دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

درس مبانی داده‌کاوی

تمرین سری دوم

توجه: پیش از شروع تمرین لطفا موارد زیر را با دقت مطالعه نمایید.

لطفا تمام فایل‌های تمرین را (از جمله فایل pdf گزارش و فایل‌های کد) در یک فایل zip/rar ذخیره کرده و نام آن را به HW۲_XXXXXXXX.zip تغییر دهید. سپس آن را در مودل بارگذاری کنید.

سوال‌ها به دو بخش نظری و برنامه‌نویسی تقسیم شده‌اند. سوال‌های نظری را می‌توانید به جای فایل word در برگه کاغذ انجام داده و تصویر آن را در فایل word قرار دهید. دقت کنید که خوانایی تمرین شرط لازم برای دریافت نمره آن است. توجه کنید که فایل گزارش را حتما به صورت pdf شده تحویل دهید.

تمرین‌های برنامه‌نویسی را می‌توانید با یکی از زبان‌های Matlab یا Python انجام دهید.

مهلت تحویل سوال ۱ تا ۶ تمرین (تمرین‌های تشریحی) تا روز یکشنبه ۲۷ آبان ساعت ۱۲:۰۰ ظهر است. مهلت تحویل سوال ۷ و ۸ (تمرین‌های برنامه‌نویسی) تا روز جمعه ۹ آذر ساعت ۲۳:۵۵ شب می‌باشد.

سوال اول

(a) درخت تصمیمی که از آموزش ۴ مثال آموزشی EnjoySport با الگوریتم hunt بدست می‌آید را بدست آورید. انتخاب معیار عدم خلوص بر عهده خودتان است.

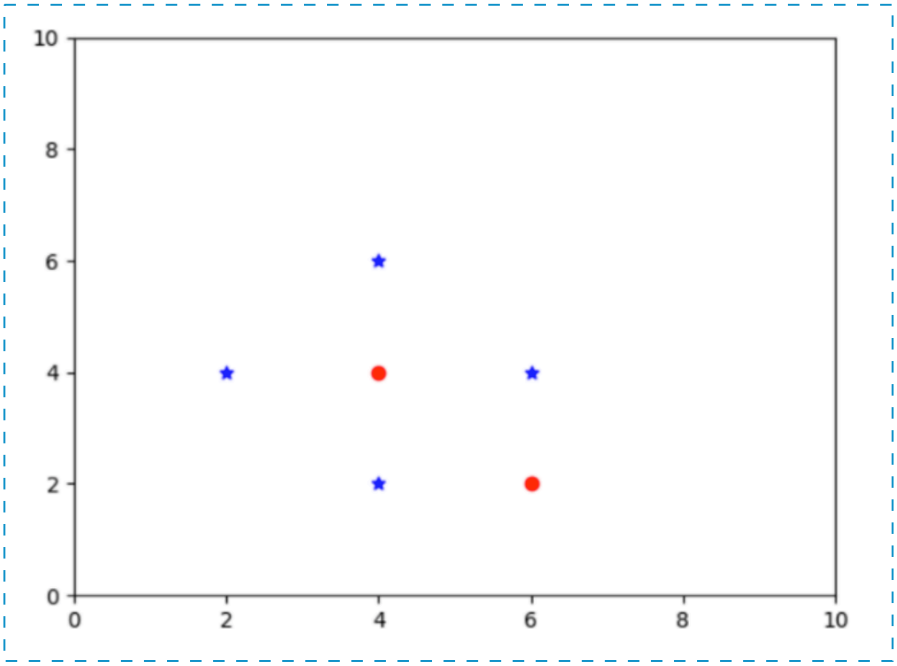
Example	Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport?
1	Sunny	Warm	Normal	Strong	Warm	Same	Yes
2	Sunny	Warm	High	Strong	Warm	Same	Yes
3	Rainy	Cold	High	Strong	Warm	Change	No
4	Sunny	Warm	High	Strong	Cool	Change	Yes

(b) مثال آموزشی زیر را به مجموعه اضافه کنید و درخت تصمیم حاصل با hunt را بدست آورید.

Sky	AirTemp	Humidity	Wind	Water	Forecast	EnjoySport?
Sunny	Warm	Normal	Weak	Warm	Same	No

سوال دوم

با در نظر گرفتن داده های نمایش داده شده در نمودار زیر به هریک از سوال های زیر پاسخ دهید. توجه کنید که داده های مشخص شده با رنگ یکسان در یک کلاس قرار دارند.



