

Машинное обучение

Вводный курс

KaltovichArtyom@gmail.com

Июнь 2018

Оглавление

Предисловие	2
Введение	3
Глава 1 Визуализация и кластеризация	8
1.1 Визуализация	8
1.1.1 Определение	8
1.1.2 Пример	8
1.1.3 Почему мы визуализируем?	9
1.1.4 Попробуем сами	10
1.1.5 Примеры кода	11
1.2 Кластеризация	11
1.2.1 Определение	11
1.3 Визуализация	11
1.3.1 Набор ирисов Фишера	11
1.3.2 Примеры кода	11
1.4 Кластеризация	11
1.4.1 Методы	11
1.5 Машинное обучение. Определение	12
Список литературы	12

Предисловие

Судя по тому, что вы решили почитать эту книгу, вам приходилось слышать о машинном обучении, пожалуй, сейчас сложнее не услышать это словосочетание, например, термин "Machine Learning" содержится в [списке buzzword'ов английской википедии](#). Модные слова (также гламурная лексика и «умные слова», англ. buzzword) — особый род новых слов и речевых конструкций, часто используемых в коммерции, пропаганде и профессиональной деятельности для оказания впечатления осведомлённости говорящего и для придания чему-либо образа важности, уникальности или новизны. Из-за неумеренного употребления смысл слова размывается, и «модные слова» можно встретить даже в контексте, не имеющем отношения к исходному значению: например, «элитные семинары [слово, в какой-то момент заменившее «элитарные» — доступные самим богатым] по умеренным ценам», «эксклюзивные часы [изготовленные штучно], выпущенные тиражом в 11111 экземпляров». [10]

Машинное обучение попало в этот список не просто так, действительно часто в СМИ появляются новости, что ещё один рубеж покорился программе, иногда можно услышать апокалиптические сценарии о скором восстании машин или наоборот, идеалистические, о скором приходе технологической сингулярности и возможности перенесения сознания в машину или воссоздания сознания умерших людей. Мы же будем в стороне от всего этого ажиотажа и подойдём непосредственно к изучению основ машинного обучения. Что это? Зачем нужно? Как и где используется? Главная же цель — помочь вам изучить основы машинных алгоритмов.

Введение



Рис. 1: ReCAPTCHA первой версии

Вы наверняка помните то время, когда интернет сайты постоянно заставляли вас разбирать, что написано на данных картинках. Сайты требовали это, чтобы избежать злоупотребления функционалом сайта со стороны ботов, предупреждая таким образом регистрацию множества "пользователей" отправку кучи сообщений или злоупотребления иным функционалом сайта или попытки перебора паролей.

Однако несколько лет назад вдруг подобные надписи пропали. Почему? Правоохранительные органы пересажали всех злоумышленников? Или они сами решили уйти в монастырь замаливать грехи. Вряд ли. Произошло то, что всегда происходит в таких случаях: всегда найдётся кто-то, кто захочет злоупотребить вашей системой во вред вам или во вред пользователям, ради кражи их данных, создания повышенной нагрузки на сервера и соответственно замедление или вообще прекращения их работы и других целей. Подобные люди никуда не делись и не денутся.

С появлением сайтов появились и скрипты, которые имитировали действия реальных пользователей, не являясь ими, создателям сайтов это не понравилось, поэтому они попытались отличить ботов от реальных пользователей. Первоначальной идеей было создать задание, с которым бы легкоправлялись люди, но плохо справлялись роботы, такой задачей стала задача распознавания символов. Действительно для вас не составляет никакого труда разобрать, что значат все эти закорючки, линии и кружочки из которых состоят слова

в этой книге (по крайней мере, я очень на это надеюсь). Но до недавних пор эта задача была для компьютеров совсем не тривиальной.

Так и появилось то, что называется капчей. Капча (от CAPTCHA — англ. Completely Automated Public Turing test to tell Computers and Humans Apart — полностью автоматизированный публичный тест Тьюринга для различия компьютеров и людей) — компьютерный тест, используемый для того, чтобы определить, кем является пользователь системы: человеком или компьютером. Термин появился в 2000 году. Основная идея теста: предложить пользователю такую задачу, которая с лёгкостью решается человеком, но крайне сложна и трудоёмка для компьютера. По состоянию на 2013 год, каждый день пользователями по всей планете вводится примерно 320 миллионов «капчей». [7]

Но разумеется злоумышленники (и не обязательно злоумышленники, но и многие исследователи из научного интереса) не сдались и решили попытаться обойти защиту. Мы рассмотрим хронологию обхода ReCAPTCHA - пожалуй самую известную реализацию, изображение которой и было представлено выше, на рис. 1.

