
سلام بر دانشجویان عزیز، چند نکته مهم:

1. کدهای ارسال شده **بدون گزارش/وویس** فاقد نمره می‌باشند.
2. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.
3. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW#_StudentNumber داشته باشد.
4. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.
5. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.
6. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.
7. در صورت داشتن سوال، از طریق ایمیل s.dashti.k@gmail.com , 1380ali.mohamadi@gmail.com سوال خود را مطرح کنید.

سوال ۱: درخت تصمیم (20 نمره)

فرخنده مدتی است که از دریافت ایمیل‌های spam کلافه شده است و هم‌اکنون پس از مطالعه مباحث مربوط به یادگیری ماشین، قصد دارد تا ایمیل‌های ورودی به آدرس خود را دسته‌بندی کرده و ایمیل‌های Spam را از ایمیل‌های Ham (ایمیل‌های غیر اسپم) جدا کند.

- جدول ۱-۱ شامل ۱۴ داده و شامل مجموعه دادگان آموزش شما می‌باشد.
- جدول ۱-۲ شامل ۶ داده و شامل مجموعه دادگان آزمون شما می‌باشد.

قسمت اول

یک طبقه‌بند درخت تصمیم مبتنی بر Information gain را با عمق ۳ (با احتساب ریشه و برگ‌ها) برای پیش‌بینی spam و یا ham بودن ایمیل ها را بر روی مجموعه دادگان جدول ۱-۱ آموزش دهید. علاوه بر نشان دادن درخت تصمیم نهایی، مراحل محاسبات خود، برای ساخت آن را بنویسید.

قسمت دوم

با استفاده از طبقه‌بند ساخته‌شده در قسمت اول، طبقه هر کدام از داده های آزمون جدول ۱-۲ را پیش‌بینی کنید.

شماره	نرخ شکایت از فرستنده	طول بدنه ایمیل	فرمت ایمیل	دامنه ایمیل	تشخیص
۱	پایین	متوسط	مشکوک	gmail	Ham
۲	بالا	کوتاه	مشکوک	gmail	Spam
۳	پایین	متوسط	مشکوک	گمنام	Ham
٤	پایین	کوتاه	مشکوک	gmail	Spam
٥	بالا	کوتاه	نامشکوک	academic	Ham
٦	پایین	طولانی	نامشکوک	gmail	Spam
۷	بالا	متوسط	نامشکوک	gmail	Spam
۸	بالا	کوتاه	مشکوک	academic	Spam
۹	پایین	طولانی	مشکوک	گمنام	Spam
۱۰	پایین	کوتاه	نامشکوک	gmail	Ham
۱۱	بالا	طولانی	نامشکوک	academic	Spam
۱۲	بالا	طولانی	مشکوک	گمنام	Spam
۱۳	بالا	متوسط	مشکوک	academic	Spam
۱٤	پایین	متوسط	نامشکوک	گمنام	Ham

جدول ۱-۱

شماره	نرخ شکایت از فرستنده	طول بدنه ایمیل	فرمت ایمیل	دامنه ایمیل	تشخیص
۱	پایین	کوتاه	مشکوک	academic	نامشکوک
۲	بالا	کوتاه	نامشکوک	gmail	مشکوک
۳	بالا	کوتاه	مشکوک	گمنام	مشکوک
٤	بالا	متوسط	مشکوک	گمنام	مشکوک
٥	پایین	متوسط	مشکوک	academic	نامشکوک
٦	بالا	طولانی	نامشکوک	gmail	مشکوک

جدول ۱-۲

سوال ۲: درخت تصمیم (20 نمره)

درخت تصمیم و جنگل تصادفی را با استفاده از مفاهیم بایاس و واریانس را مقایسه کنید. چگونه می‌توان برای رسیدن به الگوریتم بهینه، تعادل بین بایاس و واریانس را بر قرار کرد.

