Задача классификации регионов России по типу экономики с помощью knn в R: пакет class

Введение

Мы уже рассказывали как к можно использовать алгоритм ближайших соседей (knn) для прогноза значений временного ряда (пакет R tsfknn). Здесь я покажу, как использовать классический алгоритм knn для решения конкретной задачи классификации российских регионов по типу экономики. Предварительно дадим краткое описание сути алгоритма knn.

Краткое описание классического алгоритма knn

Алгоритм knn - один из самых простых алгоритмов классификации. Даже при такой простоте он может дать очень конкурентоспособные результаты. Этот алгоритм является алгоритмом с учителем, где известна цель, но неизвестна траектория движения к ней. Понимание алгоритма ближайших соседей составляет квинтэссенцию машинного обучения. Как же работает этот алгоритм ? Лучше всего это можно понять на простом конкретном примере.

Рассмотрим следующую задачу. Имеется класс из пятнадцати учеников. Учитель исходя из среднего балла учеников по гуманитарным дисциплинам (литература, языки, история и т.д.) и среднего балла по точным естественно-научным дисциплинам (математика, физика, химия и.т.д) делит учеников на следующие кластеры: положительные универсалии (отличники), лирики, физики, негативные универсалии (троечники)

Данные на основании которых учитель осуществил свою классификацию представлены ниже:

```
   x_est y_gum   cl_u

1    4.7   4.9  Yplus

2    3.5   4.9      G

3    4.7   3.6      E

4    4.6   5.0  Yplus

5    3.3   3.0 Yminus
```

```
6    4.9   3.9      E

7    4.8   3.5      E

8    3.3   4.9      G

9    3.5   4.7      G

10   4.9   4.9   Yplus

11   3.3   3.3 Yminus

12   3.7   4.8      G

13   3.9   4.6      G

14   5.0   3.3      E

15   3.0   3.1 Yminus
```

В класс пришли три новых ученика, у которых имется следующая успеваемость:

```
  xnov_est ynov_gum

1      4.0      4.0

2      4.6      4.1

3      3.5      4.2
```

Перед учителем стоит задача к какому кластеру отнести вновь прибывших учеников.

Визуально картина выглядит следующим образом (рис. 1)

Как видно из рисунка 1 раньше ученики в классе были распределены по достаточно компактным четырем кластерам (G - лирики, E- физики , Yplus -отличники, Yminus -троечники). Новые ученики (отмечены красными кружками) не попадают в привычную картинку, они располагаются между сформированных кластеров. Учитель в растеренности

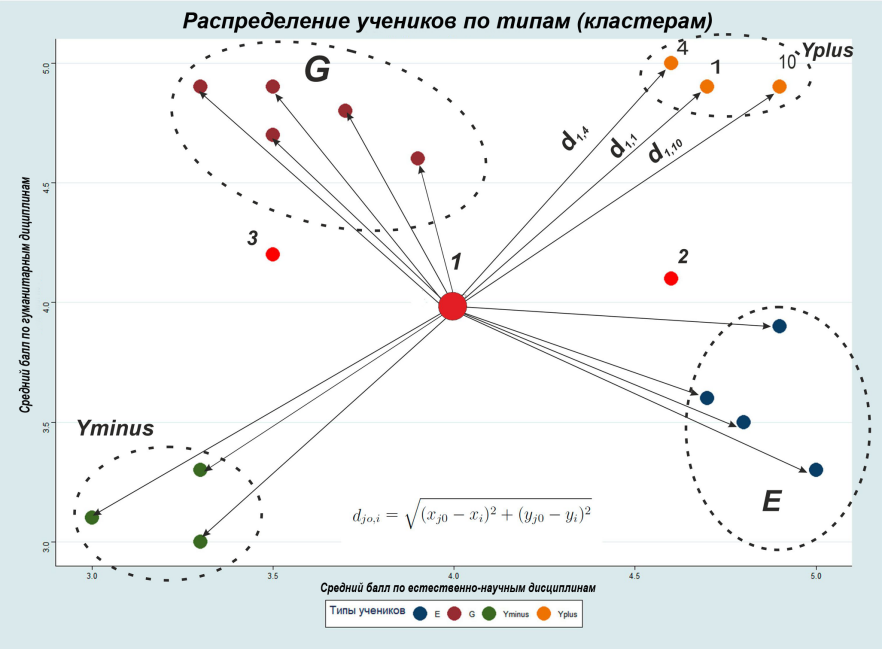


Рис. 1: Распределение учеников по кластерам в пространстве x,y

к какому кластеру отнести вновь прибывших. И здесь ему на помощь приходит алгоритм ближайших соседей.