- 1 августа 2010 Chad Houck добился точности "взлома" капчи в 31.8%.
- 26 мая 2012 Adam, C-P and Jeffball - 99.1%. Следует уточнить, что ими была взломана аудио версия, при этом за несколько часов до презентации Google выпустил обновление точность упала до 60.95%.
- 27 июня 2012 - мексиканские студенты Claudia Cruz, Fernando Uceda, and Leobardo Reyes добились точности 82%.
- Август 2017 - исследователи из университета Мериленда: Kevin Bock, Daven Patel, George Hughey, Dave Levin - 85.15%.

Исследователи утверждали, что Гугл постоянно менял свою капчу, при этом часто возвращаясь к предыдущей версии[5]. Однако, всё же сложность капчи постепенно росла в конце концов требование, с которым она создавалась, — быть сложной в решении для программ и лёгкой для людей, стало выполняться всё хуже. Так в

августе 2012 — более 90% пользователей находят капчу сложной для ввода[5]. В мае 2016 была прекращена поддержка данного типа капчи, то есть она перестала обновляться, в 31 марта 2018 выключена окончательно.[1]

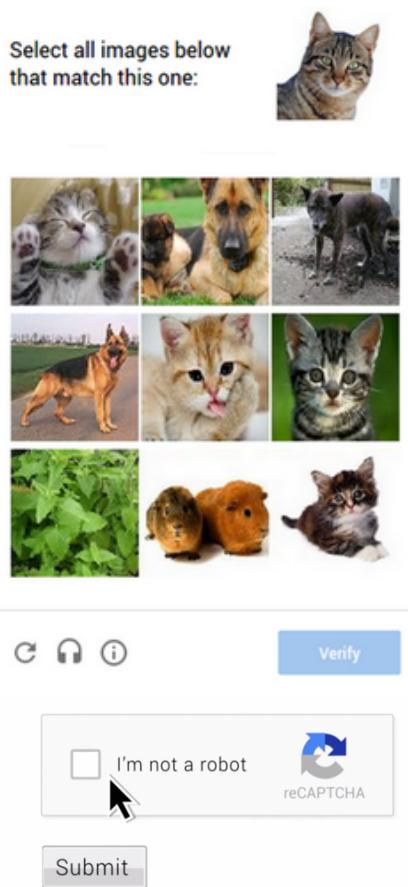


Рис. 2: ReCAPTCHA второй версии

После этого гугл поменял тип своей капчи: теперь необходимо выбрать картинки с котиками, ибо все любят котиков. Таким образом гугл убивает сразу двух зайцев: отличает людей от ботов и создаёт обучающую выборку для своих алгоритмов распознавания текста и картинок. Перед этим была промежуточная версия, в которой выводилась фотография номера дома, соответственно пользователю нужно было разобрать этот номер, во славу google.maps.

Сейчас гугл развивает систему, которая анализирует действия пользователей: движения мыши, клики, нажатия клавиш. Люди не могут похвастать монотонностью действий, если боту нет смысла вести курсор прерывисто или не по прямой, нажимать клавиши с разной скоростью, а не середине ввода вообще отвлечься на звонок или сходить за чаем (в принципе, можно вообще не вводить мышью или эмулировать физические нажатия клавиш, а использовать JS для совершения действий, но это уже совсем другая история). Разумеется создатели ботов тут же стали изменять свои алгоритмы, чтобы быть более человекоподобными (и автору самому приходилось этим заниматься), разумеется гугл тоже не стоит на месте, с каждым разом обходить защиту становится всё сложнее. Следует сказать, что эта капча, пожалуй, является самой дружелюбной

к обычным пользователям, а мы ведь не хотим отпугивать пользователей необходимостью постоянно вводить свою электронную почту, sms коды подтверждения, и выбирать на каких картинках изображено что либо (кроме котиков, разумеется). Тут же человек сразу начинает пользоваться сайтом, скрипты в фоне анализируют его действия и их характер и в случае подозрения выводят уже обычную капчу или блокируют активность.