(a) **با استفاده از روش ۱-NN و با در نظر گرفتن متریک اقلیدسی، مرز های تصمیم گیری را برای این مجموعه داده رسم کنید. توجه کنید که لازم است روش استنتاج را توضیح دهید. (راهنمایی: می توانید از این لینک کمک بگیرید. نواحی مربوط به هر کلاس را مشخص کنید.**

(b) **با در نظر گرفتن متریک اقلیدسی و با استفاده از روش ۱-NN، نقطه‌ی (۸ , ۱) را به کدام کلاس نسبت خواهید داد؟**

(c) **توضیح دهید در صورتی که از الگوریتم K-NN با مقادیر K خیلی بزرگ استفاده شود چه اتفاقی رخ خواهد داد؟**

(d) **آیا از الگوریتم K-NN می توان برای مساله رگرسیون استفاده کرد؟ توضیح دهید.**

سوال سوم

در یک مساله ی دسته بندی دو کلاسه که با وجود فقط یک ویژگی تعریف شده است، اطلاعات زیر برای هر کلاس داده شده است (منظور از w_i ، کلاس i ام می‌باشد) :

$$p(x \mid w_1) = \begin{cases} 2x & 0 \le x \le 1 \\ 0 & otherwise \end{cases} \qquad p(x \mid w_2) = \begin{cases} 2 - 2x & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$$

با فرض آنکه احتمال های اولیه دو کلاس با هم برابر باشند یا $p(w_1) = p(w_2)$ ، به هر یک از سوالات زیر پاسخ دهید:

(a) تابع توزیع را برای هر کلاس رسم کنید.

(b) نحوه ی جدا سازی این دو کلاس را با استفاده از قانون بیز بطور خلاصه توضیح دهید.

بخش امتیازی: مرز جدا کننده ی این دو کلاس را بدست آورید و آنرا در شکل نمایش دهید.

سوال چهارم

ویژگی افراد را در جدول زیر در خصوص تایید یا رد اعطای وام در یک بانک خصوصی در نظر بگیرید.

Features	Values
Income	Low, Medium, High
Education	HS, BS, MS, PhD
Dept	Low, Medium, High
Decision	Yes, No

جدول زیر مجموعه داده های آموزشی به منظور یادگیری این مساله را نمایش میدهد.

Example	Income	Education	Dept	Decision
1	High	HS	Low	Yes
2	Low	HS	Medium	Yes
3	Low	BS	High	Yes
4	Low	MS	Low	No
5	Medium	MS	High	Yes
6	Medium	MS	Low	No
7	High	MS	Medium	No
8	High	PhD	High	No

(a) با فرض استقلال ویژگی های مذکور، با استفاده از الگوریتم دسته بندی بیز ساده یا Naïve Bayes Classifier، چه برچسبی (Yes, No) برای داده ی زیر پیش بینی می کنید؟

< (Income=Low, Education=MS, Debt=High), ? >

(b) نمایش گرافی مدل مورد استفاده را ارائه دهید. (گراف احتمال شرطی)

(c) جدول احتمالاتی گره‌های گراف را رسم کنید.

سوال پنجم

یک مسئله کلاس بندی دو کلاسه (تست آزمون فرضیه باینری) را در نظر بگیرید. اگر $x_1, x_2, \ldots, x_n$ ویژگی‌های اندازه گیری شده باشند که بر اساس آنها بین دو کلاس ۰ و ۱ تصمیم گیری می شود و داشته باشیم:

$$for\ all\ i \in \{1, 2, \ldots, n\};\ x_i \mid H_0 \sim N(0, \delta)$$
$$for\ all\ i \in \{1, 2, \ldots, n\};\ x_i \mid H_1 \sim N(m, \delta)$$

