بسمه تعالی

تمرین سری چهارم　　درس پردازش تکاملی　　دکتر اثنی عشری

لطفا در حل تمارین به موارد زیر توجه داشته باشید:

- تمارین را خودتان انجام بدهید و به کاری که ارائه می‌کنید، تسلط کامل داشته باشید.
- تمارین را با هر زبان برنامه‌نویسی می‌توانید انجام دهید.
- به طور کلی، استفاده از کتابخانه‌های آماده **مجاز نیست.**
- **اسم فایل** ارسالی به صورت HW4_Name_StudentNumber باشد. همچنین اسم و شماره دانشجویی باید در فایل PDF پاسخ‌ها و داخل کدها به صورت comment وجود داشته باشد.
- فایل نهایی به صورت **فشرده شده**، باید شامل ۱ فایل PDF (پاسخ سوالات و گزارش کدها) و یک فولدر شامل همه کدها (و فایل‌های لازم) باشد. **از قرار دادن فایل‌های اضافی جدا خودداری فرمایید.**
- گزارش باید کامل و در عین حال مختصر و مفید باشد. گزارش باید شامل **توضیحات لازم** (پاسخ سوالات پرسیده شده)، نکات پیاده‌سازی، **مقادیر هایپرپارامترها** و **نتایج** اجرای کد (به همراه مقایسه نتایج با هم در صورت نیاز) باشد.
- نمره هر تمرین از ۳ بخش اصلی تشکیل می‌شود: **گزارش**، **کد** و **ارائه** تمرین. پس از تحویل هر تمرین، زمانی برای ارائه آن هماهنگ خواهد شد. تسلط شما بر موضوع و کد، بسیار مهم‌تر از صرف پاسخ دادن و پیاده‌سازی است که با پاسخ‌گویی به سوالات پرسیده شده در جلسه ارائه تمرین، سنجیده می‌شود.
- تمارین به صورت متنوع و گلچین آماده شده است تا با چالش‌های جدید و حتی مطرح نشده در کلاس روبرو شوید. هدف، افزایش قدرت تحلیل و حل مسئله شماست.
- در صورت مشاهده پاسخ‌ها و کدهای مشابه، نمره کسر می‌گردد.

سوال ۱) به سوالات زیر پاسخ دهید.

الف) تنظیم پارامتر و کنترل پارامتر را توضیح داده تفاوت ها و چالش های آن ها را بیان کنید.

ب) کدام روش نسبت به دیگری برتری دارد توضیح دهید.

ج) سه نوع اصلی کنترل پارامتر را نام ببرید و تفاوت آن‌ها را از نظر نحوه تغییر پارامترها توضیح دهید.

سوال۲) یک شرکت هوش مصنوعی قصد دارد الگوریتم جدید خود برای تشخیص چهره را با الگوریتم فعلی مقایسه کند. آزمایش‌ها به این صورت انجام شده: روی ۱۵ نفر کاربر، هر دو الگوریتم اجرا شده و نرخ موفقیت هر الگوریتم در هر نفر ثبت شده است. تیم تحقیقاتی همچنین نتایج الگوریتم خود را با یک عدد استاندارد جهانی که میانگین دقتش ٪۹۵ است، مقایسه می‌کند. علاوه بر این، تیم می‌خواهد الگوریتم جدید را با یک الگوریتم قدیمی خارجی هم مقایسه کند که روی ۱۵ تصویر متفاوت تست شده است.

الف) برای مقایسه الگوریتم جدید با الگوریتم فعلی شرکت، کدام آزمون t مناسب است؟

ب) برای مقایسه الگوریتم جدید با عدد مرجع جهانی ٪۹۵، کدام آزمون t مناسب است؟

ج) برای مقایسه الگوریتم جدید با الگوریتم خارجی که روی ۱۵ تصویر دیگر تست شده، کدام آزمون t باید استفاده شود؟

سوال ۳) به سوالات زیر پاسخ دهید.

الف) روش DPX چگونه عمل می‌کند و هدف از طراحی آن چیست؟

ب) جهش جمعی چیست و در چه مواقعی استفاده می‌شود؟

ج) مدل Boltzmann چگونه باعث تعادل بین بهره‌برداری (exploitation) و اکتشاف (exploration) در MA می‌شود؟

سوال ۴) پیاده سازی:

دیتاست تشخیص سرطان سینه ((Breast Cancer Wisconsin (Diagnostic) شامل اطلاعاتی از تومورهای خوش‌خیم و بدخیم است. هر نمونه دارای ویژگی‌های عددی است که خصوصیات فیزیکی تومورها را توصیف می‌کند (مانند اندازه، بافت، یکنواختی و...). هدف از این تمرین، یافتن زیرمجموعه‌ای از ویژگی‌هاست که

با استفاده از آن‌ها دقت طبقه‌بندی مدل SVM افزایش یابد و در عین حال، تعداد ویژگی‌های انتخاب‌شده تا حد ممکن کم باشد. برای دستیابی به این هدف، یک تابع برازندگی به صورت زیر تعریف شده است:

$$fitness = accuracy - \lambda \times \frac{number\ of\ selected\ features}{total\ number\ of\ features}$$

که در آن دقت (Accuracy) به‌دست آمده از مدل SVM با استفاده از اعتبارسنجی متقابل است.