Как мы видим, алгоритмы машинного обучения заставили гугл постоянно менять свою систему, чтобы добиться необходимой эффективности защиты, во многом добиваясь этого использованием тех же алгоритмов: для того, чтобы распознавать, на какой именно части панорамы улицы находится номер улицы, и, следовательно, какую именно, часть следует отослать пользователям на проверку, и вывести метрики для отличия ботов от людей.

Рассмотрим ещё несколько сфер, в которых сейчас широко используется машинное обучение.

Шахматы

Пожалуй первой из популярных интеллектуальных игр, которые покорились машинному интеллекту, были шахматы.

- В феврале 1996 года Гарри Каспаров победил шахматный суперкомпьютер Deep Blue со счетом 4-2. Этот матч выдающийся тем, что первую партию выиграл Deep Blue, автоматически став первым компьютером, победившим чемпиона мира по шахматам в турнирных условиях.
- В мае 1997 года Deep Blue II выигрывает матч у Гарри Каспарова со счётом $3\frac{1}{2} : 2\frac{1}{2}$.
- В 2000 году коммерческие шахматные программы Junior и Fritz смогли свести в ничью матчи против предыдущих мировых чемпионов Гарри Каспарова и Владимира Крамника.
- В ноябре-декабре 2006 года чемпион мира Владимир Крамник играл с программой DeepFritz. Матч закончился выигрышем ма-

шины со счётом 2-4.

Сейчас же даже у лучших гроссмейстеров нет шансов против среднего компьютера[4]. При этом программы продолжают играть против друг друга и совершенствоваться в уровне игры и производительности алгоритмов.

Го

Шахматы пали перед алгоритмами и вычислительной мощью компьютеров, но многие годы считалось, что уж в других сферах компьютеры ещё долго не смогут составить конкуренцию человеку. Выражались мысли, что поле для игры слишком большое, что компьютер не сможет точно оценить текущую позицию, все возможные варианты её развития, выбрать оптимальный, у него просто не хватит на это "мозгов"[3][2].

Конечно, доля правды в этих оценках была, притом немалая, однако, насчёт прогнозов, когда компьютеры смогут посоревноваться на равных с человеком, и причины, которые ему пока мешают были в корне неверными. В Го существует огромное количество возможных вариантов развития игры и позиций, для того, чтобы перебрать их все компьютеру потребуется время больше всей жизни вселенной, но ведь человек тоже не перебирает все возможные варианты, не делает этого и программа. Человек тоже весьма ограничен в своих способностях к оценке ситуации и вариантов её развития, не перебирает все возможности, выбирая лучшую, не делает этого и компьютер, для того, чтобы обойти человека не нужно играть идеально, нужно играть лишь лучше человека. На перебор всех вариантов шахмат тоже уйдёт время всей жизни вселенной, компьютер не перебирает все варианты в шахматах, не делает этого и в Го. Пожалуй, самой главной проблемой, препятствующей созданию эффективных алгоритмов, — являлось отсутствие хороших математических моделей, описывающих игру, с развитием и созданием этих моделей, появились и успехи у игровых программ:

- В октябре 2015 года программа AlphaGo, разработанная компанией DeepMind выиграла у трехкратного чемпиона Европы Фань

Хуэя (2 профессиональный дан) матч из пяти партий со счётом 4—1. Это первый в истории случай, когда компьютер выиграл у профессионала в равной игре.

- В марте 2016 года AlphaGo победила профессионала 9 дана Ли Седола в четырёх партиях из пяти.
- В мае 2017 года на саммите «Future of Go Summit» AlphaGo выиграла три партии из трёх в мини-матче с одним из сильнейших игроков в мире, лидером мирового рейтинга Эло Кэ Цзе.

