

Distribución de puntos en la esfera. Competición en Kaggle.

Daniel López García

Universidad de Granada

26 de junio de 2018

Contenidos

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

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

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.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- Algoritmos
- Resultados obtenidos.

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 $\Delta f(x) = 0$, es decir

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

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 $\mathbb{H}_n \in \mathbb{Y}_n^d$ de la siguiente forma:

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

1 Distribución de puntos en la esfera.

- Armónicos Esféricos.
- Cálculo del gradiente.
- Integración numérica.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- Algoritmos
- Resultados obtenidos.

Cálculo del gradiente.

Una base ortogonal del espacio de armónicos esféricos es

Caso particular. Esfera de dimensión 3.

Tomando coordenadas esféricas,

$$x_1 = r \sin \theta \sin \phi$$

$$x_2 = r \sin \theta \cos \phi$$

$$x_3 = r \cos \theta$$

$$0 \leq \theta \leq \pi, 0 \leq \phi \leq 2\pi, r > 0$$

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

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

La expresión de las parciales es la siguiente:

Proposición

Para $k = 0, \dots, n$

$$\partial_1 Y_{k,1}^n(x) = -\frac{(n+k)(n+k-1)}{2(2k-1)} Y_{k-1,2}^{n-1}(x) - \left(k + \frac{1}{2}\right) Y_{k+1,2}^{n-1}(x)$$

$$\partial_2 Y_{k,1}^n(x) = \frac{(n+k)(n+k-1)}{2(2k-1)} Y_{k-1,1}^{n-1}(x) - \left(k + \frac{1}{2}\right) Y_{k+1,1}^{n-1}(x)$$

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

Puntos críticos del gradiente.

Igualando las expresiones de las parciales a 0, tenemos que ha de verificarse una de las siguientes igualdades

$$\begin{cases} \sin \theta = 0 \\ \cos k\phi = 0 \\ C_{n-k-1,k+1/2}(\cos \theta) = 0 \end{cases}$$

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.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- Algoritmos
- Resultados obtenidos.

Integración de puntos dispersos.

Supongamos que tenemos N nodos, $P = \{\eta_1, \dots, \eta_N\}$ 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 $T_N = \{\triangle_1, \dots, \triangle_{M(N)}\}$ la triangulación de \mathbb{S}^2 , donde los vértices de cada triángulo son los nodos.

$$\begin{aligned} I(f) &= \sum_{k=1}^M \int_{\triangle_k} f(n) dS^2(n) \\ &\approx \sum_{k=1}^M \frac{1}{3} [f(n_{k,1}) + f(n_{k,2}) + f(n_{k,3})] \text{area}(\triangle_k) \end{aligned}$$

Integración de puntos dispersos.

Proposición

$$|I(f) - I_n(f)| \leq 4\pi c_f \max diam(\Delta) \quad \Delta \in T_N$$

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 punto están bien distribuidos.

Competición en Kaggle.

1 Distribución de puntos en la esfera.

- Armónicos Esféricos.
- Cálculo del gradiente.
- Integración numérica.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- Algoritmos
- 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

1 Distribución de puntos en la esfera.

- Armónicos Esféricos.
- Cálculo del gradiente.
- Integración numérica.

2 Competición en Kaggle.

- Introducción.
- **Preprocesamiento**
- Algoritmos
- 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.

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.

1 Distribución de puntos en la esfera.

- Armónicos Esféricos.
- Cálculo del gradiente.
- Integración numérica.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- **Algoritmos**
- 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.

1 Distribución de puntos en la esfera.

- Armónicos Esféricos.
- Cálculo del gradiente.
- Integración numérica.

2 Competición en Kaggle.

- Introducción.
- Preprocesamiento
- Algoritmos
- 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.