Estadística

Promedio n (puede ser sensible a los outliers)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Trimmed mean / Promedio recortado (quite elementos del inicio y del fin de los valores y tomando el promedio de los valores restantes

$$\bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$ donde las x están ordenadas

Promedio con peso

$$x_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Mean Absolute deviation
Error medio absoluto

$$\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$$

Desviations / errors / residuals
$\sum x_i -$ ~~x_i~~ $x_{i\,predicho}$

Varianza / mean-squared-error

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$ → el promedio

desviación estándar $\sqrt{\text{varianza}} = s$

→ ¿Qué tanto se aleja cada valor en relación al promedio

Estadística

Promedio $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ (puede ser sensible a los outliers)

Trimmed mean / Promedio recortado (quita elementos del inicio y del fin de los valores y tomando el promedio de los valores restantes

$\bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$ donde las x están ordenadas

Promedio con peso

$x_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$

Mean Absolute deviation
Error medio absoluto

$\frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}$

Desviations / errors / residuals

$\sum x_i$ ~~x_i~~ x_i predicho

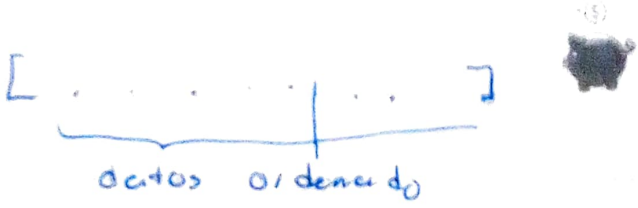
varianza / mean-squared-error

$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$ → el promedio

desviación estándar $\sqrt{\text{varianza}} = s$

→ ¿Qué tanto se aleja cada valor en relación al promedio

80^{th} percentile



50^{th} percentile = media

$IQR = 75^{th}$ percentile $- 25^{th}$ percentile

$\{1, 2, 3, 3, 5, 6, 7, 9\}$

$75^{th} = 6.5 \quad 25^{th} = 2.5$

$\Rightarrow IQR = 6.5 - 2.5 = 4$

Variables correlacionadas

Si X, Y estan correlacionados significa que X crece y al tiempo de que Y lo hace y si es en sentido contrario, estan correlacionados negativamente

- Coeficiente de correlación de Pearson estudia la correlación entre dos variables

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\, s_x s_y}$$

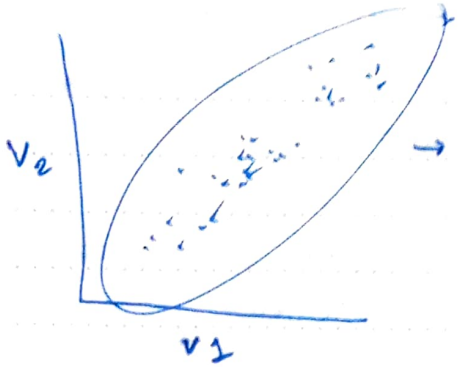
desviación estandar de ambas

Nota: Este valor va de -1 a 1 y 0 implica que no hay correlación

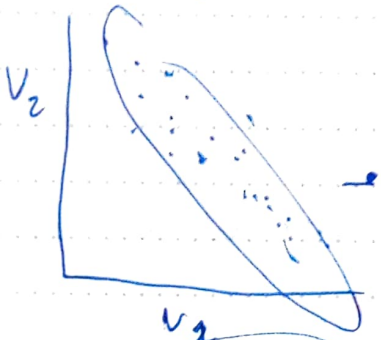
el coeficiente de correlación mide que tanto un par de variables estan relacionadas una de la otra

En un scatter plot, donde en cada eje asignemos a cada variable como algo

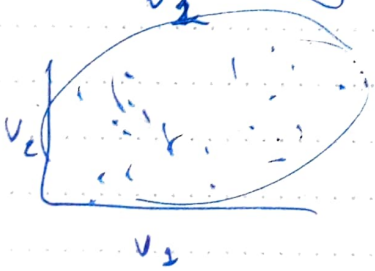
plot.scatter(x = 'V1', y = 'V2', ...)



→ Si tenemos una distribución de puntos algo así, implica que las variables sí están relacionadas, de forma positiva



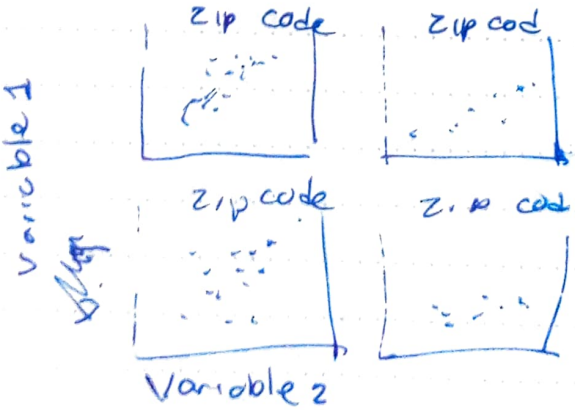
→ correlación negativa



no hay correlación

Análisis Multivariable

Se pueden hacer scatter plots con más de dos variables



- La grafica de agrupamiento hexagonal y la de contornos (como los topográficos) sirven para encontrar la correlación entre dos variables numericas