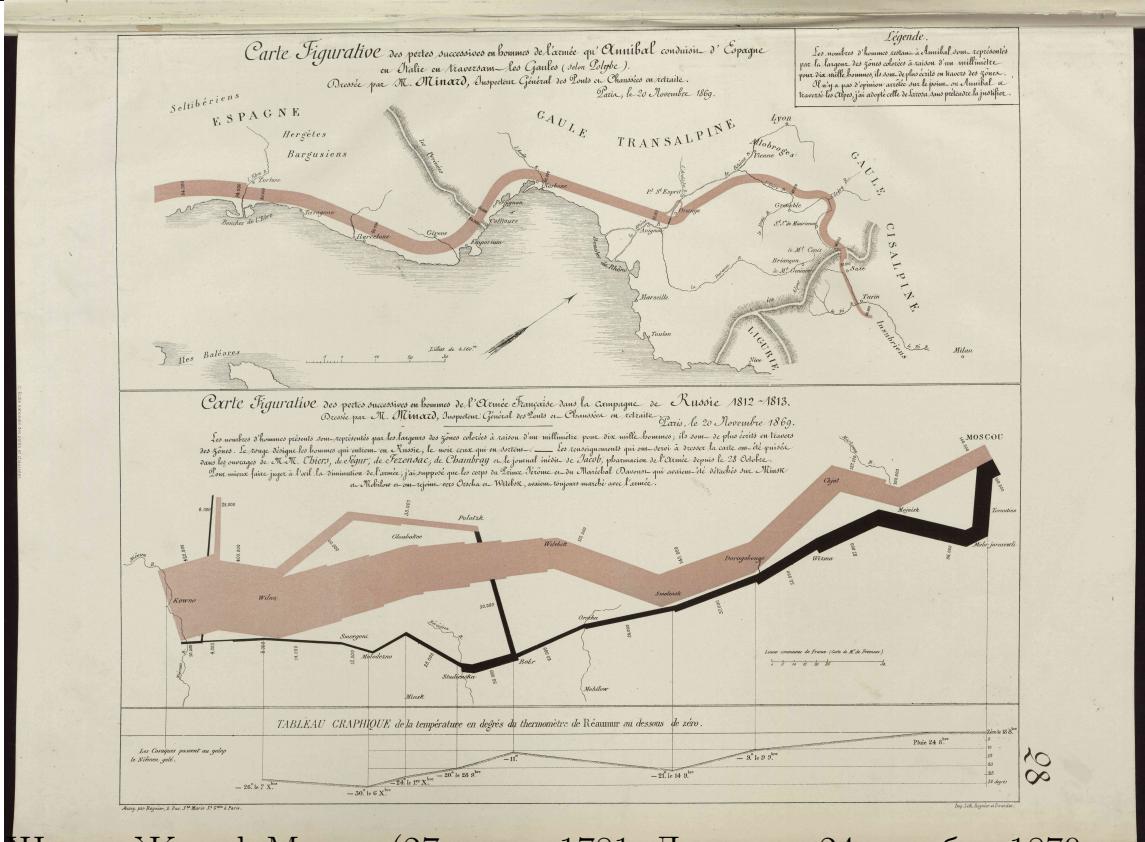
1 | Визуализация и кластеризация

1.1 | Визуализация

1.1.1 | Определение

Визуализация (от лат. *visualis*, «зрительный», англ. *Visualization*) — общее название приёмов представления числовой информации или физического явления в виде, удобном для зрительного наблюдения и анализа[6].

