

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

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

Distribución de puntos en la esfera.

Motivación

- 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.

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:

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

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.

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 puntos están bien distribuidos.

Competición en Kaggle.

Descripción del problema.

Queremos clasificar ... 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

Visualización de los datos.

Visualización de los datos.

- 1 Valores vacíos.
- 2 Distribución de los valores.
- 3 Balanceo de clases.

Preprocesamiento.

- ❶ Eliminar las columnas que no ofrecen información.
- ❷ Agrupar las variable categóricas.
 - Teniendo en cuenta el número de apariciones.
 - Usando el valor medio.

Resultados obtenidos.