به نام خدا

سری دوم تمارین درس داده کاوی

استاد درس:

دکتر مریم امیر مزلقانی

نیم سال اول ۱۴۰۲-۱۴۰۳

راه ارتباطی:

Aut.DataMining.Fall@gmail.com

توضیحات:

۱. این تمرین شامل دو بخش عملی و تئوری هست و پاسخ به هر دو بخش الزامی است.

۲. تمرین عملی در قالب یک نوت بوک آماده شده است و دیتای لازم برای این تمرین در پوشه ***Practical*** موجود است.

۳. در نوت بوک تمرین عملی حتما هر خواسته را در بلوک مربوط به خودش انجام دهید.

۴. حین حل تمرین عملی شما نیازمند به ساخت ۴ فایل جدید هستید (۱ فایل csv و ۳ فایل png)، به نحوه نامگذاری این فایل ها دقت کنید.

۵. **ملاک اصلی انجام تمارین عملی، گزارش است** و ارسال کد بدون pdf گزارش فاقد ارزش است. لذا برای این بخش یک فایل گزارش تهیه کنید و در آن برای هر بخش از تمرین عملی، توضیحات مربوط به آن را ذکر کنید.

۶. **تمرین عملی این بخش با ددلاین متفاوتی بارگزاری خواهد شد.**

۷. خوانا و مرتب بودن پاسخ های شما در نمره‌تان تاثیر مثبت خواهد داشت.

۸. مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیرمجاز بوده و برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری **الزاما با ذکر منبع** بلامانع است.

۹. فایل های ایجاد شده در تمرین عملی + نوت بوک + گزارش + پاسخ تمارین تئوری را به صورت زیپ در آورده و با فرمت **StudentID_DM01.zip** در سامانه کورسز آپلود نمایید.

۱۰. تاخیر مجاز ۷ روزه شما با جمع کردن میزان تاخیر در تمارین محاسبه خواهد شد و پس از اتمام ۷ روز هر روز تاخیر باعث کسر **۲۰** ٪ **نمره** خواهد شد.

توجه: با توجه به تاریخ میانترم(۲۹ آبان) و تاریخ تحویل تمرین(۲۵ آبان)، هرگونه تاخیر در ارسال تمرین بعد از ۲۷ام آبان ساعت ۲۲:۳۰ منجرب به از دست دادن نمره کل تمرین خواهد شد. زیرا پاسخنامه تمرین در این زمان در سایت قرار خواهد گرفت.

بخش تئوری:

سوال ۱.

الف) مجموعه داده زیرمفروض است با در نظر گرفتن کل داده‌های این مجموعه، آنتروپی آن را محاسبه کنید.

ب) از ویژگی‌های a۱ و a۲ Information gain حاصل را محاسبه کنید.

ج) برای a۳ که یک ویژگی پیوسته است Information gain را با در نظر گرفتن یک آستانه محاسبه کنید.

د) با توجه به مقادیر محاسبه شده در گام‌های قبل، درخت تصمیم یک سطحی را برای این دیتاست طراحی نمایید.

Instance	a1	a2	a3	Target class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

سوال ۲.

الف) در مسئله رگرسیون با در نظر گرفتن مربع خطا باقیمانده، به مسئله بهینه‌سازی $\min||X\beta - y||_2^2$ می‌رسیدیم . اگر به جای مربع خطا از قدرمطلق استفاده شود، مسئله بهینه سازی حاصل به چه صورتی خواهد بود و انتظار دارید در نظر گرفتن قدرمطلق سبب چه تغییری روی نتیجه به دست آمده شود.

ب) در نظر گرفتن قدرمطلق سبب چه تغییری در روند حل با رویکرد بهینه سازی descent gradient خواهد شد.

ج) ۳ کلاس داده تحت عناوین کلاس‌های A, B و C داریم. از هر کلاس، ۵ داده در اختیار داریم و هر داده شامل k ویژگی است. هدف مشخص نمودن کلاس داده جدید y است(که این داده هم شامل همان k ویژگی است). برای این منظور می خواهیم با استفاده از رگرسیون این کلاس بندی را انجام دهیم. روش پیشنهادی شما چیست. مراحل را شرح دهید.

داده های کلاس A: $X_{A1}, X_{A2}, X_{A3}, X_{A4}, X_{A5}$,

داده های کلاس B: $X_{B1}, X_{B2}, X_{B3}, X_{B4}, X_{B5}$

داده های کلاس C: $X_{C1}, X_{C2}, X_{C3}, X_{C4}, X_{C5}$

سوال ۳.