Суть алгоритма весьма проста. В пространстве признаков классифицируемых объектов вводится некоторая метрика, т.е функция, которая вычисляет для любых двух объектов по их кординатам расстояние между ними. Наиболее известной подобной метрикой является евклидово расстояние. В нашем случае, ученики расположены в двухмерном пространстве и евклидово расстояние между любым новым учеником и любым сторожилом класса может расчитываться по формуле:

$$d_{jo,i} = \sqrt{(x_{j0} - x_i)^2 + (y_{j0} - y_i)^2}$$

Используя эту формулу, можно посчитать расстояние между любым новым учеником всеми старыми учениками. В результате получили следующую матрицу:

1.	2.	3.	4.	5.	6.	7.	8.
1.14	1.03	0.81	1.17	1.22	0.91	0.94	1.14
0.81	1.36	0.51	0.90	1.70	0.36	0.63	1.53
1.39	0.70	1.34	1.36	1.22	1.43	1.48	0.73

9.	10.	11.	12.	13.	14.	15.
0.86	1.27	0.99	0.85	0.61	1.22	1.35
1.25	0.85	1.53	1.14	0.86	0.89	1.89
0.50	1.57	0.92	0.63	0.57	1.75	1.21

Рассмотрим эту матрицу внимательнее. Пусть нас интересует новый ученик с номером 1, его расстояния со всеме старыми учениками представлены в первой строке матрицы (на рис. 1 в виде соотвествующих стрелок). Найдем для него ближайшего соседа. Таким является ученик 13 (0.608), который принадлежит кластеру G. Вторым по близости является ученик 3 (0.806, E), третьим ученик 12 (0.854, G), четвертым ученик 9 (0.860, G). Этот процесс можно продолжить и дальше. Но что это дает нам с точки зрения решения задачи ? А это дает нам информацию для принятия решения об принадлежности 1-го нового ученика к соответствующему кластеру. Но чтобы использовать эту информацию мы должны сначала принять решение на каком шаге нужно прекратить процесс поиска ближайших соседей. То есть определить параметр k и сформулировать правило принятия решения о принадлежности оцениваемого объекта на основании информации о ближайших k соседях. Интуитивно понятно,что число k должно быть относительно небольшим (как минимум на порядок меньше числа используемых для оценки объектов) и желательно быть нечетным. Так как самый простое правило может состоять в следующем: оцениваемый объект относится к тому кластеру, к которому принадлежит большинство из k ближайших соседей. В нашем случае наиболее целесообразным является k=3 (G,E,G) при k=4 (G,E,G,G) , следовательно, 1-го нового ученика необходимо отнести к клатеру G. Аналогично поступаем со вторым и третьим новыми учениками. Графически сразу видно,что второй новый ученик будет отнесен к кластеру E, а третий к кластеру G.

В заключение описания сути алгоритма ближайших соседей следует отметить что величина k зависит от топографии пространства признаков объектов, для компактых кластеров удаленных друг от друга велечина k может быть существенно меньше , чем для размазанных больших класте-

ров. На практике для применения этого алгоритма необходимо располагать достаточно большой выборкой объектов, уже классифицированных внешним учителем (экспертом). Для оценки эффективности точности алгоритма эта выборка делится на две части обучающую (train) и тестовую (test). На первой алгоритм как бы настраивается, а на второй оценивается его точность. При этом целесообразно генерировать некоторое множество вариантов деления выбоки на train и test в этом случае повышется статистическая устойчивость оценок точности алгоритма и появляется возможность формальной оценки величины k. Имеется еще одна проблема, связанная с использованием этого алгоритма. Признаки объектов исходно являются неравноценными с точки зрения их масштаба, для снятия этой проблемы используются различные методы нормирования признаков.