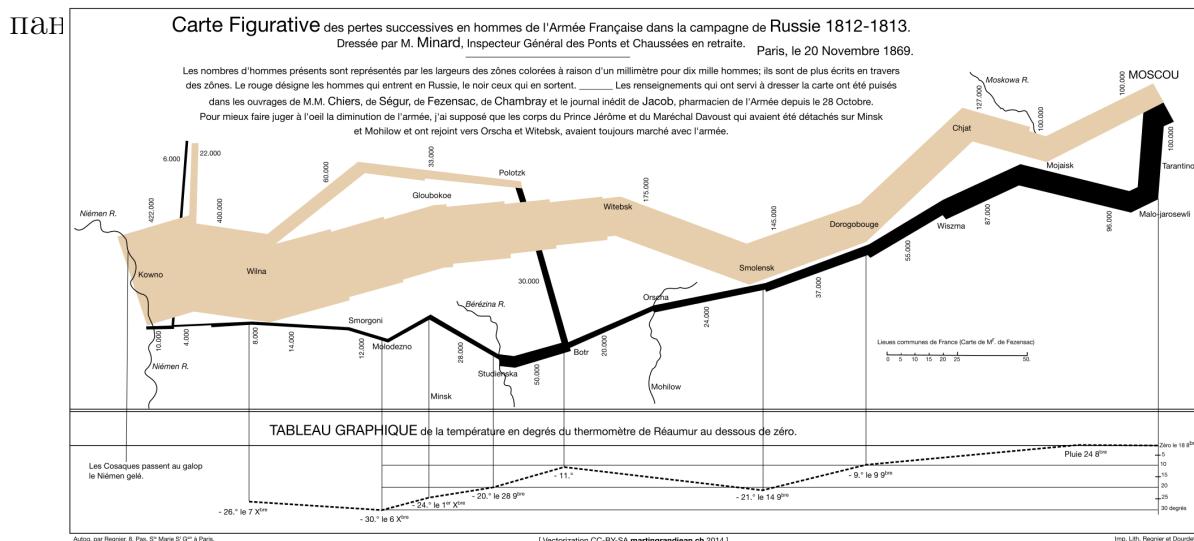
1.1.2 | Пример



ГЛАВА 1. ВИЗУАЛИЗАЦИЯ И КЛАСТЕРИЗАЦИЯ

10

Наполеона Бонапарта в Россию в 1812 году. 20 ноября 1869 года — карта, показывающая перемещение войск Ганнибала из Иберии (Испания)



1.1.3

Почему мы визуализируем?

File Home Insert Page Layout Formulas Data Review View WIP Design Tell me what you want to do

Insert Toggle Actions Pane Reset WIP Report Data Report Completed Contracts Import Export Validate Help Add-In Activation About WIP Altova on the Web WIP Add-In

WIP Report XBR L Help WIP Report Properties

A4		200																
		E	F	G	H	I	J	K	L	M	N	O	P	Q				
1			From Inception to June 13, 2016												At June 13, 2016		For the Period Ended June 13, 2016	
2	Total Contract																	
3	Estimated Revenue	Estimated Costs	Estimated Gross Profit	Earned Contract Revenue	Contract Costs	Gross Profit	Contract Billings	Estimated Costs to Complete	Percent Complete	Under (Over) Billings	Earned Contract Revenue	Contract Costs	Gross Profit (Loss)					
4	29,831,262	22,771,956	7,059,305	12,113,470	9,246,924	2,866,546	11,987,630	13,525,032	41%	125,840	3,740,588	2,855,269	885,319					
5	4,765,875	3,914,859	850,016	4,761,592	3,912,340	849,252	4,748,777	3,619	100%	12,815	319,669	185,925	133,738					
6	3,163,949	2,631,676	530,273	3,073,180	2,558,445	514,755	3,092,332	77,231	97%	(19,152)	1,212,380	1,019,868	192,512					
7	6,845,696	5,348,200	1,497,496	5,935,890	4,657,414	1,298,476	5,727,306	710,786	87%	208,584	2,985,189	2,544,782	640,407					
8	3,202,917	2,139,767	1,063,150	3,197,169	2,136,328	3,061,441	3,199,414	3,439	100%	(1,645)	386,839	241,974	144,865					
9	3,267,627	2,402,206	865,421	3,122,086	2,295,211	826,873	3,143,402	106,995	96%	(21,316)	254,751	101,060	153,691					
10	3,513,815	2,260,925	1,252,880	2,839,159	1,827,211	1,012,548	2,573,819	433,714	81%	265,940	1,823,265	1,173,159	650,106					
11	3,913,079	3,104,573	808,506	3,591,755	2,849,640	742,113	3,503,374	254,933	92%	88,381	2,651,445	2,039,028	612,417					
12	12,187,491	13,500,000	(1,312,309)	2,193,165	3,505,674	(1,312,309)	2,476,537	9,994,326	26%	(283,372)	2,193,165	3,505,674	(1,312,309)					
13	3,274,077	2,198,537	475,720	35,79	30,380	5,199	0	2,767,777	1%	35,779	30,380	5,199						
14	3,835,139	4,296,527	(461,388)	2,578,713	3,040,101	(461,388)	2,386,461	1,256,426	71%	192,252	2,578,713	3,040,101	(461,388)					
15	13,500,000	10,227,273	3,272,727	8,553,041	6,479,577	2,073,464	8,321,142	3,747,696	63%	231,899	8,553,041	6,479,577	2,073,464					
16	3,849,262	3,137,190	712,072	274,615	223,814	50,801	1,741,936	2,913,376	7%	(1,467,321)	274,615	223,814	50,801					
17	74,614,943	64,402,779	10,212,164	46,921,464	41,803,708	5,117,756	43,715,328	22,599,071		29,854,173	27,271,295	2,582,878						
18	169,767,132	142,941,288	26,825,844	99,192,278	84,546,967	14,645,311	96,617,458	58,394,321		(631,316)	56,863,606	50,512,106	6,351,500					
19																		
20																		
21																		
22																		
23																		
24																		
25																		
26																		
27																		
28																		
29																		

