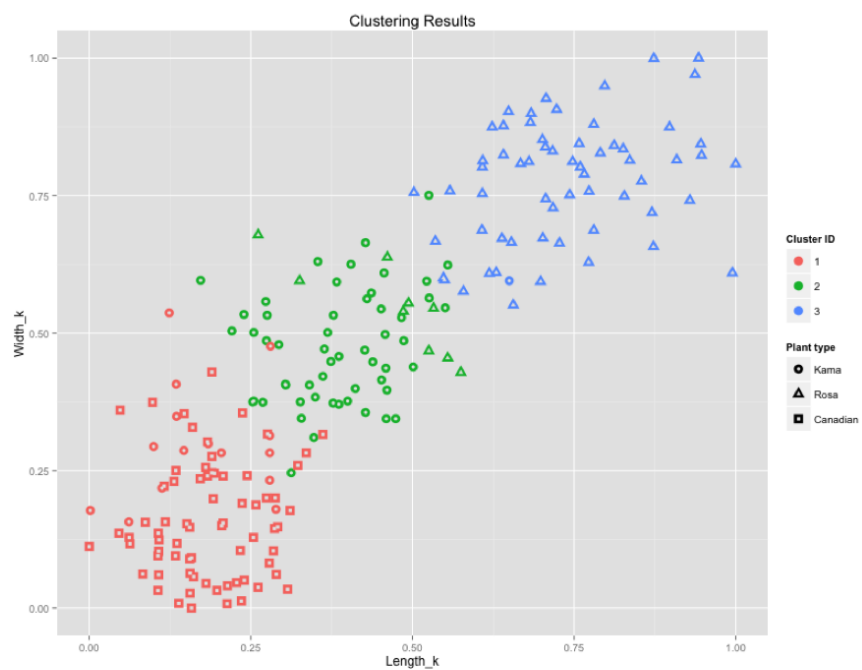
1. Площадь (A)
2. Периметр (P)
3. Компактность $C = \frac{4\pi A}{P^2}$
4. Длина ядра
5. Ширина ядра
6. Коэффициент асимметрии
7. Размер выемки

Суть кластеризации на этих данных - разделить семена, не зная их тип, на группы близких друг другу.

Очевидно, что наиболее осмысленно делить на 3 кластера.

2. Вот пример картинки, полученной на выходе алгоритма K-means.



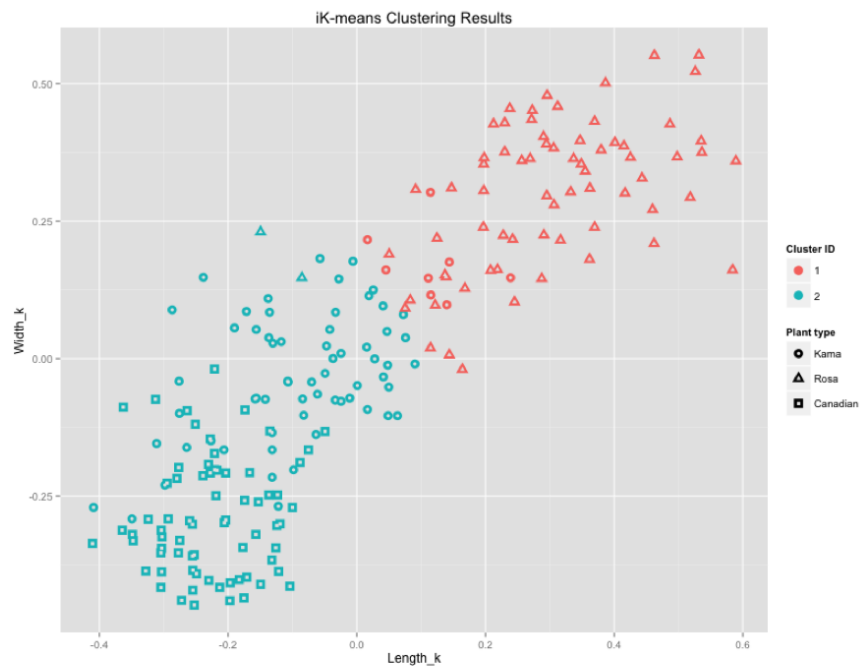


Как видно, деление на кластеры практически совпадает с делением на сорта пшеницы, к которым изначально принадлежали семена.

3. Алгоритм iK-means сделал вывод, что оптимальное число кластеров $K^* = 2$.

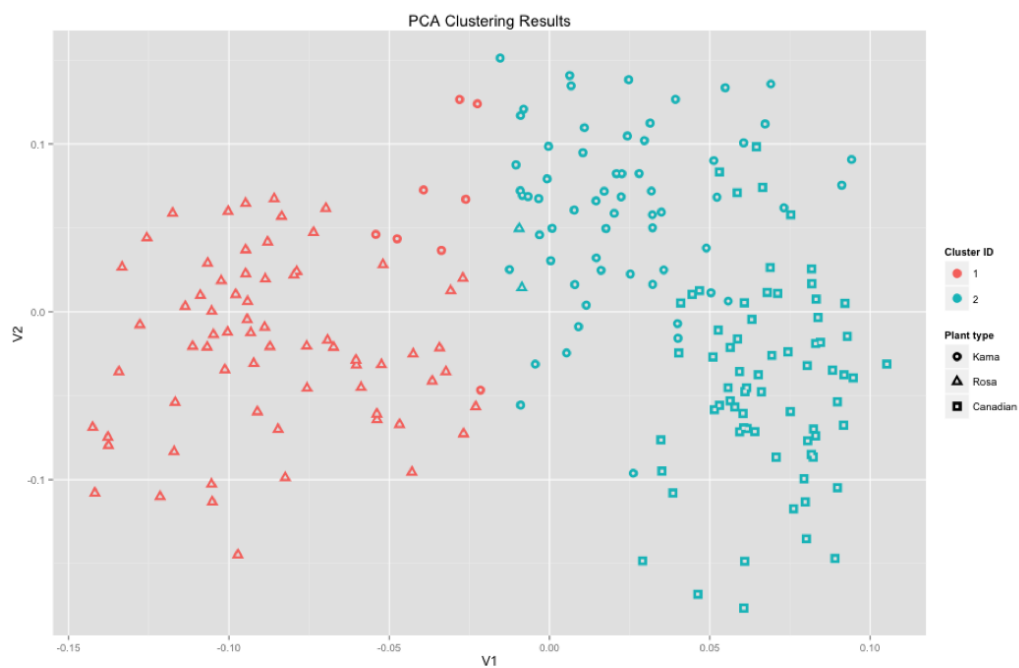
Рисунок:





Как можно заметить, семена, принадлежащие типа Kama и Canadian в большинстве своем попали в один кластер. Значит, они имеют довольно схожие параметры, в отличие от типа Rosa, объекты которого образуют отдельный кластер.

4. После реализации алгоритма SVD, мы получили такое деление на кластеры в пространстве главных компонент:



Как можно заметить, "составляющие" кластеров практически не изменились. Объекты, принадлежащие типу Rosa, входят по-прежнему в отдельный кластер. Единично туда попали объекты другого типа.