λ ضریب جریمه است که بر اساس تعداد ویژگی‌های انتخاب‌شده نسبت به کل ویژگی‌ها، جریمه‌ای اعمال می‌کند. این ضریب را ۰.۰۱ در نظر بگیرید.

این تابع سعی می کند یک توازن بین دقت مدل و تعداد ویژگی‌های انتخاب‌شده برقرار می‌کند.

در این تمرین، از شما خواسته شده است که دو الگوریتم ژنتیک و ممتیک را برای انتخاب ویژگی‌ها به‌کار گیرید و آن‌ها را به‌طور مستقل پیاده‌سازی کنید. مراحل زیر را انجام دهید:

الف) برای هر دو الگوریتم، مقادیر مناسب پارامترها را انتخاب و تنظیم کنید. این پارامترها شامل اندازه جمعیت، تعداد نسل‌ها ، نرخ جهش برای الگوریتم ژنتیک و به علاوه تعداد تکرارهای جستجوی محلی برای الگوریتم ممتیک هستند. شما می‌توانید این پارامترها را به‌صورت تجربی انتخاب کنید و با توجه به محدودیت‌های زمانی و منابع، مقادیر مناسب را آزمایش کنید. مقادیر انتخابی برای پارامترها را در گزارش خود ذکر کنید.

ب) پس از بهینه‌سازی و اجرای هر دو الگوریتم، با استفاده از آزمون آماریt-test ، نتایج دو الگوریتم را مقایسه کنید. (دقت شود که برای داشتن مقایسه معنادار، هر الگوریتم حداقل ۳۰ بار اجرا شود تا نتایج آماری معتبر و قابل استنادی بدست آید.) در ادامه، فارغ از معنادار بودن یا نبودن تفاوت، تحلیل کنید که چرا یکی از الگوریتم‌ها بهتر عمل کرده یا چرا عملکرد آن‌ها مشابه بوده است.

ج) برای هر الگوریتم، ویژگی‌های انتخابی و تعداد آن‌ها را مشخص کنید و مقادیر بهینه پارامترها را که برای بهینه‌سازی الگوریتم‌ها به‌دست آمده‌اند، ذکر کنید.

سوال ۵) پیاده سازی:

داده‌های مورد استفاده در این تمرین، مجموعه‌داده‌ی Cleveland Heart Disease از پایگاه داده‌ی UCI هستند که شامل ویژگی‌های عددی مرتبط با وضعیت سلامتی قلبی بیماران است. برچسب هر نمونه نشان‌دهنده‌ی وجود یا عدم وجود بیماری قلبی در فرد است. برای سادگی، این برچسب‌ها به یک مسئله‌ی طبقه‌بندی دودویی تبدیل شده‌اند .هدف از این تمرین، بهینه‌سازی پارامترهای SVM شامل Cو γ با استفاده از الگوریتم‌های تکاملی است. برای ارزیابی عملکرد، از معیار دقت (Accuracy) استفاده می‌شود که با استفاده از اعتبارسنجی متقابل پنج‌تایی (5-Fold Cross Validation) محاسبه می‌شود. تابع برازندگی برابر با دقت مدل SVM خواهد بود.

برای این مساله الگوریتم های زیر را پیاده سازی کنید:

الف) الگوریتم ممتیک با رویکرد Lamarckian

ب) الگوریتم ممتیک با رویکرد Baldvinian

ج) الگوریتم جستجوی محلی

د) نتایج به‌دست‌آمده از الگوریتم‌ها را با یکدیگر مقایسه کرده و تحلیل کنید که کدام‌یک عملکرد بهتری داشته و چرا، یا اینکه چرا عملکرد آن‌ها مشابه بوده است.

توجه:

-هر یک از الگوریتم‌ها را 10بار اجرا کرده و میانگین دقت و پارامترهای بهینه‌ی حاصل را گزارش دهید.

-پارامترهای مورد نیاز الگوریتم‌ها (مانند اندازه جمعیت، نرخ جهش، تعداد نسل‌ها، و تعداد تکرار جستجوی محلی در الگوریتم ممتیک) را به‌صورت تجربی و متناسب با منابع در دسترس انتخاب کرده و در گزارش خود ثبت نمایید.

موفق باشید