WIP Report Properties

Data: Accuracy, Currency

Document Information: Period Start Date, Period End Date, Fiscal Year Focus, **Fiscal Period Focus**

Entity Information: Registrant Name, Current Fiscal Year End Date, Tax Identification Number, Data Universal Numbering System (DUNS), State Registration Number

Fiscal Period Focus: Specifies the fiscal period attributed with

Cell Documentation: Select a table data cell to display a short doc

Balance table: unw								16:28 Monday, January 6, 2014 2			
Obs	row_name	tx_mn	tx_sd	ct_mn	ct_sd	std_eff_sz	stat	p	ks	ks_pval	table_name
1	unw.age	25.816	7.155	28.03	10.787	-0.309	-2.994	0.003	0.158	0.003	unw
2	unw.educ	10.346	2.011	10.235	2.855	0.055	0.547	0.584	0.111	0.074	unw
3	unw.black	0.843	0.365	0.203	0.403	1.757	19.371	0	0.64	0	unw
4	unw.hispan	0.059	0.237	0.142	0.35	-0.349	-3.413	0.001	0.083	0.317	unw
5	unw.nodegree	0.708	0.456	0.597	0.491	0.244	2.716	0.007	0.111	0.074	unw
6	unw.married	0.189	0.393	0.513	0.5	-0.824	-8.607	0	0.324	0	unw
7	unw.re74	2095.574	4886.62	5619.237	6788.751	-0.721	-7.254	0	0.447	0	unw
8	unw.re75	1532.055	3219.251	2466.484	3291.996	-0.29	-3.282	0.001	0.288	0	unw
Balance table: ks.max.ATT								16:44 Monday, January 6, 2014 3			
Obs	row_name	tx_mn	tx_sd	ct_mn	ct_sd	std_eff_sz	stat	p	ks	ks_pval	table_name
17	ks.max.ATT.age	25.816	7.155	25.764	7.408	0.007	0.055	0.956	0.107	0.919	ks.max.ATT
18	ks.max.ATT.educ	10.346	2.011	10.572	2.14	-0.113	-0.712	0.477	0.107	0.919	ks.max.ATT
19	ks.max.ATT.black	0.843	0.365	0.835	0.371	0.022	0.187	0.852	0.008	1	ks.max.ATT
20	ks.max.ATT.hispan	0.059	0.237	0.043	0.203	0.069	0.779	0.436	0.016	1	ks.max.ATT
21	ks.max.ATT.nodegree	0.708	0.456	0.601	0.49	0.235	1.1	0.272	0.107	0.919	ks.max.ATT
22	ks.max.ATT.married	0.189	0.393	0.199	0.4	-0.024	-0.169	0.866	0.01	1	ks.max.ATT
23	ks.max.ATT.re74	2095.574	4886.62	1673.666	3944.6	0.086	0.8	0.424	0.054	1	ks.max.ATT
24	ks.max.ATT.re75	1532.055	3219.251	1257.242	2674.922	0.085	0.722	0.471	0.094	0.971	ks.max.ATT
Balance table: es.mean.ATT								16:44 Monday, January 6, 2014 4			
Obs	row_name	tx_mn	tx_sd	ct_mn	ct_sd	std_eff_sz	stat	p	ks	ks_pval	table_name
9	es.mean.ATT.age	25.816	7.155	25.802	7.279	0.002	0.015	0.988	0.122	0.892	es.mean.ATT
10	es.mean.ATT.educ	10.346	2.011	10.573	2.089	-0.113	-0.706	0.48	0.099	0.977	es.mean.ATT
11	es.mean.ATT.black	0.843	0.365	0.842	0.365	0.003	0.027	0.978	0.001	1	es.mean.ATT
12	es.mean.ATT.hispan	0.059	0.237	0.042	0.202	0.072	0.804	0.421	0.017	1	es.mean.ATT
13	es.mean.ATT.nodegree	0.708	0.456	0.609	0.489	0.218	0.967	0.334	0.099	0.977	es.mean.ATT
14	es.mean.ATT.married	0.189	0.393	0.189	0.392	0.002	0.012	0.99	0.001	1	es.mean.ATT
15	es.mean.ATT.re74	2095.574	4886.62	1556.93	3801.566	0.11	1.027	0.305	0.066	1	es.mean.ATT
16	es.mean.ATT.re75	1532.055	3219.251	1211.575	2647.615	0.1	0.833	0.405	0.103	0.969	es.mean.ATT