در جدول زیر سن و فشار خون چند بیمار قلبی داده شده است. معادله رگرسیون به فرم $y = \beta_0 + \beta_1 x$ به دست آورید .همچنین با استفاده از معادله به دست آمده فشار خون یک بیمار ۴۰ ساله را پیش‌بینی کنید (متغیر x نشان‌دهنده سن و متغیر y نشان دهنده فشار خون است)

PATIENT	A	B	C	D	E	F	G
x	۴۲	۷۴	۴۸	۳۵	۵۶	۲۶	۶۰
y	۹۸	۱۳۰	۱۲۰	۸۸	۱۸۲	۸۰	۱۳۵

سوال ۴.

درک overfitting و underfitting آسان است، اما شناسایی آنها دشوار است. دو حالت را در یک تسک classification در نظر بگیرید:

(۱) دقت آموزش ۱۰۰٪ و دقت آزمون ۵۰٪ است

(۲) دقت آموزش ۸۰٪ و دقت آزمون ۷۰٪ است.

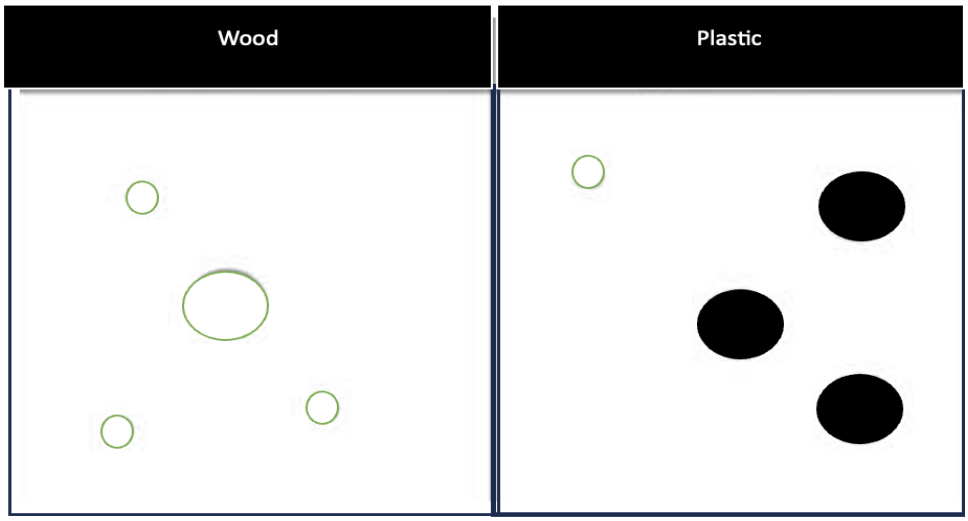
در کدام حالت احتمال وقوع overfitting بیشتر است؟

سوال ۵.

یکی از راه‌های جلوگیری از بیش‌برازش استفاده از منظمسازی است که به دو نوع L۱ و L۲ تقسیم میشود. به نوع اول Lasso Regression و به نوع دوم Ridge regression گفته میشود. تفاوت این دو روش را از نوع بهینه‌سازی بیان کرده و نحوه کار آن‌ها را توضیح دهید .

سوال ۶.

شکل زیر مثالهای آموزشی برای هدف تشخیص ماده چوب یا پلاستیکی را نشان میدهد. براساس دو ویژگی سایز و رنگ، درخت تصمیم را با استفاده از آنتروپی و Information gain بدست آورید.



سوال ۷.

قارچ ها به دو دسته سمی و غیرسمی تقسیم می شوند. اطلاعات زیر را در اختیار داریم. هدف استفاده از درخت های تصمیم برای تعیین سمی یا غیرسمی بودن داده شماره ۹ است و داده های ۱ تا ۸ داده های آموزشی هستند. از معیار GINIبرای ساخت درخت تصمیم استفاده نمایید.

HEAVY	SPOTTED	SMOOTH	POISONOUS
NO	NO	NO	NO
NO	YES	NO	NO
YES	NO	YES	NO
YES	NO	YES	YES
NO	YES	NO	YES
NO	YES	YES	YES
NO	NO	YES	YES
YES	NO	NO	YES
NO	NO	YES	**?**

الف) کدام ویژگی را به عنوان ریشه درخت انتخاب می کنید

ب) بهره به دست آمده بر اثر انتخاب ویژگی قسمت الف را محاسبه نمایید.

ج) درخت تصمیم را تا دو سطح سوال از ویژگیها به دست آورید.

د) پیش بینی درخت به دست آمده در بخش ج برای داده آخر (داده شماره ۹) چیست. در تمامی مراحل تمامی محاسبات ذکر شود.

نکته : اگردر انتخاب ویژگی ها برای تصمیم گیری در ساخت درخت، چندین ویژگی از نظر معیار ارزیابی مشابه بودند، به طور تصادفی یکی را انتخاب نمایی

در صورت هرگونه ابهام به تدریسیاران از طریق ایمیل درس و یا آیدی های تلگرامی زیر پیام دهید.

سوالات ۱و ۶:

@MoeinNasiri1379

باقی سوالات:

@HeliaHashemipour