تمرین اول درس هوش محاسباتی

استاد: دکتر حسین کارشناس

دستیاران آموزشی:

رضا برزگر

علی شاه‌زمانی

آرمان خلیلی

نیمسال دوم 1403-1404

*** این تمرین در سه بخش مبانی و مفاهیم الگوریتم ژنتیک، درک و حل مسائل با الگوریتم ژنتیک، و پیاده سازی، ارزیابی و تجزیه تحلیل الگوریتم ژنتیک جهت انتخاب بهترین ویژگی برای مسئله واقعی دسته بندی مشتریان طراحی شده است.**

**** زمان در نظر گرفته برای حل تمرین 13 روز (تاریخ دقیق ارسال تمرین تا آخر 30 اسفند 1403 روز پنجشنبه) می‌باشد.**

**** طبق برنامه‌ریزی فعلی ارائه تمرینات بصورت حضوری بوده و تاریخ آن متعاقبا اعلام خواهد شد. لذا لازم است تمامی اعضای گروه بر همه بخش‌های تمرین تسلط کامل داشته باشند.**

1. بخش اول: مبانی و مفاهیم الگوریتم ژنتیک (GA) و ویژگی‌های آن

1.1. الگوریتم‌های تکاملی را توضیح دهید. دلایل اصلی برتری الگوریتم‌های تکاملی نسبت به الگوریتم‌های یادگیری تقویتی در برخی مسائل چیست؟

راهنمایی: برای پاسخ به این سوال میتوانید به مقاله ای که «Open AI» با عنوان «استراتژی‌های تکاملی، یک روش جایگزین مقیاس‌پذیر برای یادگیری تقویتی»[1] منتشر کرد، مراجعه کنید.

2.1. الگوریتم ژنتیک چیست و چه تفاوتی با سایر الگوریتم‌های تکاملی مانند برنامه‌نویسی تکاملی(EP)یا استراتژی‌های تکامل(ES) دارد؟

3.1. عملیات جهش (**Mutation**) و ترکیب (**Crossover**) معمولاً چگونه روی رشته‌های بیتی (**Bitstring**) در یک الگوریتم ژنتیک اعمال می‌شوند؟ این عملگرها هنگام استفاده برای جایگشت‌ها یا سایر نمایش‌های غیر باینری چگونه باید تغییر یابند؟

4.1. چه خواص پایداری یا خصوصیات تغییرناپذیری (**Invariance Properties**) باید هنگام اجرای یک الگوریتم ژنتیک روی رشته‌های بیتی حفظ شوند؟ نمونه‌هایی ارائه دهید که نشان دهند این ویژگی‌ها چگونه بر کارایی و رفتار الگوریتم تأثیر می‌گذارند.

5.1. اگر یک الگوریتم ژنتیک برای رشته‌های بیتی با طول n برای یافتن جواب بهینه، **زمان مورد انتظار $O(n^3)$ را صرف کند**، این مقدار **چگونه با تعداد مراحل مورد انتظار برای جستجوی تصادفی (با توزیع یکنواخت) مقایسه می‌شود؟** علاوه بر این، **چه عواملی می‌توانند بر عملکرد الگوریتم در عمل تأثیر بگذارند؟**

1: Evolution Strategies as a Scalable Alternative to Reinforcement Learning

2. بخش دوم: درک و حل مسائل با الگوریتم ژنتیک

1.2. مسئله فروشنده دوره گرد (**TSP**) را در نظر بگیرید، که هدف آن تعیین کوتاه‌ترین مسیر ممکن است که از هر مجموعه شهر داده شده دقیقاً یک بار بازدید کرده و به شهر مبدأ بازگردد. برای حل این مسئله با بکارگیری یک الگوریتم ژنتیک(**GA**) ، فرض کنید که هر ژن در یک کروموزوم نشان‌دهنده‌ی یک یال بدون جهت بین دو شهر است. به عنوان مثال، ژن 'TI' نشان‌دهنده‌ی یک اتصال مستقیم بین تهران و اصفهان است، و با توجه به فرض بدون‌جهت بودن، 'TI' معادل 'IT' در نظر گرفته می‌شود.

