باسمه تعالی

تاریخ: ۱۴۰۴/۰۸/۱۷

پروژه‌ی اول درس مبانی علم داده

(Q۱

- فقط مجاز به استفاده از کتابخانه‌های numpy، matplotlib هستید.
- از هیچ تابع آماده برای نمونه‌گیری روی کره یا تولید جهت تصادفی استفاده نکنید.

تابعی بنویسید که n نقطه را به‌صورت یکنواخت از کره‌ی واحد $S^d \subset \mathbb{R}^{d+1}$ تولید کند. از روش استاندارد «توزیع نرمال استاندارد و نرمال‌سازی» استفاده کنید:

$$x = (x_1, x_2, \ldots, x_{d+1}) \sim \mathcal{N}(0,1), \quad \hat{x} = \frac{x}{\|x\|}$$

آزمایش‌ها (برای ابعاد $d = 2, 4, 10, 100, 1000$ تکرار کنید):

(a)

- تعداد 10,000 نقطه از S^d تولید کنید.
- فاصله‌ی ژئودزیک (فاصله‌ی کمانی) هر نقطه تا استوا (صفحه‌ی $x_1 = 0$) را محاسبه کنید:

$$\text{dist}(x) = |\arcsin(x_1)|$$

- هیستوگرام این فاصله‌ها را رسم کنید و توضیح دهید با افزایش d، شکل توزیع چه تغییری می‌کند.

(b)

- با انتخاب یک بردار نرمال واحد تصادفی $n \in \mathbb{R}^{d+1}$ صفحه‌ای تعریف کنید که از مبدا می‌گذرد

$$G = \{\, x : \langle x, n \rangle = 0 \,\}$$

- فاصله‌ی ژئودزیک هر نقطه از این صفحه را محاسبه کنید:

$$\text{dist}_G(x) = |\arcsin(\langle x, n \rangle)|$$

- هیستوگرام فاصله‌ها تا استوا و تا این صفحه‌ی تصادفی را در یک نمودار رسم کنید و مقایسه نمایید.

پرسش تحلیلی: آیا شکل دو هیستوگرام با هم متفاوت است؟ توضیح دهید که چه چیزی در هندسه‌ی کره باعث می‌شود این دو توزیع مشابه یا متفاوت باشند.

(c)

- ۱۰۰۰ جفت نقطه‌ی تصادفی روی s^d تولید کنید.
- زاویه‌ی بین هر دو نقطه را با فرمول زیر محاسبه کنید:

$$\theta(x, y) = \arccos\left(\langle x, y\rangle\right)$$

- برای هر بعد، هیستوگرام زاویه‌ها را رسم کنید.

پرسش تحلیلی: با افزایش بعد ، توزیع زاویه‌ها چه تغییری می‌کند؟ میانگین زاویه‌ها به چه مقداری نزدیک می‌شود و این چه مفهومی درباره‌ی هندسه‌ی فضاهای ابعاد بالا دارد؟

(d)

یک جدول بسازید که برای هر d:

- میانگین فاصله تا استوا،
- میانگین زاویه بین نقاط،
- و مقدار $d \times \mathbb{E}[\text{distance}]$

آیا حاصل‌ضرب $d \times \mathbb{E}[\text{distance}]$ تقریبا ثابت می‌ماند؟
توضیح دهید این موضوع از چه خاصیتی از فضای با بعد بالا ناشی می‌شود

باسمه تعالی

تاریخ: ۱۴۰۴/۰۸/۱۷

پروژه‌ی اول درس مبانی علم داده

(Q۲

کتابخانه‌های مجاز:

- numpy
- matplotlib
- scipy (فقط برای cKDTree)
- sklearn.random_projection (برای کاهش بُعد)

فرض کنید می‌خواهیم رفتار نقاط تصادفی را در فضاهای Nـبعدی مطالعه کنیم. یک مکعب Nـبعدی واحد با مختصات $[-0.5, 0.5]^N$ در نظر بگیرید و درون آن نقاط تصادفی تولید کنید. هدف، مشاهده و تحلیل اثرات «نفرین بُعد» (Curse of Dimensionality) است.

(a) برای ابعاد $N = [2, 3, 5, 10, 20]$ و تعداد نقاط 1000:

- نقاط را به‌صورت تصادفی یکنواخت در مکعب واحد $[-0.5, 0.5]^N$ تولید کنید.

(b) برای هر بعد N:

- حجم تقریبی کرهٔ Nـبعدی با شعاع 0.5 را با روش Monte Carlo محاسبه کنید.
- میانگین فاصله بین همهٔ نقاط و میانگین فاصله به نزدیک‌ترین همسایهٔ هر نقطه را محاسبه کنید.