- Las tablas de contingencia son una herramienta para observar los recuentos de dos variables categoricas

- Los diagramas de caja y de violín expresan una relación entre una variable categorica y otra numerica

Datos y distribuciones muestrales

muestra - un subconjunto de los datos (población)

estrato (división de los datos donde cada elemento comparte cierta característica)

correr experimentos en varios contextos, bajo diferentes condiciones, analizando que tanto es cierto nuestro hipotesis, nos ayuda a evitar el sesgo

Sampling distribution of a statistic

distribución de muestreo de una estadistica

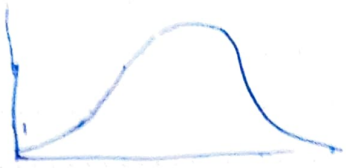
Hablando de los parametros calculables (o parametros estadisticos de una población) podemos aproximar estas medidas (como por ejemplo el promedio μ) a partir de calcular esta medida para una muestra. Sin embargo, como cada muestra que es posible tomar de la población es diferente entre sí, se calcula de cada una este parametro (digamos $\bar{x}$) y esto sera la distribución de una estadistica referente a las muestras.

Esto se hace por cuito inferencia o suposiciones de conjuntos de datos aún más grandes

distribución de los datos ≠ distribución de las muestras

Teorema Central del Limite

Los promedios de las muestras de una población se pueden reorganizar en una distribución normal



Error estandar

$$SE = \frac{S}{\sqrt{n}}$$ } mide la variabilidad de una metrica muestral es decir, que tanto varia la muestra respecto a una estadistica.

qué tanto varia o qué tanto variabilidad existe en el calculo de una estadistica en una distribución muestral

$$SE = \frac{S}{\sqrt{n}} = \frac{\sqrt{\text{la suma de diferencias de cada valor contra el promedio}}}{\sqrt{n}}$$

SE - variabilidad en estadisticos entre muestras

"g. yo tomase una muestra aparte de la misma población, este estadística que tanto variara respecto a la que calculo y en general a los demas"

es la variabilidad respecto a la distribución muestral

S - es la distribución de los datos

"que tanto se alejan del promedio"

Nivel de confianza

porcentaje de confianza en ~~[illegible]~~ que tan exacto es la metrica que estamos calculando

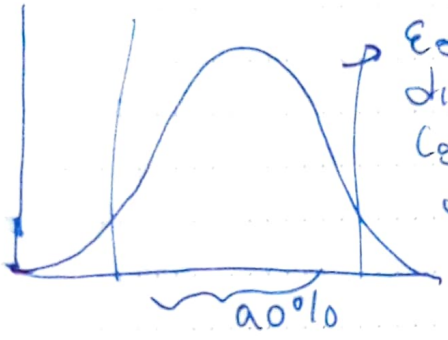
Note

Bootstrap

es una tecnica de remuestreo con reemplazo

El intervalo de confianza (X %) implica que ~~[illegible]~~ de ~~[illegible]~~ otras muestras, el X% de ellos tendrán o darán una estadística dentro de ese intervalo

Entonces



Este es la distribución de una estadística, digamos $\bar{x}$ de todas las muestras posibles (sacadas del bootstrap) así, si yo establezco un intervalo de confianza del 90%, significa que el 90% del valor de $\bar{x}$ para las muestras se encontrará en ese intervalo.

Estandarizar:

$$\frac{x_i - \bar{x}}{s}$$

Regresión y predicción

La regresión lineal se modela

$Y_i = b_0 + b_1 X_i + e_i$ → intercepto

e_i: el error entre lo predicho y el valor real

$\hat{e}_i = Y_i - \hat{Y}_i$

Calcular la regresión lineal de los datos implica ajustar una recta que cruce estos.

para un conjunto de V independiente y otros independientes

x	y
1	2
2	2.5
0	1

tenemos que encontrar los valores b_0, b_1 que más se ajusten a estos,

Tomando la forma normal de la ec. de la recta

$H = X\Theta$ (debemos encontrar Θ

Así,

$\Theta = (X^T X)^T X + Y$ (obtenido al minimizar la función de error)

En el ejemplo

$$\Theta = \left(\begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 0 \end{bmatrix} \right)^T \begin{bmatrix} 1 & 2 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 2.5 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.08 \\ 0.25 \end{bmatrix}$$

Validación cruzada

En lugar de dividir 70% para entrenamiento y 30% para prueba una sola vez y dividiendo el data set arbitrariamente se realiza este proceso ~~estratificado~~ varias diferentes particiones por el conjunto de prueba de manera en que no exista al final ningún elemento sin que haya sido parte de