(a) **تست نسبت درست‌نمایی را نوشته و تا حد ممکن ساده کنید.**

(b) **فرمول کلی P_{FA} و P_D را بنویسید.**

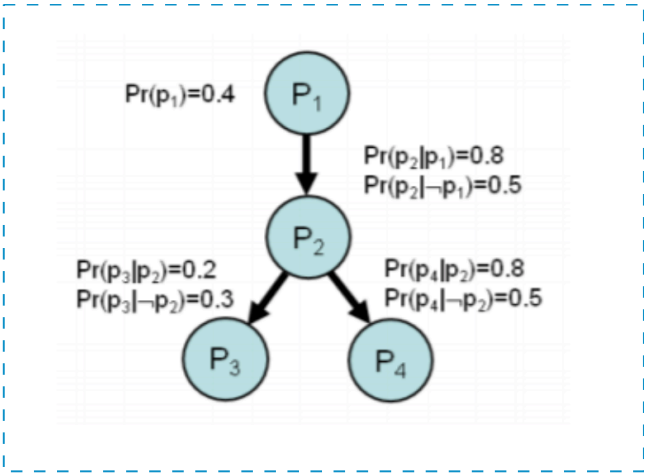
سوال ششم

با استفاده از قوانین احتمال که تاکنون آموخته‌اید، با توجه به مدل گرافی زیر مقادیر خواسته شده را به دست آورید.

دقت کنید که متغیرهای تصادفی p۱ تا p۴ از نوع دوتایی (binary) با مقادیر true و false هستند و عملگر ¬ به معنی not منطقی است. (می‌توانید p۳ را معادل P۳=true و ¬p۳ را به معنی P۳=false در نظر بگیرید)

(a) $Pr(\neg p_3)$

(b) $Pr(p_2 \mid \neg p_3)$



سوال هفتم

نمونه داده های فایل US Presidential Data را باز کنید. این فایل حاوی آمار برد و باخت انتخابات ریاست جمهوری آمریکا است که دارای دو دسته و ۱۳ ویژگی است. دو دسته، شامل برد (۱) و باخت (۰) در انتخابات ریاست جمهوری می شود. ۱۳ ویژگی شامل موارد زیر در متن های سخنرانی انتخاباتی است:

(۱) نسبت کلمات نشان دهنده:

- خوش بینی
- بدبینی
- گذشته
- حال
- آینده

(۲) تعداد دفعاتی که نامزد از حزب خود نام می برد

(۳) تعداد دفعاتی که نامزدی از حزب رقیب نام می برد

(۴) معیاری از محتوای نشان دهنده:

- صراحت
- وجدان
- برون گرایی
- موافق بودن
- عصبیت
- احساسی بودن

داده ها را به مجموعه Train و Test تقسیم کنید.

(a) داده های مجموعه Train را با الگوریتم NN-1 دسته بندی کنید. خطای دسته بندی بر روی داده Train را گزارش کنید.

(b) بخش ۱ را با الگوریتم های K-NN با مقادیر مختلف برای K امتحان کنید. نمودار خطای الگوریتم را بر اساس k های مختلف به دست آورید. تاثیر مقدار k را در دقت الگوریتم K-NN چگونه ارزیابی می کنید.

سوال هشتم

در این بخش با دادگان داده شده باید الگوریتم Hunt را برای ساختن درخت تصمیم پیاده‌سازی کنید. دادگان داده شده شامل ۳ مجموعه train، test و validation است. مجموعه train شامل دادگان نویزی نیز می‌باشد تا بتوان اثر overfitting را در آن مشاهده کرد. این مجموعه شامل مجموعه دادگان باینری است که برای طبقه‌بندی قارچ‌ها به ۲ دسته سمی و خوراکی با استفاده از ۲۲ ویژگی گسسته بکار می‌رود.

مراحل زیر را انجام دهید:

(a) **با الگوریتم hunt و مجموعه دادگان train داده شده یک درخت تصمیم بسازید. برای انتخاب بهترین ویژگی در هر مرحله باید از information gain استفاده کنید. میزان صحت طبقه‌بندی را روی مجموعه‌های train و test محاسبه کنید.**

(b) **انجام post-pruning و کاهش تعداد گره‌های درخت: به کمک دادگان validation این کار را انجام دهید، و منحنی تغییر خطای طبقه‌بندی روی هر ۳ مجموعه train، test و validation را به ازاء تعداد گره‌های مختلف درخت بکشید.**

راهنمایی: برای راحتی کار در مرحله هرس کردن، می‌توانید از یک آستانه برای تصمیم‌گیری در مورد هر گره استفاده کنید، مثلا اگر در مورد هر گره حداقل ۰.۵% بهبود عملکرد طبقه‌بندی درخت تصمیم نتیجه روی مجموعه validation داشتید، می‌توانید آن گره را هرس کنید.