Método de los k- vecinos más próximos (KNN)

Grupo: Far Away

05 de Mayo de 2024

| Integrantes |  |
| --- | --- |
| Burgos Flaqué, María Fernanda | [fernandaburgosflaque@gmail.com](mailto:fernandaburgosflaque@gmail.com) |
| Coronel, Yesica Sidanelia | [yesicacoronel.cpn@gmail.com](mailto:yesicacoronel.cpn@gmail.com) |
| Francomano, Pablo Nicolás | [pablo.francomano@gmail.com](mailto:pablo.francomano@gmail.com) |
| Maccari, Pablo Augusto | [pablomaccari@hotmail.com](mailto:pablomaccari@hotmail.com) |
| Pipinich, Joaquin | [joaquinpipinich@gmail.com](mailto:joaquinpipinich@gmail.com) |

## Kesimo Vecino Más Próximo (KNN)

### Análisis de caso

#### Introducción

El método de los Kesimos vecinos mas cercanos es una técnica de agrupamiento mediante la cual los datos analizados son asignados a cierta cantidad de grupos, dependientes de un valor k. La asignación de cada registro a cada cluster se ejecuta basándose en la distancia al centroide más cercano. Esta agrupación permite la clasificación de nuevos datos o casos, permitiendo llegar a conclusiones según el grupo al que sean asignados.

El siguiente trabajo busca utilizar este método para poder predecir en que tipo de actividad industrial o comercial se ubica una empresa de la cual solo se conocen dos variables: riesgo y rentabilidad.

#### Hipótesis

Se puede estimar dentro de que 5 actividades se encuentra una empresa cualquiera (llamada A) conociendo su riesgo y su rentabilidad.

#### Desarrollo del método KNN

##### Paso 1. Carga de datos y primer análisis

El primer paso para el análisis del caso mencionado consiste en la carga de un set de 254 datos de riesgo- rentabilidad, para empresas de Mendoza.

library(readr)  
datos <- read\_delim("https://themys.sid.uncu.edu.ar/rpalma/R-cran/Frontera\_Pareto.csv",   
 ";", escape\_double = FALSE, locale = locale(decimal\_mark = ",",   
 grouping\_mark = "|"), trim\_ws = TRUE)

## `curl` package not installed, falling back to using `url()`  
## Rows: 254 Columns: 2  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ";"  
## dbl (2): Riesgo, Rentabilidad  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

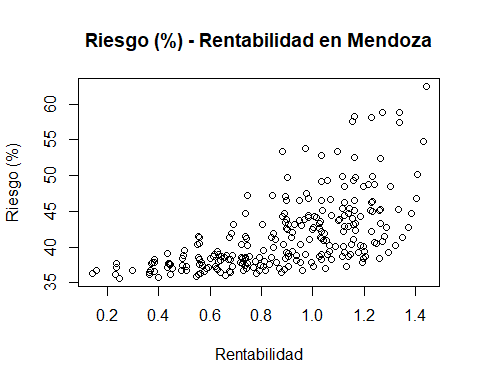
Como muestra, se imprimen las primeras filas a continuación.

head(datos)

## # A tibble: 6 × 2  
## Riesgo Rentabilidad  
## <dbl> <dbl>  
## 1 0.230 36.2  
## 2 0.139 36.4  
## 3 0.155 36.7  
## 4 0.235 37.1  
## 5 0.235 37.7  
## 6 0.362 36.2

Es buena práctica graficar los datos para poder detectar a simple vista potenciales patrones, tendencias, relaciones y estructuras de los datos. Al estar analizando dos variables continuas, un gráfico de dispersión muestra cómo están relacionadas las dos variables del análisis, colocando la variable Rentabilidad en el eje x y una segunda, Riesgo (%), en el eje y.

my <- datos$Rentabilidad   
mx <- datos$Riesgo  
  
plot(mx,my,type="p", main="Riesgo (%) - Rentabilidad en Mendoza",ylab="Riesgo (%)",xlab="Rentabilidad")



La grafica muestra una relación creciente entre ambas variables, a mayor riesgo parece haber mayor rentabilidad para la mayoría de los casos y viceversa.