الف) اگر تعداد کل شهرها ۱۰ باشد، هر یک کروموزوم به چند ژن نیاز دارد؟

ب) الفبای الگوریتم (مجموعه ژن‌های منحصربه‌فرد) شامل چند ژن یکتاست؟

2.2. فرض کنید یک الگوریتم ژنتیک از کروموزوم‌هایی به شکل $x = abcdefghx$ با طول ثابت هشت ژن استفاده می‌کند. هر ژن می‌تواند هر عددی بین 0 تا 9 باشد. مقدار برازش (**Fitness**) یک فرد/کروموزم x به‌صورت زیر محاسبه می‌شود:

$$f(x) = (a + b) - (c + d) + (e + f) - (g + h)$$

و فرض کنید که جمعیت اولیه شامل چهار فرد با کروموزوم‌های زیر باشد:

$$x1 = 65413532$$

$$x2 = 87126601$$

$$x3 = 23921285$$

$$x4 = 41852094$$

الف) برازندگی (Fitness) هر فرد/کروموزم را با نشان دادن تمام مراحل محاسبه کنید، و آن‌ها را به ترتیب از بیشترین مقدار برازش تا کمترین مرتب کنید.

ب) عملیات ترکیب (Crossover) زیر را انجام دهید:

- **دو فرد با بالاترین مقدار برازش** را با استفاده از **ترکیب تک‌نقطه‌ای (One-point crossover)** در **نقطه‌ی میانی** ترکیب کنید.
- **دومین و سومین فرد برتر** را با استفاده از **ترکیب دونقطه‌ای (Two-point crossover)** در **نقاط *b* و *f*** ترکیب کنید.
- **فرد اول و سوم برتر** را با استفاده از **ترکیب یکنواخت (Uniform Crossover)** ترکیب کنید.

ج) فرض کنید جمعیت جدید شامل **شش فرد حاصل از عملیات ترکیب** در سوال قبل باشد. مقدار برازش این جمعیت جدید را محاسبه کنید و تمامی مراحل محاسبات را نشان دهید. آیا مقدار برازش کلی بهبود یافته است؟

د) با بررسی تابع برازش و در نظر گرفتن این که ژن‌ها فقط می‌توانند اعداد 0 تا 9 باشند، کروموزومی را بیابید که بیشترین مقدار برازش ممکن را داشته باشد (**جواب بهینه**). همچنین مقدار بیشینه‌ی برازش را محاسبه کنید.

ه) با بررسی جمعیت اولیه‌ی الگوریتم، آیا می‌توان گفت که این الگوریتم بدون استفاده از **عملگر جهش (Mutation)** می‌تواند به **بهترین راه‌حل ممکن** دست یابد؟

3.2. یک الگوریتم ژنتیکی را در نظر بگیرید که روی یک جمعیت ۱۰ نفری با ترکیب زیر اعمال شده است.

- $x = 1$: دو نمونه
- $x = 2$: سه نمونه
- $x = 3$: سه نمونه
- $x = 4$: دو نمونه

تابع برازندگی به‌صورت زیر تعریف شده است:

$$f(x) = x^3 - 4x^2 + 7$$

الف) مقادیر خام برازندگی را برای هر مقدار متمایز از x محاسبه کنید.

ب) بررسی کنید که **آیا مقدار برازندگی خام برای هر x منفی است یا نه**. در صورت وجود مقادیر منفی، **یک مقدار ثابت را به تمام مقادیر برازندگی اضافه کنید** تا احتمال‌های انتخاب غیرمنفی شوند.

پ) مجموع **کل برازندگی جمعیت** (**پس از اعمال هرگونه تغییر لازم**) را محاسبه کنید.

ت) با استفاده از روش **انتخاب چرخ رولت، احتمال انتخاب یک فرد** با $x = 1$، $x = 2$، $x = 3$ و $x = 4$ را **براساس مقادیر برازندگی (اصلاح‌شده)** تعیین کنید.

