Distribución de puntos en la esfera. Competición en Kaggle.

Daniel López García

Universidad de Granada

26 de junio de 2018

Contenidos

Distribución de puntos en la esfera.

Objetivos

- Determinar conjuntos de puntos para aproximación, interpolación e integración sobre la esfera y sus propiedades geométricas.
- Simulación y visualización de distribuciones de puntos sobre la esfera.

1. Distribución de puntos en la esfera.
   - Armónicos Esféricos.
   - Cálculo del gradiente.
   - Integración numérica.

Espacios de Polinomios Homogéneos.

Consideramos $\mathcal{H}_n^d$ el espacio de polinomios homogéneos de grado n en d dimensiones. Estas funciones son de la forma:

$$\sum_{|\alpha|=n} a_\alpha x^\alpha, a_\alpha \in \mathbb{C}$$

Ejemplos

$$\mathcal{H}_2^2 = \{a_1 x_1^2 + a_2 x_1 x_2 + a_3 x_2^2\}$$
$$\mathcal{H}_3^2 = \{a_1 x_1^3 + a_2 x_2^3 + a_3 x_1^2 x_2 + a_4 x_1 x_2^2\}$$

Definición

Una función f es armónica si $\triangle f(x) = 0$, es decir

$$\frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + ... + \frac{\partial^2 f}{\partial x_n^2} = 0$$

Llamamos $\mathbb{Y}_n(\mathbb{R}^d)$ al espacio de los polinomios homogéneos de grado n en $\mathbb{R}^d$ que son armónicos.

Armónicos Esféricos.

Definición

Se llama espacio de armónicos esféricos de orden n en d dimensiones a $\mathbb{Y}_n^d = \mathbb{Y}_n(\mathbb{R}^d)_{|\mathbb{S}^{d-1}}$

De la definición se deduce que un armónico esférico $\mathbb{Y}_n \in \mathbb{Y}_n^d$ está asociado a un armónico homogéneo $\mathbb{H}_n \in \mathbb{Y}_n$ de la siguiente forma:

$$\mathbb{H}_n(r\xi) = r^n \mathbb{Y}_n(\xi)$$

Base ortogonal

Teorema

Sean, $T_n(t), U_n(t)$ los polinomios de Chebyshev de 1ª y 2ª clase respectivamente. Y definimos

$$g_{0,n}(x_1, x_2) = (x_1^2 + x_2^2)\, T_n^2(x_2(x_1^2 + x_2^2)^{-1/2})$$

$$g_{1,n-1}(x_1, x_2) = x_1(x_1^2 + x_2^2)^{\frac{n-1}{2}}\, U_{n-1}(x_2(x_1^2 + x_2^2)^{-1/2})$$

entonces, si tomamos $\boldsymbol{n} = (n_1, ..., n_d)$ con $n_1 = \{0, 1\}$ se define

$$Y_{\boldsymbol{n}} = g_{n_1,n_2}(x_1, x_2) \prod_{j=3}^{d} (x_1^2 + ... + x_j^2)^{n_j/2}\, C_{n_j,\lambda_j}(x_j(x_1^2 + ... + x_j^2)^{-1/2})$$

donde $\lambda_j = \lambda_j(n_1, ..., n_{j-1}) = \sum_{i=1}^{j-1} n_i + \frac{j-2}{2}$. Entonces $\{Y_n, |\boldsymbol{n}| = n\}$ es una base de $\mathbb{Y}_n^d$.

1. Distribución de puntos en la esfera.
   - Armónicos Esféricos.
   - Cálculo del gradiente.
   - Integración numérica.

Caso particular. Esfera de dimensión 3.

Tomando coordenadas esféricas,

$$x_1 = r \operatorname{sen}\theta \operatorname{sen}\phi$$
$$x_2 = r \operatorname{sen}\theta \operatorname{sen}\phi$$
$$x_3 = r \cos\theta$$
$$0 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi, r > 0$$

una base ortogonal de $\mathbb{Y}_n^3$ viene dada por

$$\begin{cases} Y_{k,1}^n = (\operatorname{sen}\theta)^k C_{n-k,k+1/2}(\cos\theta)\cos(k\phi), & 0 \leq k \leq n \\ Y_{k,2}^n(x) = (\operatorname{sen}\theta)^k C_{n-k,k+1/2}(\cos\theta)\operatorname{sen}(k\phi), & 1 \leq k \leq n \end{cases}$$

La expresión de las parciales es la siguiente:

Proposición

Para $k = 0, \ldots, n$

$$\partial_1 Y^n_{k,1}(x) = -\frac{(n+k)(n+k-1)}{2(2k-1)} Y^{n-1}_{k-1,2}(x) - (k+\frac{1}{2}) Y^{n-1}_{k+1,2}(x)$$

$$\partial_2 Y^n_{k,1}(x) = \frac{(n+k)(n+k-1)}{2(2k-1)} Y^{n-1}_{k-1,1}(x) - (k+\frac{1}{2}) Y^{n-1}_{k+1,1}(x)$$

$$\partial_3 Y^n_{k,1}(x) = (n+k) Y^{n-1}_{k,1}(x)$$

Puntos críticos del gradiente.