Ниже я покажу, как решаются эти проблемы при решении конкретной экономической задачи средствами R

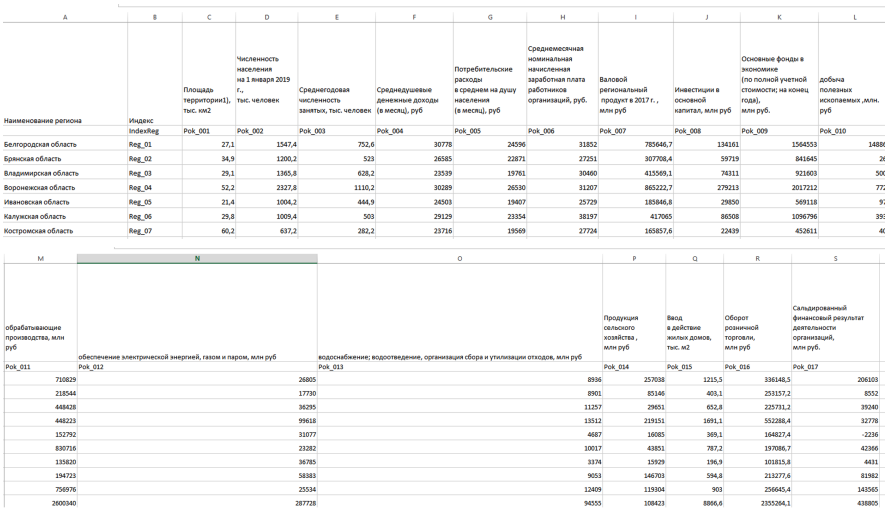
Решение задачи классификации российских регинов средствами R

Постановка задачи

Россия огромная страна, которая отличается большим разнообразием территориальных, природно-климатических и демографических условий. В настоящее время в нашей стране выделено 85 субъектов (областей, республик, краев), чтобы эффективно управлять таким количеством территориальных объектов необходимо уметь классифицировать их, выделяя относительно небольшое число кластеров однородных по тем или иным признакам субъектов. Одной из подобных задач является задача кластеризация регионов России по типу региональной экономики с возможностью анализа ее временной динамики.

В качестве доступной и регулярной информации для решения этой задачи может использоваться информация ежегодно публикуемая Ростатом по основным социально-экономическим показателям субъектов РФ. Ниже на рисунке представлен фрагмент исходной социально экономической информации о 81 субъекте РФ за 2018 год (несколько регионов учтены в составе более крупных регинов и исключен как отдельный регион город Севастополь, как несопоставимый с другими регионами)

Эта таблица используется как исходные данные, которые подверглись нами предворительной обработке без привлечения инструментов R. Были выделены: Рок_007 - валовой региональный продук, скорректированный на 2018 год; Рок_010 -добыча полезных ископаемых;Рок_011



A	B	C	D	E	F	G	H	I	J	K	L
Наименование региона	Индекс IndexReg	Площадь территории1), тыс. км2	Численность населения на 1 января 2019 г., тыс. человек	Среднегодовая численность занятых, тыс. человек	Среднедушевые денежные доходы (в месяц), руб	Потребительские расходы в среднем на душу населения (в месяц), руб	Среднемесячная номинальная начисленная заработная плата работников организаций, руб.	Валовой региональный продукт в 2017 г. , млн руб	Инвестиции в основной капитал, млн руб	Основные фонды в экономике (по полной учетной стоимости; на конец года), млн руб.	добыча полезных ископаемых ,млн. руб
		Pok_001	Pok_002	Pok_003	Pok_004	Pok_005	Pok_006	Pok_007	Pok_008	Pok_009	Pok_010
Белгородская область	Reg_01	27,1	1547,4	752,6	30778	24596	31852	785646,7	134161	1564553	1488(
Брянская область	Reg_02	34,9	1200,2	523	26585	22871	27251	307708,4	59719	841645	2(
Владимирская область	Reg_03	29,1	1365,8	628,2	23539	19761	30460	415569,1	74311	921603	50(
Воронежская область	Reg_04	52,2	2327,8	1110,2	30289	26530	31207	865222,7	279213	2017212	77.
Ивановская область	Reg_05	21,4	1004,2	444,9	24503	19407	25729	185846,8	29850	569118	9:
Калужская область	Reg_06	29,8	1009,4	503	29129	23354	38197	417065	86508	1096796	39:
Костромская область	Reg_07	60,2	637,2	282,2	23716	19569	27724	165857,6	22439	452611	4(

