

به نام خدا



دانشکده مهندسی کامپیوتر

هوش مصنوعی و سیستم های خبره

پروژه اول – درخت تصمیم

دکتر آرش عبدی

پاییز 1403

طراحان :

محمدصادق نعمت‌پور

نیایش خانی



- در صورت وجود هر گونه ابهام در سوالات تنها به طراح آن سوال پیام دهید.
- با توجه به تنظیم شدن ددلاین تمارین توسط خود شما امکان تمدید وجود ندارد.
- زبان برنامه‌نویسی و قالب تمپلیت پایتون است ولی برای تمرین‌های اول می‌توانید از C# نیز استفاده کنید.
- کل محتوا ارسالی را زیپ کرده و نام آن را شماره دانشجویی خود قرار دهید.
- نیازی به نوشتن داک نیست.
- به سوالات انتهایی با دقیق و به صورت کامل پاسخ دهید.
- انجام تمرین تک نفره است. لطفاً به تنهایی انجام شود، در غیر اینصورت نمره منفی در نظر گرفته خواهد شد.
- بدیهی است که نه قرار بر پر کردن هر صفحه سؤال است نه تک کلمه‌ای و انتزاعی مشوش از افکارتان، پس حتی اگر جواب شما بله یا خیر است دلیل یا توضیحی برای آن ارائه کنید (نه سخنان قصار گاندی نه قصه هزار و یک شب). معیار را دوستستان قرار بدهید که بعد خواندن از شما سوال نپرسد.

آیدی تلگرام طراحان :

[@msnp1381](https://t.me/msnp1381)  
[@mainlynia](https://t.me/mainlynia)

## سوالات تمرین

1. در روند پروژه با چه چالش هایی مواجه شدید؟

در هر دو پروژه با چالش امده سازی داده و رفع کردن ویژگی های که مقدار `nan` دارند روبرو بودیم و بعد از آن نرمالایز کردن و در اخر پیاده سازی خود درخت تصمیم و استفاده از هرس کردن و بهینه سازی های مختلف و یادگیری تجمیعی.

تایتانیک / کرونا:

مشکلاتی که داده داشت مثل تبدیل `int 64` به `float 64`

`Nan` بودن مقادیر بعضی ویژگی ها

پیدا کردن بهترین ویژگی های برای کشیدن درخت تصمیم

پیدا کردن بهترین های پرپارامتر ها



۲. درخت به دست آمده برای هر کدام از دیتاست‌ها به چه صورت بوده است؟

تاپتانیک / کرونا:

دیتایی که به ما دادن دزدی بود (یعنی خیلی کشیف و داغون بود)

ماهم یاعلی گفتیم شروع کردیم اول او مدیم مقادیر بولین صفر و یک کردیم بعد مقادیری که نبود بر اساس  
صلاح دید خودمون پر کردیم شاید بپرسید چجوری منم خوبیم شما چطورید (خیلی بامزه بود)

اگر پیوسته بود میانگین گرفتیم ناپیوسته مد گرفتیم و ...

داده های پیوسته رو به چند بخش تقسیم می کنیم که من اومدم به تعداد دسته تقسیم کردم که هر کدام تعدادش برایر باشد.



### ۳. دو معیار آنتروپی و Gini index را مقایسه کنید.

هردو را درست استفاده کنیم در جواب زیاد فرقی وجود ندارد و هر دو معیار پیدا کردن بهترین Gain می باشند.

به معنای بی نظمی است و جینی هم تقریباً دنبال بی نظمی است.

تایتانیک:

Gini

Accuracy: 0.8513

F1 Accuracy: 0.7801

Entropy

Accuracy: 0.8513

F1 Accuracy: 0.7832

کرونا:

Gini

Accuracy: 0.9576

F1 Accuracy: 0.9655

Entropy

Accuracy: 0.9576

F1 Accuracy: 0.9655

۴. برای افزایش دقت چه ایده‌ای دارید؟

تایتانیک / کرونا:

هرس کردن درخت پیدا کردن بهترین الگا

استفاده کردن بهترین هاپیرپارامترها GridsearchCV

کشیدن نمودار داده ها و بررسی مقادیر

انتخاب کردن حذف داده های یا پر کردن آن

وابستگی داده ها را پیدا کنیم.



۵. آیا بیشبرازش داشته اید؟ توضیح دهید.

تایتانیک / کرونا:

بله 😔 اشتباهی داده تست دادم به مدل باعث نشت اطلاعات شد

یکی دیگه هم برداشتم جواب رو باینری کردم که باینری نبود

هرس زیاد درخت

6. چه نکات و کارهایی پروژه شما را متمایز می کند؟

مهمنترین نکته من خیلی خوشگلم 😊

استفاده کردن از درخت های تصمیم پیشرفته که دقت را به 100% نزدیک می کند.

استفاده از نمودار های مختلف برای پیدا کردن هایپر پارامتر

برای پیدا کردن هایپر پارامتر روش های مختلف استفاده کردم greedyCV

هرس کردن و نشون دادن الفا های مختلف

از روش های مختلف Boosting استفاده کردم

AdaBoostClassifier

RandomForestClassifier

GradientBoostingClassifier

تازه رگرسیونشم بلدم جای شنا نبود و گرنه شناگر خوبیم 😊



## 7. مفهوم cross-validation چیست و در چه موقعی استفاده می شود؟

Cross-validation (اعتبارسنجی متقطع) روشی در یادگیری ماشین و آمار است که برای ارزیابی عملکرد یک مدل روی داده‌ها استفاده می‌شود. در این روش، داده‌ها به چند بخش تقسیم می‌شوند تا مدل به شکل بهتری بر روی داده‌ها آزمایش شود و از overfitting (بیش‌برازش) جلوگیری شود.

### **Cross-Validation** نحوه عملکرد

در رایج‌ترین روش آن، یعنی **k-fold cross-validation**، داده‌ها به  $k$  بخش تقسیم می‌شوند. سپس مراحل زیر اجرا می‌شود:

1. مدل آموزش داده می‌شود: یکی از بخش‌ها به عنوان مجموعه اعتبارسنجی (validation) کنار گذاشته می‌شود و مدل با استفاده از بقیه بخش‌ها آموزش داده می‌شود.
2. مدل ارزیابی می‌شود: مدل آموزش دیده بر روی بخش اعتبارسنجی آزمایش می‌شود.
3. تکرار مراحل: این مراحل برای تمام بخش‌ها تکرار می‌شود، به طوری که هر بخش به نوبت به عنوان بخش اعتبارسنجی استفاده می‌شود.
4. محاسبه میانگین دقیقت: در پایان، دقیقت‌های به دست آمده از همه تکرارها میانگین گرفته می‌شوند تا دقیقت نهایی مدل به دست آید.

### **Cross-Validation** موارد استفاده از

این روش عمدتاً در شرایط زیر استفاده می‌شود:

- ارزیابی عملکرد مدل: قبل از پیاده‌سازی مدل نهایی، با استفاده از cross-validation می‌توان عملکرد مدل را بررسی کرد.
- جلوگیری از overfitting: به دلیل استفاده از بخش‌های مختلف داده‌ها به عنوان اعتبارسنجی، مدل از بیش‌برازش به داده‌های خاص جلوگیری می‌کند.
- مقایسه مدل‌ها: با cross-validation، می‌توان چندین مدل را با داده‌های مشابه تست کرد و بهترین مدل را انتخاب کرد.