دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم‌های خبره

ورکشاپ اول (درخت تصمیم)

دکتر آرش عبدی

پاییز 1404

طراحان ورکشاپ: امین شیروانی ، مرتضی جوادیان

لطفا نکات تکمیلی زیر را در طول ترم در نظر بگیرید:

1. مهلت تحویل (ددلاین)

تمامی مهلت‌های تحویل به صورت سخت‌گیرانه (Hard Deadline) و در سطح دقت ثانیه تعیین می‌شوند. چنانچه پروژه‌ای حداکثر تا ساعت "00:05" روز مقرر تحویل داده نشود (صرف نظر از علت، از جمله اختلالات اینترنتی یا سایر موارد)، آن پروژه قابل پذیرش نبوده و دانشجو می‌بایست تمرکز خود را بر پروژه‌های بعدی معطوف نماید.

2. مجوز تأخیر:

در طول ترم، حداکثر ۵ روز تأخیر مجاز در مجموع پروژه‌ها قابل استفاده است. در صورت بهره‌برداری از کل این ظرفیت در پروژه اول، هیچ گونه تأخیری برای پروژه‌های بعدی مجاز نخواهد بود. این مدت قابلیت تمدید (Extension) ندارد.

3. تعیین ددلاین‌ها:

مهلت‌های تحویل، با هماهنگی جمعی دانشجویان حاضر در کلاس و در اولین جلسه پس از انتشار (پروژه/تمرین) تعیین می‌گردند و پس از تصویب، تغییرناپذیر هستند.

4. ارائه پروژه:

از کلیه پروژه‌ها ارائه شفاهی اخذ خواهد شد. تاریخ‌های ارائه در بازه‌ای متشکل از ۲ تا ۳ روز متوالی تنظیم می‌شوند. عدم ارائه به منزله دریافت نمره صفر در آن پروژه خواهد بود.

5. شیوه ارائه:

ارائه‌ها به صورت حضوری یا مجازی (با انتخاب دانشجو) قابل اجرا هستند. مدت زمان هر ارائه بین ۲۰ تا ۳۰ دقیقه متغیر است.

6. مستندسازی (داکیومنت‌نویسی):

ارائه مستندات کامل برای هر پروژه الزامی است. عدم رعایت ضوابط مستندسازی موجب کسر نمره به شرح زیر خواهد شد:

- عدم ارسال مستند: کسر ۳۰ نمره از ۱۰۰
- استفاده از مستند تولیدشده توسط چت‌بات‌ها: کسر ۵۰ نمره از ۱۰۰
- مستند بسیار ناقص: کسر ۲۰ نمره
- نواقص جزئی: کسر ۵ تا ۱۰ نمره

7. ضوابط مستندسازی:

مستندات می‌بایست موارد زیر را پوشش دهند:

- توضیح بلاک‌های کد مورد استفاده در پروژه با ذکر نقش و عملکرد آن‌ها (بدون نیاز به توضیح خط‌به‌خط)

- ضمیمه کردن اسکرین‌شات از هر بخش کد الزامی است.

- استفاده از فونت B Nazanin سایز ۱۴ و رعایت کامل اصول نگارشی.

- تشریح مفاهیم به کار رفته در پروژه که در کلاس یا تمرینات توضیح داده نشده‌اند.

- تهیه مستند در قالب Word یا LaTeX و تحویل نهایی به صورت PDF (تحویل دست‌نویس مجاز نیست).

- ذکر منبع در موارد استفاده از منابع خارجی (برای مثال ChatGPT) الزامی است. این امر موجب کسر نمره نخواهد شد.

8. پاسخ‌های تئوری:

پاسخ سوالات تئوری می‌بایست تایپ‌شده و در قالب PDF تحویل داده شوند. تحویل پاسخ‌های دست‌نویس مجاز نمی‌باشد.

9. مطرح کردن پرسش‌ها:

کلیه پرسش‌های مربوط به تمرین‌ها می‌بایست در گروه درس مطرح شوند تا سایر دانشجویان نیز از پاسخ بهره‌مند گردند. از ارسال پیام‌های خصوصی (Pv) خودداری فرمایید.