M	N	O	P	Q	R	S
обрабатывающие производства, млн руб	обеспечение электрической энергией, газом и паром, млн руб	водоснабжение; водоотведение, организация сбора и утилизации отходов, млн руб	Продукция сельского хозяйства , млн руб	Ввод в действие жилых домов, тыс. м2	Оборот розничной торговли, млн руб	Сальдированный финансовый результат деятельности организаций, млн руб.
Pok_011	Pok_012	Pok_013	Pok_014	Pok_015	Pok_016	Pok_017
710829	26805	8936	257038	1215,5	336148,5	206103
218544	17730	8901	85146	403,1	253157,2	8552
448428	36295	11257	29651	652,8	225731,2	39240
448223	99618	13512	219151	1691,1	552288,4	32778
152792	31077	4687	16085	369,1	164827,4	-2236
830716	23282	10017	43851	787,2	197086,7	42366
135820	36785	3374	15929	196,9	101815,8	4431
194723	58383	9053	146703	594,8	213277,6	81982
756976	25534	12409	119304	903	256645,4	143565
2600340	287728	94555	108423	8866,6	2355264,1	438805

Рис. 2: Фрагмент таблицы основных социально-экономических показателей регионов РФ за 2018 год

- обрабатывающие производства; Рок_014 - продукция сельского хозяйства; Рок_016 - оборот розничной торговли.

Мы априори выделили следующие типы региональной экономики: добывающая (D); промышленная (P) ; аграрная (A), торговая (T). Было необходимо экспертно распределить 81 регион по этим кластерам.

Нами реаливовалась следующая технология этого разбиения субъектов. На основе исходных показателей в млн. р. расчитываем относительные показатели как отношения показателя, характерезующего вид экономической деятельности к величине валового регионального продукта. Затем эти относительные показатели ранжировались от большего к меньшему. В реультате для каждого региона был сформиров вектор , например, для Белгородской области такого вида [27,16,7,56], где кажый элемент это место региона по соотвествующему относительному показателю.

Дальше использовали интуитивно логичное правило, регион относился к тому кластеру по показателю которого он занимал наивысшее место, например, для Белгородской области наивысшее место 7 по показателю сельское хозяйство, следовательно, она относится к классу A. В результате получена следующая таблица.

Второй столбец этой таблицы мы присоединили к исходной инфор-

Белгородская область	[27,16,7,56]	A
Брянская область	[78,30,10,8]	T
Владимирская область	[53,11,54,33]	P
Воронежская область	[57,46,14,16]	A
Ивановская область	[63,20,49,6]	T
Калужская область	[56,1,41,47]	P
Костромская область	[73,21,46,22]	P
Курская область	[22,47,4,30]	A
Липецкая область	[51,2,16,39]	P
Московская область	[69,31,69,19]	T
Орловская область	[77,42,6,20]	A
Рязанская область	[64,19,28,34]	P
Смоленская область	[59,28,50,23]	T
Тамбовская область	[79,41,1,14]	A
Тверская область	[71,22,44,26]	P
Тульская область	[54,6,36,38]	P
Ярославская область	[70,24,56,48]	P
Республика Карелия	[18,48,73,45]	D
Республика Коми	[8,60,74,76]	D
Архангельская область	[10,61,76,69]	D
Вологодская область	[75,5,60,66]	P
Калининградская область	[37,4,51,60]	P
Ленинградская область	[48,8,47,58]	P
Мурманская область	[25,56,79,64]	D
Новгородская область	[66,27,45,55]	P
Псковская область	[52,32,15,10]	A
Республика Адыгея	[41,44,18,1]	T
Республика Калмыкия	[81,80,2,70]	A
Республика Крым	[38,65,34,11]	T
Краснодарский край	[45,50,26,21]	T
Астраханская область	[6,68,43,59]	D
Волгоградская область	[32,7,27,43}	P
Ростовская область	[44,35,22,13]	T
Республика Дагестан	[58,74,19,4]	T
Республика Ингушетия	[47,77,24,50]	A
Кабардино-Балкарская Республика	[74,66,5,3]	T
Карачаево-Черкесская Республика	[39,55,3,40]	A
Республика Северная Осетия	[62,67,23,5]	T
Чеченская Республика	[42,75,29,2]	T
Ставропольский край	[50,51,9,9]	A
Республика Башкортостан	[24,17,37,18]	P
Республика Марий Эл	[67,13,13,41]	P
Республика Мордовия	[80,18,8,54]	A
Республика Татарстан	[19,15,40,53]	P
Удмуртская Республика	[13,33,36,57]	D
Чувашская Республика	[68,29,31,29]	P
Пермский край	[20,12,68,51]	P
Кировская область	[65,25,32,17]	T
Нижегородская область	[76,10,61,24]	P
Оренбургская область	[9,54,33,65]	D
Пензенская область	[72,43,17,28]	A
Самарская область	[21,22,58,46]	D
Саратовская область	[34,40,21,35]	A
Ульяновская область	[46,26,38,31]	P
Курганская область	[49,45,20,27]	A
Свердловская область	[40,14,67,36]	P
Тюменская область	[4,64,78,81]	D
Челябинская область	[36,9,48,63]	P
Республика Алтай	[35,71,11,25]	T
Республика Тыва	[12,81,42,62]	D
Республика Хакасия	[17,52,57,61]	D
Алтайский край	[55,38,12,12]	A
Красноярский край	[15,37,65,74]	D
Иркутская область	[11,53,62,73]	D
Кемеровская область	[2,39,64,68]	D
Новосибирская область	[33,49,55,52]	D
Омская область	[61,3,30,42]	P
Томская область	[14,59,59,71]	D
Республика Бурятия	[28,62,52,7]	T
Республика Саха (Якутия)	[5,76,70,77]	D
Забайкальский край	[16,72,53,32]	D
Камчатский край	[31,36,66,75]	D
Приморский край	[43,63,63,37]	T
Хабаровский край	[29,58,71,44]	D
Амурская область	[23,69,25,15]	T
Магаданская область	[7,78,75,78]	D
Сахалинская область	[1,73,77,79]	D
Еврейская автономная область	[26,70,39,49]	D
Чукотский автономный округ	[3,79,72,80]	D
г. Москва	[30,57,80,72]	D
г. Санкт-Петербург	[60,34,81,67]	P