ث) فرض کنید که اکنون احتمال‌های انتخاب با استفاده از تابع برازندگی تغییریافته زیر محاسبه می‌شوند:

$$g(x) = [f(x)]^2$$

مزیت تابع برازندگی جدید چیست؟ **احتمال انتخاب هر فرد را با استفاده از $g(x)$ مجدداً محاسبه کنید.**

ج) توضیح دهید که استفاده از مقادیر برازندگی به توان دو، یعنی $g(x)$ به‌جای $f(x)$، **چگونه فشار انتخاب** (Selection Pressure) **را تحت تأثیر قرار می‌دهد. این موضوع چه تأثیری بر همگرایی و تنوع جمعیت در الگوریتم ژنتیکی خواهد داشت؟**

3. بخش سوم: پیاده سازی، ارزیابی و تجزیه تحلیل الگوریتم ژنتیک جهت انتخاب بهترین ویژگی‌ها برای بهبود مدل‌های طبقه‌بندی مشتریان فروشگاه

در این تمرین، شما باید **الگوریتم ژنتیک (GA) را از ابتدا پیاده‌سازی کنید** تا مهم‌ترین ویژگی‌ها را برای مسئله‌ی **طبقه‌بندی مشتریان یک فروشگاه** انتخاب کنید. اهداف این تمرین شامل موارد زیر است:

- طراحی و پیاده‌سازی **الگوریتم ژنتیک برای انتخاب ویژگی** با انتخاب پارامترهای کلیدی مانند اندازه‌ی جمعیت، تابع برازش، نرخ جهش و غیره.
- شناسایی **بهترین 3، 5 و 8 ویژگی** از میان ویژگی‌های مجموعه داده.
- مقایسه‌ی عملکرد یک مدل طبقه‌بندی (در این تمرین **Decision Tree Classifier**) با استفاده از هریک **3 ویژگی منتخب برتر، 5 ویژگی منتخب برتر، 8 ویژگی منتخب برتر** و **همه‌ی ویژگی‌ها.**

دستورالعمل‌های گام‌به‌گام

۱) آشنایی با مجموعه داده

ابتدا مجموعه داده را از این لینک دانلود کنید. این مجموعه داده شامل **اطلاعات مشتریان** میباشد و هدف آن **پیش‌بینی دسته‌بندی آن‌ها (A، B، C یا D)** است.

- **دسته بندی‌ها یا کلاس‌های هدف:** ستون **Segmentation** (A، B، C، D)
- **ویژگی‌ها**: تمام ستون‌ها به‌جز **ID**، **Var_1** و **Segmentation**

توضیحات کامل دیتاست در سند **Data Description** ارائه شده است.

---

۲) پیش‌پردازش داده‌ها

قبل از اعمال **الگوریتم ژنتیک برای انتخاب ویژگی‌ها**، باید مجموعه داده را پردازش کنید و مراحل زیر را انجام دهید:

- ❖ **مدیریت داده‌های از دست‌رفته:** مقداردهی به مقادیر گمشده (مثلاً در ستون Work_Experience) یا حذف ردیف‌های دارای داده‌های گمشده.
- ❖ **حذف داده های پرت (Outliers)**
- ❖ **رمزگذاری ویژگی‌های دسته‌ای (Categorical):** تبدیل ستون‌های دسته‌ای(Gender، Ever_Married، Graduated، Profession، Spending_Score) به فرمت عددی با استفاده از **Label Encoding** یا **One-Hot Encoding**.

---

۳) پیاده‌سازی الگوریتم ژنتیک (GA) برای انتخاب ویژگی‌ها

شما باید **الگوریتم ژنتیک را از ابتدا (from Scratch) کدنویسی کنید** و از کتابخانه‌های آماده‌ی GA استفاده نکنید.

1.3) تعریف اجزای الگوریتم ژنتیک

- **نمایش کروموزوم**: نوع نمایش خود را مشخص و تعریف کنید.
- **جمعیت اولیه**: مجموعه‌ای تصادفی از زیرمجموعه‌های ویژگی‌ها را به‌عنوان جمعیت اولیه تولید کنید.

2.3) تعریف تابع برازش

تابع برازش میزان عملکرد یک مجموعه‌ی ویژگی را در طبقه‌بندی Decision Tree Classifier ارزیابی می‌کند.

- شما باید **استراتژی تابع برازش را خودتان تعریف کنید** که می‌تواند بر اساس **دقت مدل طبقه‌بندی** باشد **یا ترکیبی از چندین معیار.**

3.3) استراتژی انتخاب

روشی را برای **انتخاب بهترین افراد** از جمعیت برای نسل بعدی مشخص کنید و براساس آن پیاده سازی نمایید.