لینک راهنمایی

سوال ۳: کلاسترینگ (20 نمره)

یکی از ایده‌های اصلی در clustering استفاده از فاصله‌ی بین نقاط است. آیا این روش همیشه جواب می‌دهد؟ در چه شرایطی این روش می‌تواند نتیجه منفی بدهد.

الگوریتم DBSCAN را توضیح دهید. همچنین توضیح در کدام دسته از الگوریتم‌های clustering قرار می‌گیرد. در ادامه تفاوت آن را با الگوریتم OPTICS شرح دهید.

سوال ۴: درخت تصمیم (**شبیه سازی**، 25 نمره)

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، برای خالی شدن میز صبر می‌کنند یا نه؟

داده‌های ثبت شده از ۱۲ مراجعه کننده، جنبه‌های مختلف و اینکه صبر می‌کنند یا نه را در جدول زیر نشان می‌دهد.

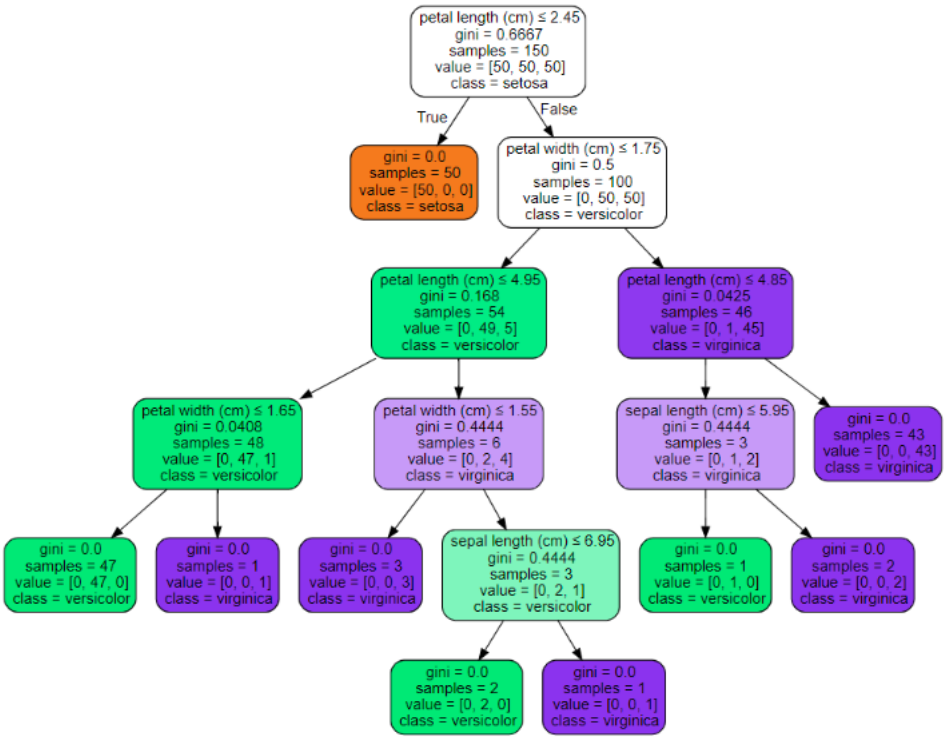
Example	Input Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
$\mathbf{x}_1$	*Yes*	*No*	*No*	*Yes*	*Some*	*$$$*	*No*	*Yes*	*French*	*0–10*	$y_1 = Yes$
$\mathbf{x}_2$	*Yes*	*No*	*No*	*Yes*	*Full*	*$*	*No*	*No*	*Thai*	*30–60*	$y_2 = No$
$\mathbf{x}_3$	*No*	*Yes*	*No*	*No*	*Some*	*$*	*No*	*No*	*Burger*	*0–10*	$y_3 = Yes$
$\mathbf{x}_4$	*Yes*	*No*	*Yes*	*Yes*	*Full*	*$*	*Yes*	*No*	*Thai*	*10–30*	$y_4 = Yes$
$\mathbf{x}_5$	*Yes*	*No*	*Yes*	*No*	*Full*	*$$$*	*No*	*Yes*	*French*	*>60*	$y_5 = No$
$\mathbf{x}_6$	*No*	*Yes*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Italian*	*0–10*	$y_6 = Yes$
$\mathbf{x}_7$	*No*	*Yes*	*No*	*No*	*None*	*$*	*Yes*	*No*	*Burger*	*0–10*	$y_7 = No$
$\mathbf{x}_8$	*No*	*No*	*No*	*Yes*	*Some*	*$$*	*Yes*	*Yes*	*Thai*	*0–10*	$y_8 = Yes$
$\mathbf{x}_9$	*No*	*Yes*	*Yes*	*No*	*Full*	*$*	*Yes*	*No*	*Burger*	*>60*	$y_9 = No$
$\mathbf{x}_{10}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$$$*	*No*	*Yes*	*Italian*	*10–30*	$y_{10} = No$
$\mathbf{x}_{11}$	*No*	*No*	*No*	*No*	*None*	*$*	*No*	*No*	*Thai*	*0–10*	$y_{11} = No$
$\mathbf{x}_{12}$	*Yes*	*Yes*	*Yes*	*Yes*	*Full*	*$*	*No*	*No*	*Burger*	*30–60*	$y_{12} = Yes$

