بسم الله الرحمن الرحیم

برنامه‌سازی پیشرفته

پروژه

دانشکده ریاضی و علوم کامپیوتر

دانشگاه علم و صنعت ایران

نیمسال دوم ۰۴-۰۳

Decision Tree

استاد:

دکتر کارلو آبنوسیان

مهلت ارسال:

پایان روز ۱۰ تیر ۱۴۰۴

نکته‌های قابل توجه

- پروژه به زبان پایتون و در قالب **۱** نفره انجام می‌شود.
- حداکثر مهلت تحویل پروژه پایان روز **۱۰ تیر ۱۴۰۴** می‌باشد. اما در صورت زودتر به اتمام رساندن پروژه امکان ارائه دادن زودتر از موعد مشخص شده وجود دارد.
- استفاده از ChatGPT و دیگر ابزار های هوش مصنوعی برای پیاده سازی کد در صورتی مجاز است که فرد توانایی درک و توضیح بخش های مختلف کد را داشته باشد اما برای نوشتن داکیومنت و توضیحات استفاده از ابزار های هوش مصنوعی **مجاز نیست**.

پروژه

هدف از این پروژه، طراحی و پیاده‌سازی یک مدل درخت تصمیم برای تحلیل و پیش‌بینی نتایج بر اساس داده‌های ورودی است. درخت تصمیم یکی از الگوریتم‌های پرکاربرد در حوزه یادگیری ماشین طبقه‌بندی (Classification) است که با استفاده از ساختار درختی، فرآیند تصمیم‌گیری را به صورت گام به گام و قابل فهم انجام می‌دهد.

در این پروژه تلاش شده است تا با استفاده از الگوریتم درخت تصمیم، در مرحله ی آموزش مدل درخت ساخته شده و بر اساس آن، پیش‌بینی‌هایی با دقت بالاتر و قابل تفسیر انجام شود. یکی از مزایای کلیدی این روش، سادگی در تفسیر خروجی‌ها و امکان استخراج قوانین تصمیم‌گیری از مدل نهایی است.

مراحل اجرایی پروژه

انتخاب و آماده‌سازی دیتاست

- دیتاست باید شامل حداقل **۱۰,۰۰۰ دیتاسمپل** و **۲۰ ویژگی (Feature)** باشد.
- توجه داشته باشید که دیتاست انتخاب‌شده، برای مدل‌سازی با درخت تصمیم مناسب باشد.

پیش‌پردازش و تقسیم داده‌ها

- ابتدا داده‌ها را به دو بخش تقسیم کنید:
- ۸۰٪ برای **آموزش (Training)**
- ۲۰٪ برای **آزمون (Testing)**
- می توانید از تکنیک hold-out validation با کمک مجموعه‌ی اعتبارسنجی (Validation) که در ورکشاپ توضیح داده شده است جهت تنظیم (Tune) هایپرپارامترها استفاده کنید.

گسسته‌سازی ویژگی‌های پیوسته (Discretization)

- برای کاهش پیچیدگی محاسبات مدل و بهبود عملکرد الگوریتم، از روش **چارک‌بندی (Quartile Binning)** جهت گسسته‌سازی ویژگی‌های پیوسته یا ویژگی‌هایی با دامنه مقادیر بسیار متنوع استفاده کنید. شرح چارک‌بندی به صورت زیر است:

  این روش داده‌های پیوسته را به چهار گروه تقسیم می‌کند:

  چارک اول (Q1): از پایین‌ترین مقادیر 25%

  چارک دوم (Q2) یا میانه: مقادیر پایین‌تر از 50%

  چارک سوم (Q3) : از پایین‌ترین مقادیر 75%

  چارک چهارم: شامل مقادیر بالاتر از Q3

پ.ن: می‌توانید از ایده‌های دیگر برای گسسته‌سازی نیز استفاده کرده، آن‌ها را پیاده‌سازی و با چارک‌بندی مقایسه کنید.

پیاده‌سازی درخت تصمیم

- پیاده‌سازی باید **کاملاً بدون استفاده از توابع آماده درخت تصمیم** انجام شود.
- اجزای اصلی که باید توسط شما پیاده‌سازی شود:

محاسبه **information gain**

محاسبه **آنتروپی**

محاسبه **Gini Index**

الگوریتم بازگشتی ساخت درخت

پس هرس (Pruning)

پ.ن: استفاده از توابع آماده تنها برای موارد عمومی مانند خواندن فایل، نمایش درخت یا ترسیم گرافیکی مجاز است.

خروجی‌های مورد انتظار پروژه

- کد اجرایی با توضیحات کافی جهت اجرا
- دیتاست مورد استفاده
- فایل پاسخ به سؤالات ورکشاپ (به صورت PDF یا در قالب نوت‌بوک)
- نمایش ساختار نهایی درخت تصمیم، شامل:

  ویژگی‌های تست‌شده در هر گره

  مقادیر **Information Gain** و **Gini Index** برای هر گره و شاخه

نکات امتیازی

- افزایش دقت مدل با انتخاب هوشمندانه داده‌ها
- گسسته‌سازی حرفه‌ای‌تر یا نوآورانه
- تحلیل آماری پیشرفته روی فیچرها
- ارائه نمودارها یا ابزارهای بصری برای تفسیر بهتر مدل

موفق باشید!

تیم تدریس‌یاری برنامه‌سازی پیشرفته