##### Paso 2. Generación de clusters

La generación de clusters comienza seleccionando el parámetro K, buscando que la cantidad de muestras que quedan en cada grupo sean aproximadamente las mismas. La elección del valor K para este método es crítica, en sus extremos un valor K de 1 asignaría un cluster a cada dato, lo que significa un error cero, pero un modelo no útil y sobre ajustado (overfitting). Para 5 clusters el modelo selecciona para cada dato los mejores resultados utilizando el la suma de las distancias entre los datos y el correspondiente centroide de cada cluster (WCSS).

set.seed(123)  
library(cluster)  
fit <- kmeans(datos, 5)  
fit

## K-means clustering with 5 clusters of sizes 54, 53, 14, 105, 28  
##   
## Cluster means:  
## Riesgo Rentabilidad  
## 1 0.9434583 40.31409  
## 2 1.0620127 43.74186  
## 3 1.1985407 56.04491  
## 4 0.6755260 37.52853  
## 5 1.1075578 47.95315  
##   
## Clustering vector:  
## [1] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4  
## [38] 4 4 1 4 4 4 4 4 4 1 4 4 4 4 4 4 4 4 4 4 1 1 1 1 4 4 4 4 4 4 4 4 4 4 4 4 4  
## [75] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 1 1 1 1 1 1 1 1 4 4 4 4 1 1 4 1 4 1 1 4 4 1 1  
## [112] 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 5 5 5 5 5  
## [149] 5 5 5 2 2 2 2 2 2 1 2 5 2 2 1 3 3 5 5 5 3 5 5 5 3 3 3 5 5 5 5 5 3 3 3 3 3  
## [186] 5 2 5 2 2 2 5 5 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 1 1 1 1 1 4 4 4 4 1 1 1  
## [223] 1 1 1 2 2 2 2 2 5 2 5 3 4 4 4 4 4 4 4 4 4 4 4 1 1 1 2 2 5 5 3 3  
##   
## Within cluster sum of squares by cluster:  
## [1] 48.77914 52.30852 127.64933 82.01858 49.56641  
## (between\_SS / total\_SS = 94.5 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

##### Paso 3. Asignación de nombres para los clusters

A continuación se le asigna un nombre a cada uno de los clusters. Para ello se utiliza la denominación “Group\_N” donde N representa un número secuencial. Se podrían también asignar palabras o rubros que faciliten el análisis.

set.seed(123)  
aggregate(datos,by=list(fit$cluster),FUN=mean)

## Group.1 Riesgo Rentabilidad  
## 1 1 0.9434583 40.31409  
## 2 2 1.0620127 43.74186  
## 3 3 1.1985407 56.04491  
## 4 4 0.6755260 37.52853  
## 5 5 1.1075578 47.95315

mi\_cluster <- data.frame(datos, fit$cluster)

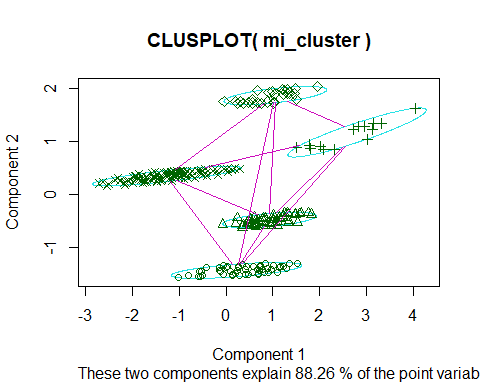
Se puede obtener a continuación la asignacion de cada dato de la problación a cada el cluster en función de riesgo y rentabilidad.