توضیح فیچرهای مختلف نیز 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).

- مرحله (یا لایه) اول درخت تصمیم را با استفاده از معیار آنتروپی به صورت دستی حل کنید.
- طبقه بند درخت تصمیم را بدون استفاده از پکیج آماده و با در نظر گرفتن معیار آنتروپی کرده پیاده‌سازی کرده و نتایج پیاده‌سازی را گزارش کنید.
- طبقه‌بند درخت تصمیم را با استفاده از پکیج‌های آماده و با در نظر گرفتن معیار آنتروپی پیاده‌سازی کرده و آن را رسم کنید. شکل ۲ نشان‌دهنده نمونه از تصویر رسم شده برای درخت تصمیم می‌باشد.

- تفسیر خود از این درخت تصمیم را شرح دهید.



شکل ۲ – درخت تصمیم رسم شده توسط پکیج scikit learn

سوال ۵: کلاسترنیگ (**شبیه سازی**، 25 نمره)

مجموعه داده مشتریان بازار یک مجموعه داده نمونه است که برای تقسیم بندی مشتریان طراحی شده است. این شامل اطلاعاتی درباره 200 مشتری یک مرکز خرید، از جمله جزئیات جمعیتی و الگوهای خرید آنها است. این مجموعه داده معمولاً در یادگیری ماشینی بدون نظارت برای شناسایی گروه‌های مشتریان متمایز بر اساس رفتارهای مخارج و سطوح درآمد استفاده می‌شود.

فیچرهای دیتاست:

شناسه مشتری: یک شناسه منحصر به فرد برای هر مشتری.

جنسیت: جنسیت مشتری (مرد / زن)

سن: سن مشتری بر حسب سال.

درآمد سالانه :(k$) درآمد سالانه مشتری به هزار دلار.

امتیاز هزینه (1-100): امتیازی که مرکز خرید بر اساس هزینه و رفتار مشتری اختصاص می دهد. نمرات بالاتر نشان دهنده هزینه های مکرر و/یا زیاد است.

ابتدا به بررسی دیتاست و ویژگی‌های آن بپردازید و سپس با انتخاب ویژگی‌ها و الگوریتم خوشه‌بندی k-means را با مقدار اولیه k=3 پیاده‌سازی کنید. خوشه‌ها را در یک نمودار پراکندگی دوبعدی، با درآمد سالانه روی محور x و امتیاز خرید روی محور y و خوشه‌ها با رنگ‌های مختلف، نمایش دهید. و در انتها چگونه می‌توان مقدار بهینه k را تعیین کرد؟ درباره این موضوع تحقیق کنید و یک راه را انجام دهید.