Igualando las expresiones de las parciales a 0 tenemos que

$$\begin{cases} \operatorname{sen}\theta = 0 \\ \cos k\phi = 0 \\ c_{n,k} C_{n-k}^{k-1/2}(cos\theta) + d_k sen^2\theta C_{n-k-2}^{k+3/2}(cos\theta) = 0 \end{cases}$$

De estas condiciones se deduce que existen, $2k(n-k)+2$ puntos que anulan el gradiente.

1. Distribución de puntos en la esfera.
   - Armónicos Esféricos.
   - Cálculo del gradiente.
   - Integración numérica.

Integración de puntos dispersos.

Supongamos que tenemos N nodos, $P = \{\eta_1, ..., \eta_N\}$ y sus valores aproximados $f_i \approx f(\eta_i)$. Queremos aproximar la integral

$$I(f) = \int_{\mathbb{S}^2} f(\eta) dS^2(\eta)$$

.

Proposición

Sea $T_N = \{\triangle_1, ..., \triangle_{M(N)}\}$ la triangulación de $\mathbb{S}^2$, donde los vértices de cada triángulo son los nodos.

$$I(f) = \sum_{k=1}^{M} \int_{\triangle_k} f(n) dS^2(n)$$

$$\approx \sum_{k=1}^{M} \frac{1}{3}[f(n_{k,1}) + f(n_{k,2}) + f(n_{k,3})] area(\triangle_k)$$

Integración de puntos dispersos.

Proposición

$$|I(f) - I_n(f)| \leq 4\pi c_f \text{ máx } diam(\triangle) \quad \triangle \in T_N$$

La bondad de la aproximación depende de la triangulación y del conjunto de nodos elegidos. Es conocido que se obtienen buenos resultados tomando un conjunto en el que los puntos están bien distribuidos.

Competición en Kaggle.

2. Competición en Kaggle.
   - Introducción.
   - Preprocesamiento
   - Algoritmos usados.
   - Resultados obtenidos.

Descripción del problema.

Queremos detectar cuando se registran clicks en los anuncios y estos no conllevan la instalación de la app a partir de los siguientes datos.

- ip: dirección IP de click.
- app: id de la aplicación
- device: identificación del tipo de dispositivo del teléfono móvil del usuario
- channel: id del canal del editor publicitario móvil
- so: id de la versión del OS del teléfono móvil del usuario
- click_time: marca de tiempo del click
- attributed_time : momento de la descarga de la aplicación
- is_attributed : el objetivo que se va a pronosticar, indica si la aplicación se descargó

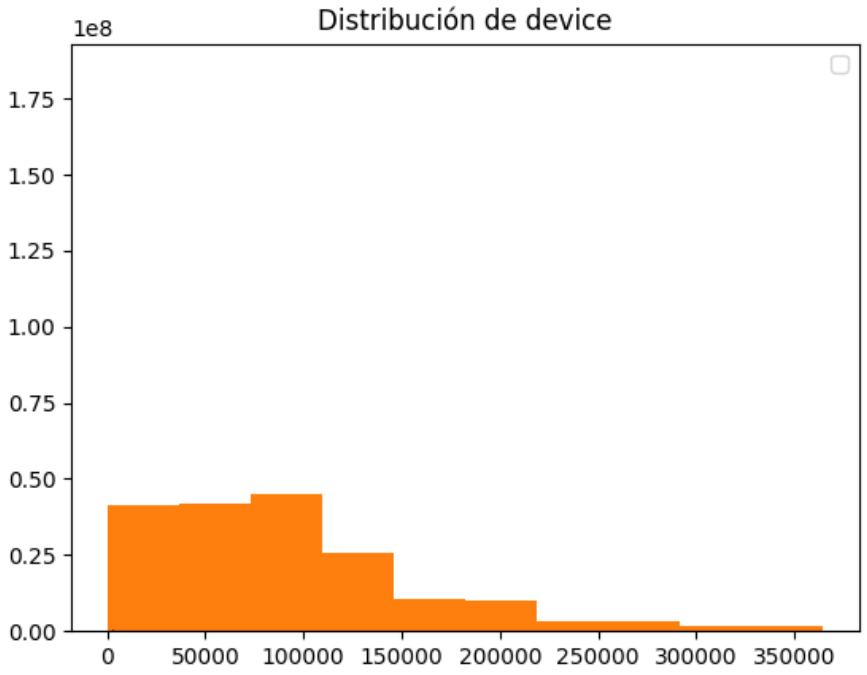
2. Competición en Kaggle.
   - Introducción.
   - Preprocesamiento
   - Algoritmos usados.
   - Resultados obtenidos.