10. رفع مشکلات درسی:

در صورت بروز هرگونه مشکل در طول ترم، فقط با Head TAهای درس تماس حاصل فرمایید:

آقای امیر حسین حسینی جبلی ([@Amir_Jebbeli](https://t.me/Amir_Jebbeli))

آقای امیرعلی دستوری ([@amirali_dst](https://t.me/amirali_dst))

تاریخ ریلیز پروژه: شنبه 19 مهرماه ساعت 20

تاریخ تحویل پروژه: دانشجویان مشخص می‌کنند

آیدی طراحان در تلگرام : @aminkte و @mor1383teza

پیاده سازی درخت تصمیم:

در این بخش، هدف شما پیاده‌سازی کامل الگوریتم درخت تصمیم (Decision Tree) و تحلیل نتایج آن است. فایل DT_Library.py شامل کلاس های DecisionTree و Node است که متدهای ضروری را شامل می‌شوند که باید تکمیل شوند.

در ادامه در فایل نوت‌بوک در پارت های اول تا پنجم برای متدها و کلاس‌هایی که نوشتید؛ تست‌هایی طراحی شده است که اطمینان حاصل کنید که پاس بشوند. همچنین باید بخش های آموزش مدل را تکمیل کنید(به کمک دیتاست داده شده و متدهایی که نوشتید)

شما باید مراحل زیر را به دقت انجام دهید:

* **تکمیل توابع اصلی و کلاس‌ها و متدها** (**در فایل** DT_Library)

- کلاس ها شامل تعدادی متد ناقص است که شما باید آن‌ها را بر اساس تئوری الگوریتم ID3 یا CART کامل کنید.

* **اجرای تست‌ها** + **تکمیل قسمتی دیگر از سلول‌ها** (**در فایل** NoteBook)

- در پارت های 1 تا 5 چند تست در نوت‌بوک طراحی شده است.
- شما باید اطمینان حاصل کنید که تمامی این تست‌ها پاس شوند.
- این مرحله باعث می‌شود مطمئن شوید پیاده‌سازی اولیه شما صحیح است و آماده‌ی مرحله‌ی بعد می‌باشد.

*** تیون کردن هایپرپارامترها (Hyperparameter Tuning)**

- برای بهبود عملکرد مدل درخت تصمیم و جلوگیری از Overfitting یا Underfitting، لازم است هایپرپارامترهای اصلی درخت را تنظیم کنیم. این مرحله شامل بررسی ترکیب‌های مختلف هایپرپارامترها و انتخاب بهترین حالت بر اساس معیار دقت مدل (Accuracy) است *(اگر علاقه مند بودید، متریک‌های Recall, precision, f1-score را نیز بررسی کنید.)* از جمله هایپرپارامترهای مهم می‌توان به max_depth , min_sample_split اشاره کرد.برای انجام این کار از روش هایی مثل GridSearch و RandomizedSearch روی دیتای Validation که با روش hold out ایجاد شده اشاره کرد.

- نکته: درواقع با هایپرپارامتر تیونینگ قبل از ساخت کامل درخت، با محدود کردن این هایپرپارامترها از رشد بیش از حددرخت جلوگیری می‌کنیم. نتیجه‌ی این مرحله کاهش احتمال Overfitting و افزایش قابلیت تعمیم مدل می‌شود.

*** پیش‌پردازش یا Pre-Processing**

در پارت 7 نوت بوک باید دیتاست خام EuroRail_Survey را برای مدل‌سازی آماده کنید. این کار شامل مجموعه‌ای از مراحل زیر است که می‌توانید با روش‌های مناسب انجام دهید:

- مدیریت رکورد های ناقص (Handling Missing Values)
- مدیریت رکورد های تکراری (Duplicate Records)
- تشخیص و حذف داده‌های پرت و نویزی (Outlier Detection) –-Outlier نتیجه‌ی نویز شدید است--
- استانداردسازی ویژگی‌ها (Feature Scaling) با روش هایی مثل MinMaxScaler یا StandardScaler
- انتخاب ویژگی‌های مؤثر(Feature Selection) با روش هایی مثل Heatmap یا Correlation Matrix
- دسته‌بندی ویژگی‌های عددی گسترده (Binning)
- کدگذاری ویژگی‌های متنی (Encoding Categorical Data) با روش هایی مثل OneHotEncoder یا LabelEncoder

- نکته مهم: در این بخش تمرکز بر درک فرآیندها و انتخاب روش مناسب است. هر گروه می‌تواند با توجه به تحلیل داده‌ی خود، از روش‌های متفاوت استفاده کنند.استفاده از مراحل بالا اختیاری است و لزومی بر به کار بردن همگی نیست.

*** آموزش و ارزیابی مدل (Training & Evaluation)**

داده‌های پیش‌پردازش شده را به بخش‌های Train و Test تقسیم کنید. مدل را روی داده‌های آموزش train کنید و روی داده‌های تست آموزش کنید.

نتایج اولیه را بررسی کنید و مشاهده کنید که آیا مدل Overfitting یا Underfitting دارد. در صورت بدست آوردن دقت بالای 90 درصد در دیتای Test نمره امتیازی به شما تعلق می گیرد.

*** پس-هرس(Post-Pruning)**

پس از ساخت کامل درخت، گره‌هایی که به دقت مدل کمک چندانی نمی‌کنند را حذف کنید .این کار باعث می‌شود درخت ساده‌تر و قابل فهم‌تر شود و همچنان دقت بالایی داشته باشد.

*** رسم درخت تصمیم (Tree Visualization)**

درخت تصمیم خود را با استفاده از یکی از کتابخانه‌های پایتون مانند:

1. Network
2. graphviz

به‌صورت ویژوال و گرافیکی ترسیم کنید.

با اجرای کامل این مراحل، شما یک درخت تصمیم آماده، پاکسازی‌شده و بهینه خواهید داشت که قادر است داده‌های پیش‌پردازش‌شده را به‌درستی دسته‌بندی کرده و پیش‌بینی رضایت مشتریان (Satisfaction) را انجام دهد.

در مرحله‌ی بعدی می‌توانید اهمیت ویژگی‌ها (Feature Importance) را تحلیل کنید و ببینید کدام ویژگی‌ها بیشترین تأثیر را بر تصمیم‌گیری دارند شبیه کاری که در مقدمه در Feature Selection کردید.

- نکته: خلاقیت شما برای افزایش دقت درخت مثل افزایش داده‌های آموزشی یا هرگونه انتخاب هوشمندانه از میان آنها ، روش‌های جدیدتر و حرفه‌ای تر گسسته‌سازی و یا حتی فعالیت‌های اضافه تر حرفه‌ای مانند تحلیل‌های آماری جداگانه از فیچرها و ... می‌تواند نمره امتیازی داشته باشد.

آنچه باید در نهایت تحویل دهید:

1. کد اجرای برنامه با توضیحات لازم برای اجرا
2. دیتاست‌نهایی
3. درخت را به شکل visualized آنطور که در فایل Notebook گفته شده است، در فرمت مناسب نمایش دهید.
4. در داکیومنت خود علاوه بر موارد گفته شده در Policy درس؛ گزارشی کامل از مسیر انجام کار و چالشهایی که با آن مواجه شدید، همچنین اجراهای گرفته شده و روند پیشرفت پروژه و به‌علاوه توضیحاتی در مورد دقت خود در داده های تست ارائه دهید!

   آیا با overfit مواجه شدید؟ چه روش هایی را برای حل این مشکل در پیش گرفتید؟ (این بخش نمره ی قابل توجهی دارد پس از چت بات استفاده نکنید و خودتون بنویسید)
5. هرگونه تحلیل اضافه مفید و خلاقیت می تواند نمره امتیازی داشته باشد.
6. نکته بسیار مهم: 5 مورد بالا را zip کرده و نام آن را بصورت DT_STudentID.zip خود در LMS قرار دهید.