(c) نمودارهای زیر را رسم کنید:

- پراکنده سازی ۲ بعد اول نقاط (scatter plot)
- حجم کره در ابعاد مختلف
- میانگین فاصله بین نقاط
- میانگین فاصله به نزدیک‌ترین همسایه

پرسش تحلیلی:

- با افزایش N ، چه اتفاقی برای حجم کره می‌افتد و چرا؟
- چگونه پراکندگی نقاط در فضا تغییر می‌کند؟ آیا نقاط به مرکز نزدیک هستند یا به گوشه‌ها؟
- میانگین فاصله بین نقاط و میانگین فاصله به نزدیک‌ترین همسایه چه روندی دارند؟ چه نتیجه‌ای می‌توان گرفت؟
- یک تحلیل شهودی ارائه دهید که چرا با افزایش ابعاد، فضا خالی‌تر می‌شود حتی اگر محدوده مختصات کوچک باشد.

(d) اکنون فرض کنید داده‌های 20ـ بعدی خود را با استفاده از Random Projection به فضاهایی با بُعد کمتر کاهش می‌دهید. بُعدهای جدید: $[2, 5, 10]$

- داده‌های ۲۰ بعدی را به بُعدهای جدید تصویر کنید.
- مجدداً میانگین فاصله‌ها و میانگین فاصلهٔ نزدیک‌ترین همسایه را محاسبه کنید.
- نتایج را با حالت اصلی (بدون کاهش بُعد) مقایسه کنید.
- سپس نموداری از تغییر میانگین فاصله‌ها قبل و بعد از کاهش بُعد رسم کرده و توضیح دهید: آیا کاهش بُعد باعث کاهش اثر نفرین بُعد شد؟

باسمه تعالی

تاریخ: ۱۴۰۴/۰۸/۱۷

پروژه‌ی اول درس مبانی علم داده

(Q۳

دیتاست: از scikit-learn مجموعه‌دادهٔ 20 Newsgroups را با گزینهٔ subset='all' دریافت کنید (بدون تفکیک train/test).
بردارسازی متن: با TfidfVectorizer متن‌ها را به بردار تبدیل کنید. تنظیمات پیشنهادی:

- max_features = 20000
- dtype = float32
- نرمال‌سازی L_2 (پیش‌فرض norm='l2' کافی است)

نمادگذاری: ماتریس ویژگی‌ها را با $X \in \mathbb{R}^{n \times d}$ نمایش دهید که n تعداد سندها و $d = 20000$ تعداد ویژگی‌هاست.
تکرارپذیری: هر جا پارامتر تصادفی وجود دارد از random_state = 0 استفاده کنید.
محاسبهٔ فاصله: برای هر دو فضا (اصلی و تصویرشده)، فاصلهٔ اقلیدسی را مبنا قرار دهید و نسبت زیر را محاسبه کنید:

$$\frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2}$$

(a) با استفاده از تابع sklearn.random_projection.johnson_lindenstrauss_min_dim و پارامترهای

$$n \in \{2000,\ 5000,\ 10000,\ 15000\}, \quad \varepsilon \in \{0.1,\ 0.2,\ 0.3,\ 0.5\}$$

بعد کمینه‌ی m را محاسبه کنید
خروجی‌ها:

- جدول 4×4: سطرها = مقادیر n، ستون‌ها = مقادیر ε، در هر سلول مقدار m درج شود.
- نمودار: m بر حسب $\log_{10} n$ (برای هر ε یک منحنی رسم شود).

پرسش تحلیلی: روند m نسبت به n و ε را توضیح دهید و بگویید این با رابطه‌ی $O\left(\frac{\log n}{\varepsilon^2}\right)$ چگونه هم‌خوان است؟

(b) با استفاده از تابع sklearn.random_projection.GaussianRandomProjection برای

$$m \in \{50,\ 100,\ 200,\ 500,\ 1000,\ 2000,\ 5000\}$$

و

$$\varepsilon \in \{0.02,\ 0.05,\ 0.10\}$$

و۵۰۰۰۰زوج تصادفی، مراحل زیر را انجام دهید:

- برای هر مقدار m:
  نگاشت (GaussianRandomProjection(n_components=m, random_state=0)) را بسازید و ماتریس ویژگی X را به X_m تصویر کنید.

- از میان نمونه‌ها، ۵۰۰۰۰ زوج تصادفی متمایز (i, j) انتخاب کنید با شرط $i \neq j$. زوج‌هایی که فاصلهٔ اصلی‌شان صفر باشد ($\|x_i - x_j\|_2 = 0$) حذف شوند.

- برای هر زوج انتخاب‌شده، نسبت فاصله‌ها را محاسبه کنید:

$$\frac{\|f(x_i) - f(x_j)\|_2}{\|x_i - x_j\|_2},$$

- برای هر m:
میانگین (mean)، میانه (median)، صدک‌های ۵ و ۹۵ (p5, p95) نسبت‌ها را محاسبه کنیدو درصد زوج‌های داخل $[1 \pm \varepsilon]$, برای $\varepsilon \in \{0.02, 0.05, 0.10\}$ قرار دارند را گزارش دهید.