mi\_cluster

## Riesgo Rentabilidad fit.cluster  
## 1 0.2299914 36.17758 4  
## 2 0.1392425 36.35505 4  
## 3 0.1547155 36.71515 4  
## 4 0.2348895 37.14919 4  
## 5 0.2348895 37.67857 4  
## 6 0.3618077 36.18910 4  
## 7 0.4453633 36.17758 4  
## 8 0.3618077 36.37949 4  
## 9 0.3842360 36.62135 4  
## 10 0.3663192 36.74245 4  
## 11 0.2972553 36.75617 4  
## 12 0.5527457 36.16610 4  
## 13 0.5606940 36.30674 4  
## 14 0.5082500 36.70159 4  
## 15 0.4875774 36.72878 4  
## 16 0.4539005 37.05795 4  
## 17 0.4324673 37.16463 4  
## 18 0.3708177 37.64321 4  
## 19 0.3842360 37.80500 4  
## 20 0.4367781 37.67857 4  
## 21 0.4410767 37.53901 4  
## 22 0.4917347 37.43762 4  
## 23 0.5920631 37.11851 4  
## 24 0.5764627 36.63460 4  
## 25 0.6416343 36.45394 4  
## 26 0.6786307 36.41649 4  
## 27 0.6749742 36.62135 4  
## 28 0.6265626 36.86804 4  
## 29 0.5842838 37.01319 4  
## 30 0.6227698 37.05795 4  
## 31 0.6189669 37.32273 4  
## 32 0.5685996 37.48798 4  
## 33 0.5567252 37.60816 4  
## 34 0.5646521 37.87914 4  
## 35 0.5082500 37.35520 4  
## 36 0.4453633 37.53901 4  
## 37 0.4410767 37.76846 4  
## 38 0.3842360 38.33430 4  
## 39 0.3753033 37.84189 4  
## 40 0.4324673 39.06271 1  
## 41 0.4958806 38.79837 4  
## 42 0.4917347 38.41889 4  
## 43 0.5567252 38.31342 4  
## 44 0.5527457 38.50519 4  
## 45 0.5998011 38.41889 4  
## 46 0.6113311 38.66058 4  
## 47 0.6151541 39.01345 1  
## 48 0.6639477 38.84527 4  
## 49 0.6565487 38.61562 4  
## 50 0.6786307 38.31342 4  
## 51 0.6416343 38.09045 4  
## 52 0.6341182 37.99302 4  
## 53 0.6713082 38.33430 4  
## 54 0.7253181 38.37637 4  
## 55 0.7393741 38.59331 4  
## 56 0.6859154 38.84527 4  
## 57 0.6453775 38.63804 4  
## 58 0.6341182 39.01345 1  
## 59 0.6265626 39.44984 1  
## 60 0.5000151 39.55871 1  
## 61 0.5487554 40.43002 1  
## 62 0.7358735 37.38795 4  
## 63 0.7463489 37.64321 4  
## 64 0.7182362 37.84189 4  
## 65 0.7773041 37.95468 4  
## 66 0.8108925 37.97380 4  
## 67 0.8597496 37.78669 4  
## 68 0.8306507 37.53901 4  
## 69 0.7942025 37.47113 4  
## 70 0.8075710 37.24282 4  
## 71 0.7670633 37.10327 4  
## 72 0.7428659 37.04297 4  
## 73 0.7288455 36.79769 4  
## 74 0.7908396 36.79769 4  
## 75 0.8175110 36.78379 4  
## 76 0.8913464 36.81164 4  
## 77 0.8693087 37.02806 4  
## 78 0.9037744 37.22703 4  
## 79 1.0019580 37.27457 4  
## 80 0.9933401 37.82340 4  
## 81 0.9523244 37.80500 4  
## 82 0.9373488 38.19030 4  
## 83 0.7217817 38.52705 4  
## 84 0.7288455 37.99302 4  
## 85 0.7253181 37.67857 4  
## 86 0.6859154 