

تمرین سری دوم

درس: یادگیری ماشین- پاییز ۱۴۰۳

استاد درس: دکتر فاطمه زمانی

دستیاران آموزشی: حسین آقاگل زاده، ابوالفضل حسینی فرد
دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی نوشیروانی بابل

- این تمرین دارای سه بخش تئوری، پیاده سازی و مطالعاتی است.
- پاسخنامه بخش تئوری می تواند به صورت دست نویس یا تایپی باشد. در هر صورت اما می بایست در قالب یک فایل pdf تحویل گردد.
- پاسخ بخش پیاده سازی می بایست حاوی کد پیاده سازی شده به همراه گزارش آن در قالب فایل pdf باشد. از آوردن کد در گزارش اجتناب کنید (مگر آوردن بخشی از آن از نظر شما ضرورت داشته باشد).
- پاسخ بخش مطالعاتی می بایست به صورت تایپ شده در قالب یک فایل pdf باشد.
- گزارش بخش پیاده سازی می بایست میزان تلاش شما رو با ۴ اولویت مهم ۱- شفافیت ۲- درستی ۳- زیبایی ۴- کوتاهی، نشان بد.
- بخش تئوری بدون استفاده از کد انجام بشه مگر دلیل موجهی وجود داشته باشه.
- کل پاسخ (کد و فایل های pdf) در یک فایل zip با فرمت زیر ارسال شود.

HW2.شماره دانشجویی_.zip

تئوری

مسئله ۱

الف) یک متغیر تصادفی به چه معناست؟

ب) ایا ممکن است entropy یک متغیر تصادفی گستته منفی شود؟ برای یک متغیر تصادفی پیوسته چطور؟

پ) در یک مسئله ۵ نمونه زیر که هر کدام ۴ attributes دارند وجود دارد که دو نمونه از آن ها دارای missing value (با علامت "?") مشخص شده اند) هستند. بر اساس روش virtual values چگونه باید با این دو نمونه بر خورد کنیم؟

$$\begin{array}{ccccc} \left\{ \begin{array}{l} a \\ 1 \\ 5 \\ z \end{array} \right. & \left\{ \begin{array}{l} b \\ 8 \\ 2 \\ x \end{array} \right. & \left\{ \begin{array}{l} a \\ ? \\ 3 \\ x \end{array} \right. & \left\{ \begin{array}{l} a \\ -2 \\ 2 \\ w \end{array} \right. & \left\{ \begin{array}{l} ? \\ 4 \\ 1 \\ y \end{array} \right. \end{array}$$

ت) در مورد زمان اجرای دو روش knn و decision tree در مراحل training و test بحث کنید.

مسئله ۲

ما از داده های زیر استفاده خواهیم کرد تا یک درخت تصمیم را یاد بگیریم که پیش بینی می کند آیا افراد در درس یادگیری ماشین قبول می شوند (بله یا خیر)، بر اساس معدل قبلی آنها (بالا، متوسط یا پایین) و اینکه آیا مطالعه کرده اند یا خیر.

GPA	Studied	Passed
L	F	F
L	T	T
M	F	F
M	T	T
H	F	T
H	T	T

الف) موارد زیر را محاسبه کنید

$$H(\text{Passed}) = -$$

$$H(\text{Passed}|GPA) = -$$

$$H(\text{Passed}|Study) = -$$

ب) درخت تصمیمی که برای این مجموعه داده یاد گرفته می شود، رسم کنید.

مسئله ۳

یک رستوران بر این است که بررسی نماید با توجه به عوامل موثر، افرادی که به رستوران مراجعه میکنند در صورتی که تمام میزها پر باشد، آیا برای خالی شدن میز صبر میکنند یا نه؟

نمونه های ثبت شده از ۱۲ مراجعه کننده، جنبه های مختلف و اینکه صبر میکنند/نمیکنند را در جدول زیر مشاهده می فرمایید. برای این مساله decision tree را بر اساس معیار information gain رسم نمایید. سپس بر اساس درخت حاصله برای یک نمونه دلخواه موجود (از بین ۱۲ مورد) قضاوت نمایید که آیا مراجعه کننده در صورت پر بودن میزها صبر میکند یا خیر.

Example	Input Attributes										Goal WillWait
	Alt	Bar	Fri	Hun	Pat	Price	Rain	Res	Type	Est	
x_1	Yes	No	No	Yes	Some	\$\$\$	No	Yes	French	0-10	$y_1 = Yes$
x_2	Yes	No	No	Yes	Full	\$	No	No	Thai	30-60	$y_2 = No$
x_3	No	Yes	No	No	Some	\$	No	No	Burger	0-10	$y_3 = Yes$
x_4	Yes	No	Yes	Yes	Full	\$	Yes	No	Thai	10-30	$y_4 = Yes$
x_5	Yes	No	Yes	No	Full	\$\$\$	No	Yes	French	>60	$y_5 = No$
x_6	No	Yes	No	Yes	Some	\$\$	Yes	Yes	Italian	0-10	$y_6 = Yes$
x_7	No	Yes	No	No	None	\$	Yes	No	Burger	0-10	$y_7 = No$
x_8	No	No	No	Yes	Some	\$\$	Yes	Yes	Thai	0-10	$y_8 = Yes$
x_9	No	Yes	Yes	No	Full	\$	Yes	No	Burger	>60	$y_9 = No$
x_{10}	Yes	Yes	Yes	Yes	Full	\$\$\$	No	Yes	Italian	10-30	$y_{10} = No$
x_{11}	No	No	No	No	None	\$	No	No	Thai	0-10	$y_{11} = No$
x_{12}	Yes	Yes	Yes	Yes	Full	\$	No	No	Burger	30-60	$y_{12} = Yes$