Рис. 3: Распределение регионов РФ в 2018 год по типу экономики

мационной таблице (рис.2) и полученные таким образом данные использовали в дальнейшем при решении задачи средствами R (в принципе вся предедущая технология могла бы быть также реализована в R, но мы имитировали экспертную оценку)

Решение задачи и исследование полученных результатов

Ввод исходных и их трансформация для использования классификатора

Начальный фрагмент R скрипта:

```
library(openxlsx)
library(class)
library(gmodels)

rm(list=ls(all=TRUE))

#ЗАГРУЗКА ИЗ EXSEL ПЕВИЧНЫХ ДАННЫХ
N=1 # номер листа с данными
x <-read.xlsx("Region_Data4.xlsx", sheet = N)
#x

xreg <-data.frame(Reg=x$IndexReg,P1=x$Pok_001,P2=x$Pok_002,

P3=x$Pok_003,P4=x$Pok_004,

P5=x$Pok_005,P6=x$Pok_006,P7=x$Pok_007,P8=x$Pok_008,

P9=x$Pok_009,P10=x$Pok_010,

P11=x$Pok_011,P12=x$Pok_012,P13=x$Pok_013,P14=x$Pok_014,

P15=x$Pok_015,P16=x$Pok_016,

P17=x$Pok_017,P18=x$Pok_018)

#xreg
w=c(11,12,15,17,19)
xreg1 <-xreg[,w]
xreg2 <-xreg[,8]
```

```
xreg11 <-xreg1[,-5]
xreg_f <-xreg1[,5]
xreg_str <-xreg11/xreg2
xreg_f <- as.factor(xreg_f)
xreg_f
```

Сначала загружаются три необходимых пакета (для ввода данных их таблицы excel, для реализации классификатора ближайших соседей и оценки результатов по кросс таблице). затем загружается исходная таблица x и из нее формируется фрейм R xreg.

Из фрейма xreg выбираются столбы 11,12,15,17,19 (все необходимые дальше данные) из этих данных формируется фрейм xreg1 из пяти столбцов. Из исходного врема выдергивается стобец 8 (валовой региональный продукт) и формируется соответствующий фрейм xreg2. Из фрейма xreg1 исключается 5-ый стобец (содержащий обозначения типов региональной экономики) и формируется фрейм xreg11. Из фрейма xreg1 выбирается пятый стобец и формируется фрейм (вектор) xreg_f обозначений кластеров. И наконец путем деления соотвествующих элементов фрейма xreg11 на соответствующие элементы вектора xreg2 получаем фрейм относительных значений экономических показателей, а также формируем фактор классов xreg_f. Теперь почти все готово для начала реализации классификации по алгоритму ближайших соседей.