- استراتژی‌های انتخابی می‌توانند شامل موارد زیر اما نه محدود به آنها باشد:
  - **انتخاب تورنمنتی(Tournament Selection)**
  - **انتخاب چرخ رولت(Roulette Wheel Selection)**
  - **انتخاب رتبه‌بندی‌شده(Rank-Based Selection)**
- **حداقل دو روش را بکار بگیرید و نتایج را مقایسه کنید.**

4.3) عملگر ترکیب(Crossover)

- یک **مکانیسم ترکیب** پیاده‌سازی کنید که در آن دو والد انتخاب‌شده فرزندان جدیدی را با ترکیب رشته‌هایشان/ژن‌شان ایجاد کنند.
- حداقل دو نوع ترکیب را پیاده سازی و نتایج را مقایسه کنید. به عنوان مثال عملگرهای ترکیب زیر:
  - **ترکیب چندنقطه‌ای(Multi-point Crossover)**
  - **ترکیب یکنواخت(Uniform Crossover)**

5.3) جهش(Mutation)

- تغییرات تصادفی کوچک در فرزندان ایجاد کنید تا تنوع حفظ شود.
- **نرخ جهش** را تعیین کنید (مثلاً **۱ تا ۵ درصد بیت‌ها به‌طور تصادفی معکوس شوند**).

6.3) معیار توقف

- الگوریتم باید در شرایط زیر متوقف شود:
  - رسیدن به **تعداد نسل‌های ثابت**.
  - **همگرایی جمعیت** (عدم بهبود قابل توجه).

---

٤) انتخاب بهترین ویژگی‌ها

GA را سه بار اجرا کنید:

1. برای انتخاب **3 ویژگی برتر**.
2. برای انتخاب **5 ویژگی برتر**.
3. برای انتخاب **8 ویژگی برتر**.

❖ زیرمجموعه‌های ویژگی‌هایی را که بالاترین امتیاز برازش را دارند، استخراج و با رسم نمودار توزیع هر ویژگی تحلیل کنید.

---

۵) آموزش و ارزیابی مدل طبقه‌بندی

پس از انتخاب **3، 5 و 8 ویژگی برتر** توسط GA :

❖ **مدل Decision Tree Classifier را برای طبقه‌بندی** با استفاده از هر یک از **3 ویژگی منتخب، 5 ویژگی منتخب، 8 ویژگی منتخب** و **همه‌ی ویژگی‌ها آموزش دهید.**

❖ پس از آموزش مدل‌ها، **عملکرد آن‌ها را** با استفاده از **معیارهای دقت** (**Accuracy**) **ارزیابی کنید**.

❖ نتایج بدست آمده هر یک به همراه نمایش در نمودار مقایسه کنید.

مواردی که باید ارسال کنید:

1. **کد پیاده‌سازی، ترجیحا پایتون (Python) در پلتفرم Google Colab** (پیاده سازی با دیگر زبان‌ها چون **MATLAB،R**، و غیره نیز قابل قبول است).
2. **نتایج استخراج ویژگی های منتخب با رسم نمودار توزیع آن‌ها. نتایج طبقه‌بندی** شامل جدول یا نموداری که معیارهای عملکرد را مقایسه کند.
3. **گزارش تحلیلی** شامل توضیح فرآیند، مقایسه‌ی نتایج و تحلیل تأثیر پارامترهایGA، پاسخ به سوالات زیر.

**** نکات کلیدی برای بحث که باید به آنها پاسخ بدهید:**

- **کدام مجموعه‌ی ویژگی بهترین عملکرد طبقه‌بندی را داشت؟**
- **آیا استفاده از GA برای انتخاب ویژگی باعث بهبود مدل شد، یا عملکرد مشابهی با همه‌ی ویژگی‌ها داشت؟**
- **مزایا و معایب استفاده از تعداد ویژگی‌های کمتر در مقایسه با همه‌ی ویژگی‌ها چیست؟**
- **کدام پارامترهای GA (اندازه‌ی جمعیت، روش انتخاب، نرخ جهش) بیشترین تأثیر را بر عملکرد داشتند؟**

محدودیت‌های پیاده‌سازی

🚨 **از کتابخانه‌های آماده‌ی GA** مانند **DEAP یا PyGAD استفاده نکنید** .شما باید **الگوریتم ژنتیک را از ابتدا پیاده‌سازی کنید**.

✅ می‌توانید از **NumPy**، **Pandas**، **Scikit-learn** و **Matplotlib**برای پردازش داده‌ها، آموزش مدل Decision Tree Classifier، ارزیابی و نمایش نتایج استفاده کنید.