1.1.4 | Попробуем сами
 Рональд Фишер в 1936 году продемонстрировал работу разработанного им метода анализа. Данные были собраны американским ботаником Эдгаром Андерсоном.

Признаки:

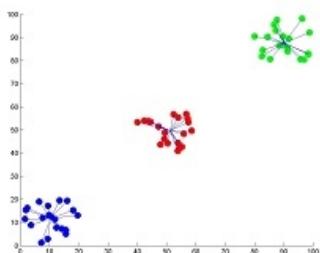
- Длина наружной доли околоцветника (англ. sepal length);
- Ширина наружной доли околоцветника (англ. sepal width);
- Длина внутренней доли околоцветника (англ. petal length);
- Ширина внутренней доли околоцветника (англ. petal width).



1.1.5 | Примеры кода

1.2 | Кластеризация

1.2.1 | Определение



Кластерный анализ (англ. cluster analysis)
— многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы.[8]

1.3 | Визуализация

1.3.1 | Набор ирисов Фишера



1.3.2 | Примеры кода

Задание:

- Выбрать любые две пары признаков
- Отобразить на одном рисунке две зависимости
- Подписать график и оси
- Добавить легенду.

1.4 | Кластеризация

1.4.1 | Методы

Методы:

- К-средних (k-means)
- Иерархические

1.5 | Машинное обучение. Определение

Машинное обучение (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач. Для построения таких методов используются средства математической статистики, численных методов, методов оптимизации, теории вероятностей, теории графов, разработанных на базе алгоритмов эффективности форм [9].

Список литературы

- [1] developers.google.com. *Choosing the type of reCAPTCHA*. [Онлайн доступ 30-июня-2018]. 2018. URL: <https://developers.google.com/recaptcha/docs/versions#v1>.
- [2] gouti.ru. *Робот в панике*. [Онлайн доступ 30-июня-2018]. 2011. URL: <http://gouti.ru/2011/07/17/1/comments>.
- [3] habr.com. *Программа Zen обыграла в го профессионального игрока 9 дана с форой в 4 камня*. [Онлайн доступ 30-июня-2018]. 2012. URL: <https://habr.com/post/140244/>.
- [4] Popular Mechanics. *Checkmate, Human: How Computers Got So Good at Chess*. [Онлайн доступ 30-июня-2018]. 2016. URL: <https://www.popularmechanics.com/technology/a19914/chess-computers/>.
- [5] Википедия. *ReCAPTCHA*. [Онлайн доступ 27-июня-2018]. 2018. URL: <https://en.wikipedia.org/wiki/ReCAPTCHA>.
- [6] Википедия. *Визуализация*. [Онлайн доступ 21-февраля-2018]. 2018. URL: <https://ru.wikipedia.org/wiki/%D0%92%D0%88%D0%B7%D1%83%D0%B0%D0%BB%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D1%8F>.
- [7] Википедия. *Капча*. [Онлайн доступ 27-июня-2018]. 2018. URL: <https://ru.wikipedia.org/wiki/%D0%9A%D0%B0%D0%BF%D1%87%D0%BD%D0%BE>.
- [8] Википедия. *Кластерный анализ*. [Онлайн доступ 21-февраля-2018]. 2018. URL: <https://ru.wikipedia.org/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9%D0%B0%D0%BD%D0%BD%D0%BB%D0%B8%D0%B7>.