37.30661 4  
## 87 0.7358735 38.41889 4  
## 88 0.7532885 38.68324 4  
## 89 0.5567252 40.52932 1  
## 90 0.5527457 41.43655 1  
## 91 0.6749742 41.31953 1  
## 92 0.6822778 41.84481 1  
## 93 0.7358735 41.43655 1  
## 94 0.7428659 41.16713 1  
## 95 0.7217817 40.43002 1  
## 96 0.7463489 40.17418 1  
## 97 0.7908396 38.57111 4  
## 98 0.8371724 38.54901 4  
## 99 0.8944646 38.41889 4  
## 100 0.9037744 38.75196 4  
## 101 0.9068630 39.03800 1  
## 102 0.9191439 39.36977 1  
## 103 0.9373488 38.89269 4  
## 104 0.9700658 39.01345 1  
## 105 1.0274596 38.75196 4  
## 106 1.0330561 38.94061 1  
## 107 1.0579300 38.94061 1  
## 108 1.0358449 38.27198 4  
## 109 1.0742357 38.25139 4  
## 110 1.0633895 39.42301 1  
## 111 1.0633895 40.20547 1  
## 112 1.0524459 40.94614 1  
## 113 1.0358449 41.24283 1  
## 114 1.0469371 41.92997 1  
## 115 1.0246519 42.56228 2  
## 116 1.0302611 43.53594 2  
## 117 1.0076704 44.23259 2  
## 118 0.9817568 44.41815 2  
## 119 1.0019580 42.75546 2  
## 120 0.9552986 43.00538 2  
## 121 0.9493433 43.70361 2  
## 122 0.9068630 43.16015 2  
## 123 0.8975753 42.65810 2  
## 124 0.8944646 43.42636 2  
## 125 0.8819471 44.23259 2  
## 126 0.8913464 44.67314 2  
## 127 0.8042413 43.21255 2  
## 128 0.8533381 41.05553 1  
## 129 0.8371724 41.39727 1  
## 130 0.8468954 41.92997 1  
## 131 0.8850877 40.36487 1  
## 132 0.8468954 39.96040 1  
## 133 0.8075710 39.47682 1  
## 134 0.8975753 39.26514 1  
## 135 0.9523244 40.36487 1  
## 136 0.9875618 41.09248 1  
## 137 1.0414035 40.98236 1  
## 138 1.0302611 41.39727 1  
## 139 1.0358449 42.14837 2  
## 140 0.9990920 42.42122 2  
## 141 1.0246519 43.26536 2  
## 142 1.0133567 44.17182 2  
## 143 1.0386273 45.07307 2  
## 144 1.0358449 46.49689 5  
## 145 1.0688247 46.66254 5  
## 146 0.9846626 46.74662 5  
## 147 0.9493433 46.49689 5  
## 148 0.9759248 47.54354 5  
## 149 0.8975753 46.57930 5  
## 150 0.8944646 47.09169 5  
## 151 0.8436623 47.18017 5  
## 152 0.8882208 44.73837 2  
## 153 0.8975753 43.87555 2  
## 154 0.9282786 43.16015 2  
## 155 0.9006785 42.51490 2  
## 156 0.9671261 43.81772 2  
## 157 0.9846626 44.17182 2  
## 158 0.9160846 41.28105 1  
## 159 0.9191439 42.10406 2  
## 160 0.7463489 47.18017 5  
## 161 0.7358735 44.73837 2  
## 162 0.6895436 43.21255 2  
## 163 0.5567252 41.31953 1  
## 164 0.8819471 53.34421 3  
## 165 0.9700658 53.84761 3  
## 166 0.9006785 49.73460 5  
## 167 1.0358449 49.16468 5  
## 168 1.0579300 49.38861 5  
## 169 1.0330561 52.86177 3  
## 170 1.1218721 48.41998 5  