Probabilidad

Principio Básico del Conteo

y = • o = +

importa el orden
- con reemplazo — $n O_k$ — n^k
- sin reemplazo — $n P_k$ — $\frac{n!}{(n-k)!}$

no importa el orden
- sin reemplazo — $n C_k$ — $\frac{n!}{(n-k)!k!}$

Conjuntos

$C \cup (A \cap B) = (C \cup A) \cap (C \cup B)$

$(A \cap B)^c = A^c \cup B^c$

$(A \cup B)^c = A^c \cap B^c$

Función de probabilidad

$\mathbb{P} = f \longrightarrow [0,1]$

$\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$

$\mathbb{P}(A^c) = 1 - \mathbb{P}(A) \quad \mathbb{P}(\emptyset) = 0$

$\mathbb{P}(B \setminus A) = \mathbb{P}(B) - \mathbb{P}(A)$

$B/A =$ 

$\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C)$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(A|B)P(B)$$

Eventos independientes

$$P(A \cap B) = P(A)P(B)$$

Hadoop

Cluster hadoop

Los masters dividen las operaciones en trozos y los distribuye (las operaciones) entre los trabajadores hadoop (

Entorno distribuido para aplicaciones distribuidas que trabajan con Big data

Procesamiento de datos de forma distribuida

Cuenta con tres partes

HDFS - hadoop file system

Planificación de tareas y negociación de recursos: YARN

Procesamiento distribuido MapReduce (programación distribuida)

Docker

Imagen: archivo que contiene para crear el contenedor

volumen: espacio de disco que pueden compartirse entre contenedores o entre un contenedor y el host

Redes: conecta contenedores

Docker file

HDFS

almacena ficheros grandes

se dividen en bloques que se replican en diferentes nodos del cluster

Namenode - el servidor master

mantiene la información de donde se encuentra los datos

Datanodes - tiene los datos

Rack

una colección (40-50) Datanodes usando el mismo switch de red

Rackawareness selecciona los datanodes más próximos para evitar el tráfico de red

Operaciones de hdfs

read

client interactua con el namenode para la metadata

Comandos de hdfs

hdfs dfs -ls
" " -cat
-appendToFile

} de la misma manera con el linux con el prefijo hdfs

MapReduce

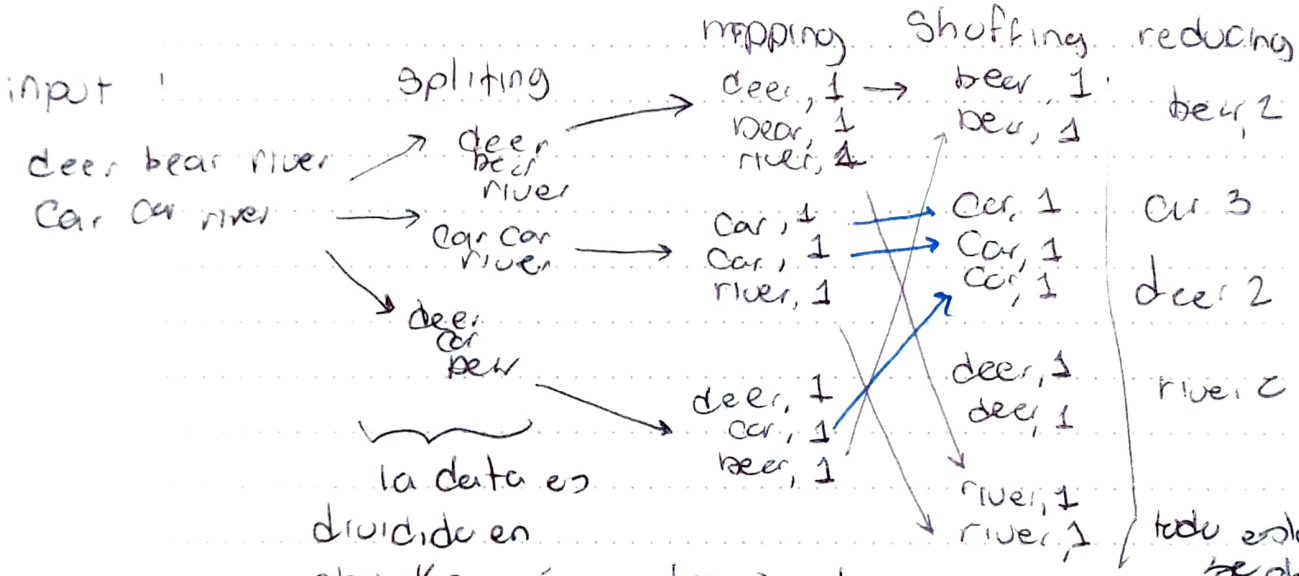
La capa de proceso de hadoop

Terminología

map - toma el valor de la llave como entrada y genera un valor de salida asociado a esa llave
el valor donde debemos de procesar

reduce - procesar la data
lo que sale es el final output y se almacena en hdfs

Job - todo el proceso de mapreduce
task una parte del procesamiento de la data en map o reduce



la data es dividida en chunks más pequeños de data

luego, el mapeo actua sobre la función que le fue pasada como parametro en este caso la función es el conteo de las palabras asignadas

todo esto se almacena en los data nodes

la data se pasa lo más cerca posible (rack awarness)

cuenta las instancias de cada palabra