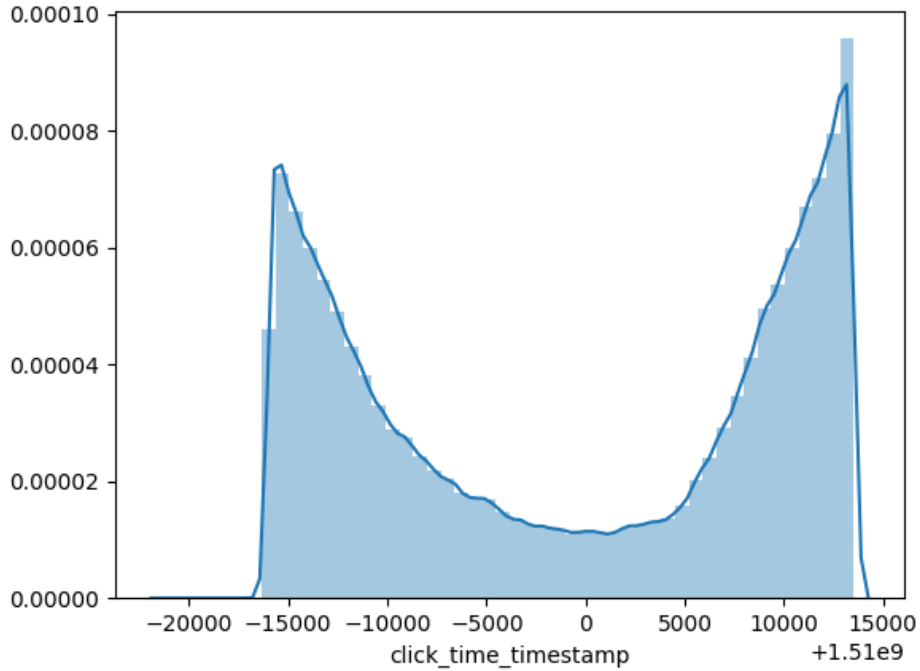
Visualización de los datos.

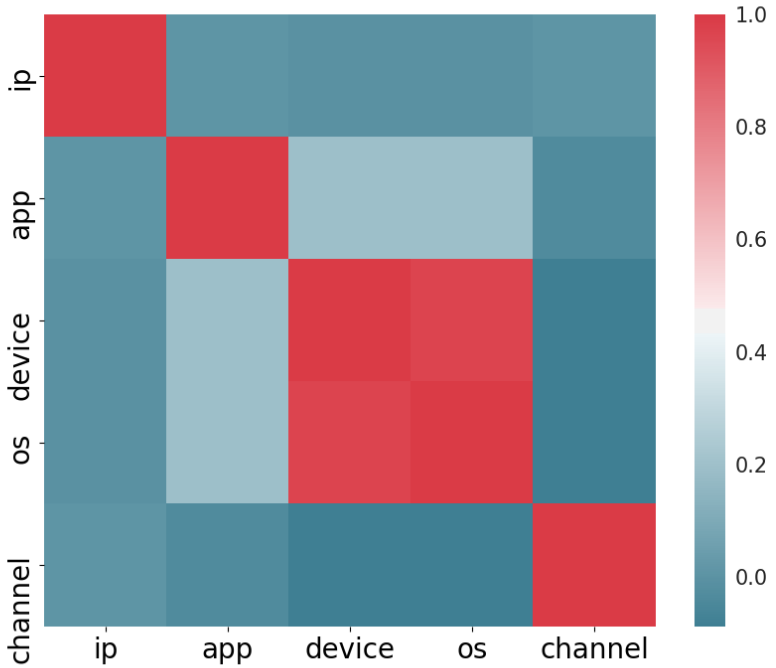
Para poder elegir una estrategia para el preprocesamiento es necesario realizar una visualización de los datos.De esta forma, podremos obtener cómo están distribuidos los valores de cada uno de los atributos o si existe alguna relación de correlación entre ellos.