## 171 1.1166718 49.85265 5  
## 172 1.1601629 49.38861 5  
## 173 1.1626719 52.55108 3  
## 174 1.0956410 53.34421 3  
## 175 1.1551289 57.61880 3  
## 176 1.1651754 49.73460 5  
## 177 1.1972402 48.52279 5  
## 178 1.2189279 48.73189 5  
## 179 1.2402109 48.73189 5  
## 180 1.2331608 49.85265 5  
## 181 1.2656901 52.39887 3  
## 182 1.3387933 58.81818 3  
## 183 1.2725396 58.81818 3  
## 184 1.2308010 58.08651 3  
## 185 1.1626719 58.32624 3  
## 186 1.3017417 48.41998 5  
## 187 1.2611004 45.07307 2  
## 188 1.2308010 46.25449 5  
## 189 1.2331608 44.93738 2  
## 190 1.1874683 44.23259 2  
## 191 1.1626719 45.14184 2  
## 192 1.1626719 46.57930 5  
## 193 1.1244638 46.25449 5  
## 194 1.1424485 45.42339 2  
## 195 1.1398959 44.67314 2  
## 196 1.1140631 45.35204 2  
## 197 1.1218721 44.35579 2  
## 198 1.0742357 44.48107 2  
## 199 1.1035709 43.81772 2  
## 200 1.1551289 43.99256 2  
## 201 1.1676736 43.31859 2  
## 202 1.1398959 42.90425 2  
## 203 1.1218721 42.90425 2  
## 204 1.1009334 43.76043 2  
## 205 1.1192748 42.23796 2  
## 206 1.1601629 43.10817 2  
## 207 1.1651754 42.28326 2  
## 208 1.1576486 41.09248 1  
## 209 1.1972402 40.05093 1  
## 210 1.1874683 40.23698 1  
## 211 1.1526037 40.30053 1  
## 212 1.1192748 40.08146 1  
## 213 1.0903251 40.08146 1  
## 214 1.1218721 39.29108 1  
## 215 1.1475371 38.94061 1  
## 216 1.1398959 38.66058 4  
## 217 1.1140631 38.68324 4  
## 218 1.1972402 38.37637 4  
## 219 1.2069299 38.72894 4  
## 220 1.2020953 39.44984 1  
## 221 1.1800848 39.29108 1  
## 222 1.2472172 40.52932 1  
## 223 1.2425512 40.69908 1  
## 224 1.2702611 40.73369 1  
## 225 1.2816072 41.51596 1  
## 226 1.2861133 42.80467 2  
## 227 1.2656901 43.31859 2  
## 228 1.2284363 42.14837 2  
## 229 1.2656901 45.28135 2  
## 230 1.2260666 45.07307 2  
## 231 1.2331608 46.33449 5  
## 232 1.3344993 45.21131 2  
## 233 1.3017417 48.41998 5  
## 234 1.3366484 57.39085 3  
## 235 0.2446419 35.62279 4  
## 236 0.3975398 35.72958 4  
## 237 0.5447543 35.88192 4  
## 238 0.6602530 36.04259 4  
## 239 0.7807006 36.24737 4  
## 240 0.8787989 36.45394 4  
## 241 0.9612261 36.66123 4  
## 242 1.0496946 37.02806 4  
## 243 1.1296302 37.30661 4  
## 244 1.1948050 37.78669 4  
## 245 1.2587984 38.44030 4  
## 246 1.2950709 39.29108 1  
## 247 1.3236891 40.17418 1  
## 248 1.3515739 41.39727 1  
## 249 1.3725422 42.75546 2  
## 250 1.3869782 44.73837 2  
## 251 1.4032384 46.83164 5  
## 252 1.4092716 50.09305 5  
## 253 1.4310995 54.73636 3  
## 254 1.4428147 62.48594 3