Исследования влияния на ошибку велечины K

Но для реализации этого алгоритма (и очень многих других алгоритмов) необходимо нормализовать данные (отобразить все все показатели в в метрическую шкалу 0,1). Для этого создаем номирующую функцию norm и с ее помощью получаем нормированные данные xstr_nor. Осталось только разбить все регионы на две части обучающую и тестовую. Обычно обучающая выборка содержит 70-80 процентов от исходных данных (возмем 80). воспользуемся оператором sample, который позволет случайно выбирать номера объектов, включаемых в обучающую и тестоваю выборки. Получаем reg_train и reg_test, очищаем их от данных о классах и формируем очищенные reg_train, reg_test а формируем обучающий и тестовый вектора классов xreg_ftrain, xreg_ftest.

Теперь все готово, остался один, но очень важный вопрос. Какое значение нужно взять для k ? И тут есть проблема однозначного ответа нет. В литературе предлагают брать величину k как корень квадрат от числа объектов y на это порядка 9 , но насколько это обоснованно трудно

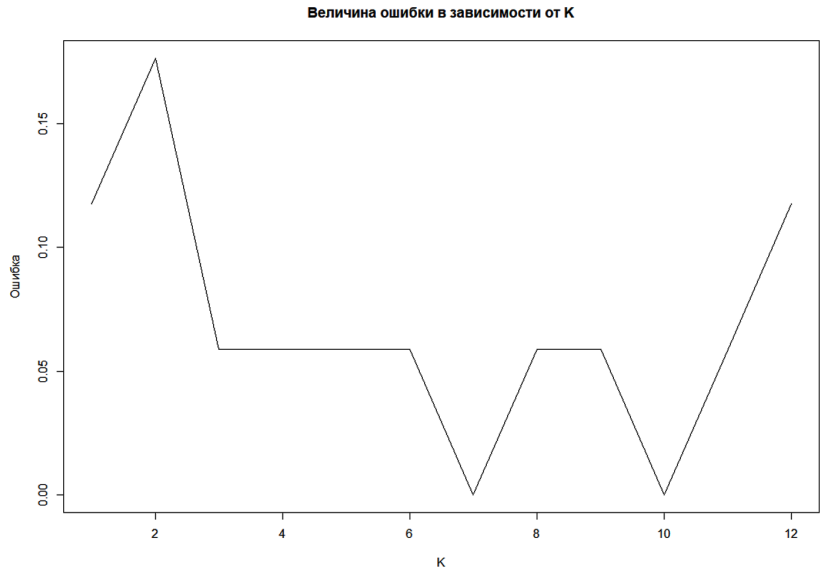


Рис. 4: Величина ошибки в зависимости от k

сказать.

Мы реализуем процедуру позволяющую оценивать влияние величины k на величину ошибки (k изменяется от 1 до 12) Результаты представляем в виде графика (рисунок 4)

Часть скрипта реализующего выше описанные процедуры имеет вид:

```
#нормирующая функция--------------------------
norm <-function(x){
return ((x-min(x))/(max(x)-min(x)))
}

xstr_nor <-as.data.frame(lapply(xreg_str[1:4],norm))
xstr_nor <-cbind(xstr_nor,xreg_f)

#Исследования влияния на ошибку велечины K-----

set.seed(4948493)
reg_sample<-sample(1:nrow(xstr_nor),size=nrow(xstr_nor)*.8)
```

```
reg_train<-xstr_nor[reg_sample,] #Выбор 80 % строк
reg_test<- xstr_nor[-reg_sample,] #Выбор 20 % строк

xreg_ftrain <-reg_train[,5]
xreg_ftest <-reg_test[,5]
reg_train <-reg_train[,-5]
reg_test <-reg_test[,-5]
#reg_test

reg_acc<-numeric()
for(i in 1:12){
 predict<-knn(reg_train ,reg_test,
 xreg_ftrain,k=i)
 reg_acc<-c(reg_acc,
 mean(predict==xreg_ftest))
}

reg_acc
plot(1-reg_acc,type="l",ylab="Ошибка",
 xlab="K",main="Величина ошибки в зависимости от K")

reg.TEST <-cbind(reg_test,xreg_ftest,predict)
reg.TEST
```

Посмотрим на график полученных результатов и видим ,что нулевая ошибка получена при k=7 и k=10. Но можно ли доверять этим результатам. Статистически эти результататы мало значимы, так как мы их получили, используя единственный способ разбиения субъектов на обучающую и тестовую последовательности, а таких способов может быть сгенерировано очень много.