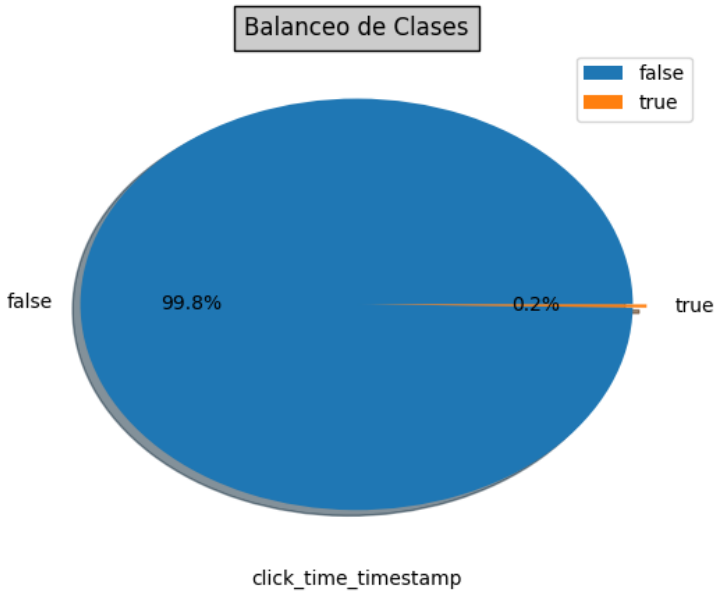
Tras realizar el estudio de los datos se obtienen las siguientes conclusiones:

- La columna *attributed_time* puede ser eliminada ya que la mayoría de sus valores son vacíos.
- De la variable *click_time* podemos obviar los datos relativos al mes y al año.
- Las variables *os* y *device* representan la misma información.
- Existe un gran desbalanceo de clases.


Distribución de device
1e8
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
50000
100000
150000
200000
250000
300000
350000


0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
−20000
−15000
−10000
−5000
0
5000
10000
15000
click_time_timestamp
+1.51e9


ip
app
device
os
channel
1.0
0.8
0.6
0.4
0.2
0.0


Balanceo de Clases
false
true
false
99.8%
0.2%
true
click_time_timestamp

Atributo	Total	%
ip	0	0
app	0	0
os	0	0
chanel	0	0
device	0	0
click_time	0	0
attributed_time	184447044	99.7529

Cuadro: Valores perdidos.

Preprocesamiento.

1. Eliminar la columna attributed_time,
2. Obtener los datos timestamp y día de la variable click_time.
3. Agrupar las variables categóricas.
   - Teniendo en cuenta el número de apariciones.
   - Usando el valor medio.

La siguiente tabla recoge los resultados obtenidos en las distintas fases del preprocesamiento.

Cambio realizado	**Resultado**
Conjunto inicial	0.8385028
Eliminar mes,año y click_time	0.8471159
Añadir día y hora	0.8471159
Cambiar día por día de la semana	0.8471159
count(channel) tras agrupar ip-app e ip-app-os	0.8289424
count(channel) tras agrupar ip-day e ip-day-hour	0.8569927
media(channel) tras agrupar ip-app e ip-app-os	0.8559159
media(channel) tras agrupar ip-day e ip-day-hour	0.8628768
Conjunto final	0.8649459

Cuadro: Pruebas realizadas durante el preprocesamiento.

2. Competición en Kaggle.
   - Introducción.
   - Preprocesamiento
   - Algoritmos usados.
   - Resultados obtenidos.

Algoritmos Usados.

Los algoritmos elegidos para construir un clasificador han sido los siguientes:

- Como punto de partida usaremos el algoritmo Boosting.
  - Se han estudiado los parámetros que lo componen y se han optimizado algunos de ellos.
  - Se ha balanceado el conjuntos de datos de entrenamiento.
- Se han estudiado algunas variantes como RUSBoosting y CUSBoosting.

La siguiente tabla recoge los resultados obtenidos.

Algoritmo	**Resultado**
Boosting base	0.9509598
Boosting con parámetros optimizados	0.9769301
RUSBoosting	0.8763027
CUSBoosting	-
Boosting con conjunto balanceado	0.9441415

Cuadro: Resultados obtenidos con los diferentes algoritmos.

2. Competición en Kaggle.
   - Introducción.
   - Preprocesamiento
   - Algoritmos usados.
   - Resultados obtenidos.

Resultados obtenidos.

- Los resultados obtenidos en esta última sección no han sido tan satisfactorios como a priori esperaba, ya que la aplicación de algoritmos alternativos no ha mejorado los resultados.
- Tras obtener los mejores parámetros del algoritmo Boosting hemos mejorado la puntuación un 2,73 %.
- Durante la fase de preprocesamiento hemos mejorado el rendimiento de nuestro clasificador un 3 % aproximadamente.