توضیح جنبه های مختلف که در Input attributes آمده است به شرح زیر است:

1.	Alternate: whether there is a suitable alternative restaurant nearby.
2.	Bar: whether the restaurant has a comfortable bar area to wait in.
3.	Fri/Sat: true on Fridays and Saturdays.
4.	Hungry: whether we are hungry.
5.	Patrons: how many people are in the restaurant (values are None, Some, and Full).
6.	Price: the restaurant's price range (\$, \$\$, \$\$\$).
7.	Raining: whether it is raining outside.
8.	Reservation: whether we made a reservation.
9.	Type: the kind of restaurant (French, Italian, Thai or Burger).
10.	WaitEstimate: the wait estimated by the host (0-10 minutes, 10-30, 30-60, >60).

• آوردن محاسبات مربوطه در پاسخنامه تا عمق دوم کافی است اما درخت باید کامل رسم شود.

مسئله ۴

در یک مسئله دو کلاسه، الگوهای آموزشی زیر را که هر کدام دارای چهار ویژگی باینری هستند، در نظر بگیرید.

ω_1	ω_2
0110	1011
1010	0000
0011	0100
1111	1110

الف) با استفاده از entropy impurity، یک دسته‌بند درخت تصمیم برای این داده ایجاد کنید.

ب) هر دسته را با simplest logical expression OR و AND بیان کنید، یعنی با کمترین تعداد

پیاده سازی

مسئله ۱

مجموعه داده iris را در نظر بگیرید.

https://scikit-learn.org/1.5/auto_examples/datasets/plot_iris_dataset.html

این مجموعه داده دارای ۱۵۰ نمونه است که هرکدام ۴ ویژگی دارند. هر نمونه متعلق به یکی از ۳ برچسب (کلاس) است .(Setosa, Versicolour, and Virginica)

(الف) مجموعه داده را به برنامه اضافه کنید. به صورت تصادفی ۳۰ نمونه آن را به عنوان مجموعه test و مابقی را به عنوان مجموعه training در نظر بگیرید.

(ب) از ۴ ویژگی را به صورت دلخواه انتخاب کنید و نمونه ها را بر روی یک نمودار (نمودار scatter) نشان دهید طوری که نمونه ها با برچسب یکسان دارای رنگ مشخص دلخواه باشند (هر برچسب یک رنگ). این بخش را فقط برای مجموعه training انجام دهید.

(پ) مازول پیاده سازی decision tree در کتابخانه sklearn را در نظر بگیرید:

<https://scikit-learn.org/dev/modules/generated/sklearn.tree.DecisionTreeClassifier.html>

هر یک از پارامتر ("Parameters") های تعریف این کلاس را به طور مختصر شرح دهید.

```
class sklearn.tree.DecisionTreeClassifier(*, criterion='gini', splitter='best',
max_depth=None, min_samples_split=2, min_samples_leaf=1,
min_weight_fraction_leaf=0.0, max_features=None, random_state=None,
max_leaf_nodes=None, min_impurity_decrease=0.0, class_weight=None, ccp_alpha=0.0,
monotonic_cst=None)
\[source\]
```

A decision tree classifier.

(ت) به کمک کتابخانه scikit learn مرحله آموزش decision tree را بر روی مجموعه training انجام دهید (شامل ۴ ویژگی). با درخت حاصل شده یک بار برچسب نمونه های training و یک بار برچسب نمونه های test را پیش بینی کنید. برچسب های پیش بینی شده را برای دو مجموعه به طور مجزا به برچسب های واقعی آن ها مقایسه کنید و درصد برچسب های پیش بینی شده درست را گزارش کنید (accuracy test و accuracy train).

* این بخش را برای هر سه criterion موجود در این مازول انجام دهید

(ث) بخش ت را این بار با الگوریتم Random Forest انجام دهید (با کمک کتابخانه sklearn و فقط با یک criterion دلخواه).

مسئله ۲

الگوریتم ID3 را به صورت دستی بر روی دیتای موجود در لینک زیر پیاده کنید.

<https://www.kaggle.com/datasets/tareqjoy/trainplaytennis>

- لینک زیر شامل پیاده سازی دستی این الگوریتم هست و لی حق استفاده مستقیم از آن (با چنین کدهایی) را ندارید و کد باید توسط خودتان نوشته و گزارش شود. کمک گرفتن از آن مشکلی ندارد.

<https://medium.com/geekculture/step-by-step-decision-tree-id3-algorithm-from-scratch-in-python-no-fancy-library-4822bbfdd88f> (نیاز به فیلتر شکن)

مطالعاتی

مقاله‌ی پیوست شده را مطالعه کنید و برداشت خود از آن را شرح دهید (حداقل ۲ صفحه).

- بخشی از مقاله که از قبل در کلاس فرا گرفته اید را کوتاه‌تر بیاورید.

Capcom باشید 😊