##### Paso 4. Analisis de componenetes prinicpales (PCA)

clusplot(mi\_cluster,mi\_cluster$fit.cluster)

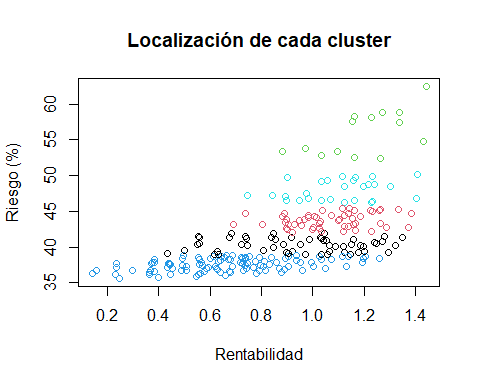


El análisis de componentes principales es técnica estadística de síntesis de la información, o reducción de la dimensión (número de variables). En el caso analizado se utiliza para entender cuan bien explican las dos variables seleccionadas el comportamiento de las empresas. Para este ejemplo, las variables riesgo y rentabilidad explican el 88.26%. Es interesante mencionar que si este valor fuera menor al 50% deberían agregarse mas variables a los datos iniciales y repetir los pasos ya mencionados.

##### Paso 5.Grafica de asignaciones (clusters)

Al graficar nuevamente los datos, pero esta vez diferenciando por color la asignación a cada cluster, se observa la misma relación entre las dos variables analizadas pero esta vez con división por actividad (clusters).

plot(mi\_cluster$Riesgo , mi\_cluster$Rentabilidad , col=mi\_cluster$fit.cluster, main = "Localización de cada cluster",ylab="Riesgo (%)",xlab="Rentabilidad")



##### Paso 5. Entrenamiento KNN

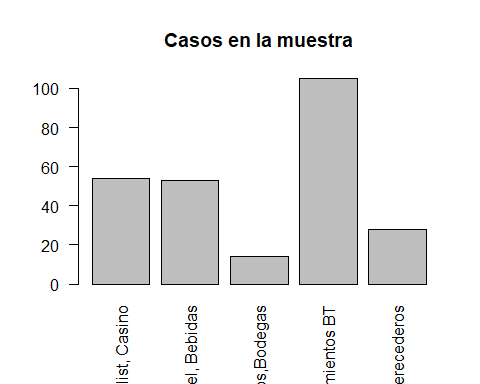
El entrenamiento de un modelo de vecino próximo puede ser una buena opción cuando la población es muy grande. Es como trabajar con una muestra de la población. Para el análisis de este caso se tiene conocimiento de a qué actividad pertenece cada una de las empresas que fueron encuestadas sobre riesgo percibido de sus actividad y rentabilidad.

train <- mi\_cluster  
cl <- factor(mi\_cluster$fit.cluster, levels = c("1", "2","3","4","5"),labels = c("Prost, Penalist, Casino","Bank, Cred\_pers, Tel\_Cel, Bebidas", "Manufactura, Metal\_Mec, Servicios ,Seguros,Bodegas","MOA, Restaurant, Emprendimientos BT","Transporte, Pasajeros, Indumentaria, Perecederos"))

summary(cl)

## Prost, Penalist, Casino   
## 54   
## Bank, Cred\_pers, Tel\_Cel, Bebidas   
## 53   
## Manufactura, Metal\_Mec, Servicios ,Seguros,Bodegas   
## 14   
## MOA, Restaurant, Emprendimientos BT   
## 105   
## Transporte, Pasajeros, Indumentaria, Perecederos   
## 28

plot(cl, main="Casos en la muestra", las = 2)



#### Caso de estudio e interpretación de resultados

Teniendo los datos de rentabilidad y riesgo de una empresa A, se busca conocer a que tipo de actividad pertenece, segun su asignación a un cluster.

* Rentabilidad 50.000 %
* Riesgo 62,23%
* 3 Vecinos más próximos

mi\_caso <- cbind (50,62.23,3 )  
mi\_caso

## [,1] [,2] [,3]  
## [1,] 50 62.23 3

A continuación se analiza a que cluster pertenecen los 20 vecinos más próximos.

set.seed(123)  
library(class)  
knn(train, mi\_caso, k=20, cl, prob=TRUE)

## [1] Manufactura, Metal\_Mec, Servicios ,Seguros,Bodegas  
## attr(,"prob")  
## [1] 0.7  
## 5 Levels: Prost, Penalist, Casino ... Transporte, Pasajeros, Indumentaria, Perecederos

Con este resultado, se puede asegurar con un 70% de nivel de confianza que se trata de una empresa del siguiente grupo:

* Manufactura
* Matalmecánica
* Servicios profesionales
* Agente de seguros
* Bodegas pequeñas

#### Conclusiones

El método de los k- vecinos mas próximos (KNN) es un método fácil de implementar, con pocos hiperparametros (valor k y métrica de distancia) y de fácil adaptación a nuevos datos. Para el tipo de análisis realizado en este trabajo es un método muy ventajoso, permitiendo obtener de manera rápida una buena estimación sobre el caso nuevo a analizar. Por otro lado, es un modelo muy propenso al sobre ajuste al ser tan relevante la elección del valor k.

#### Bibliografia

Vandeput, Nicolas. Data Science for Supply Chain Forecasting. Walter de Gruyter GmbH, 2021.

Grus, Joel. Data Science from Scratch. O´Reilly Media, 2015.

IBM, <https://www.ibm.com/mx-es/topics/knn>, (consultado el 5 de Mayo de 2024)