Исследования влияния на ошибку величины k при вариации выборок обучения и теста

Исследуем влияние k на величину ошибки на множестве вариантов обучающих и тестовых последовательностей (500 выборок). В этом случае при каждом значении k ошибка представляет собой случайную величину с определенным законом распределения.

В качестве ее оценки используем величину среднего числа ошибочных классификаций на множестве выборок обучения и теста для каждого

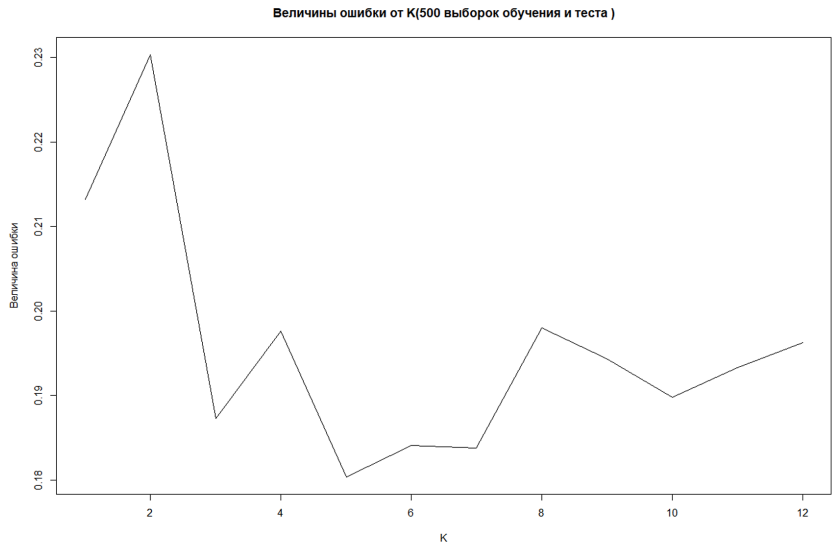


Рис. 5: Величины ошибки от k (500 выборок обучения и теста)

значения $k = 1 : 12$. Графические результаты предствлены на рисунке 5.

Как видно из графика имется довольно четкая область минимальных ошибок при k от 5 до 7.

Часть скрипта реализующего выше описанные процедуры имеет вид:

```
#Исследование влияния на ошибку величины K
#при вариации обучения и теста
Kn=12
trial_sum<-numeric(Kn)
trial_n<-numeric(Kn)
set.seed(6033850)

for(i in 1:500){
reg_sample<-sample(1:nrow(xstr_nor),size=nrow(xstr_nor)*.8)
reg_train<-xstr_nor[reg_sample,] #Выбор 80% строк
reg_test<- xstr_nor[-reg_sample,] #Выбор 20% строк
test_size<-nrow(reg_test)

xreg_ftrain <-reg_train[,5]
```

```
xreg_ftest <-reg_test[,5]
reg_train <-reg_train[,-5]
reg_test <-reg_test[,-5]

for(j in 1:Kn){
predict<-knn(reg_train,reg_test,xreg_ftrain,k=j)
trial_sum[j]<-trial_sum[j]+sum(predict==xreg_ftest)
trial_n[j]<-trial_n[j]+test_size
}
}

x11()
plot(1-(trial_sum / trial_n),type="l",ylab="Величина ошибки",
xlab="K",main="Величины ошибки от K(500 выборок обучения и теста )")

CrossTable(x=xreg_ftest,y=predict,prob.chisq = FALSE)
```

В заключение отметим следующее. Построенный классификатор на данных для момента t (2018 г.) может использоваться в качестве обучающей выборки для момента t+1, данные за t и t+1 в качестве обучающей выборки для момента t+2. В результате можно увидеть наличие или отсутствие изменений в типе экономики каждого субъекта РФ за определенный период времени.

Преимущества и недостатки алгоритма knn

Преимущества: алгоритм имеет беспристрастный характер и не делает предварительных предположений о данных. Будучи простым и эффективным, он легко реализуем и приобрел большую популярность.

Недостатки: данный метод не создает каких-либо моделей или правил, обобщающих предыдущий опыт, в выборе решения они основываются на всем массиве доступных исторических данных, поэтому невозможно сказать, на каком основании строятся ответы. При использовании метода возникает необходимость полного перебора обучающей выборки при распознавании, следствие этого -вычислительная трудоемкость.

Автор Варюхин А.М.

ВЕРСТКА LATEX 20.12.2020