به نام خدا

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

۷. 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.

۸. 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.

۹. شایان ذکراست هر روز تاخیر باعث کسر **۲۰ ٪ نمره** خواهد شد.

بخش تئوری:

سوال ۱.

درجدول داده شده زیر با استفاده از قانون بیز برچسب داده زیر را به دست آورید.در صورت صفر شدن احتمال، از هموارسازی لاپلاس (Laplace smoothing)استفاده کنید.

(معدل = عالی ، مطالعه = بله ، حضور = خیر)

پاس شدن	حضور در کلاس‌ها	مطالعه برای امتحان	معدل
خیر	خیر	خیر	ضعیف
بله	بله	بله	ضعیف
خیر	خیر	خیر	متوسط
بله	بله	بله	متوسط
بله	خیر	خیر	عالی
بله	بله	بله	عالی

سوال ۲.

مجموعه داده زیر را در نظر بگیرید.

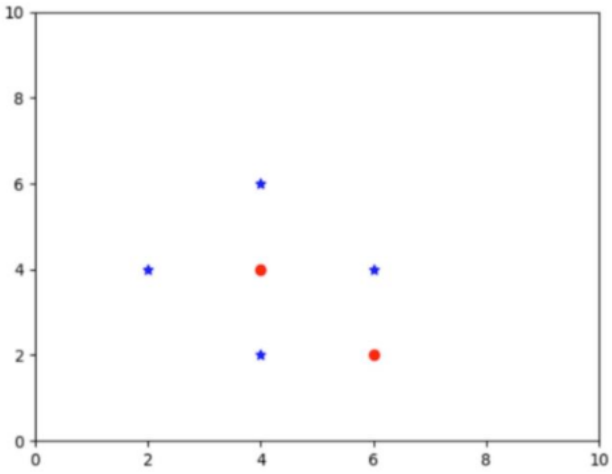
الف) احتمالات شرطی برای$P(A \mid +)$ ، $P(B \mid +)$ ، $P(C \mid +)$ ، $P(A \mid -)$، $P(B \mid -)$ و$P(C \mid -)$ را بصورت تخمینی محاسبه کنید.

ب) با استفاده از محاسبات قسمت قبل برچسب داده($A = ۰, B = ۱, C = ۰$) را با استفاده از روش بیز ساده پیش بینی کنید.

شماره داده	A	B	C	برچسب کلاس
۱	0	0	0	+
۲	0	0	1	-
۳	0	1	1	-
۴	0	1	1	-
۵	0	0	1	+
۶	1	0	1	+
۷	1	0	1	-
۸	1	0	1	-
۹	1	1	1	+
۱۰	1	0	1	+

سوال ۳.

داده های نمایش داده شده زیر را در نظر گرفته و به سوالات زیر پاسخ دهید.



الف) با استفاده از روش KNN و در حالت $K = ۱$ ، مرزهای تصمیم‌گیری را برای این مجموعه داده مشخص نمایید. معیار فاصله اقلیدسی در نظر گرفته و روش کار خود را توضیح دهید.

ب) آیا میتوان از الگوریتم KNN برای مساله رگرسیون استفاده کرد؟ توضیح دهید.

ج) آیا استفاده از knn بر روی دیتاست بزرگ پیشنهاد میشود؟ چرا؟

سوال ۴.

در جدول زیر مختصات تعدادی نقطه در دنیای دو بعدی و کلاس مربوط به هر کدام از آنها مشخص شده است (x: محور افقی، y: محور عمودی). فرض کنید می‌خواهیم کلاس نقطه جدیدی با مختصات x=1 و y=1 را با استفاده از الگوریتم 3-NN و 7-NN تعیین کنیم.

x	y	Class
−1	1	−
0	1	+
0	2	−
1	−1	−
1	0	+
1	2	+
2	2	−
2	3	+

الف) با در نظر گرفتن فاصله اقلیدسی (majority vote) مشخص کنید این نقطه به کدام کلاس تعلق دارد؟

ب) قسمت الف را با استفاده از distance-weighted voting تکرار کرده و نتایج را مقایسه کنید.

سوال ۵.

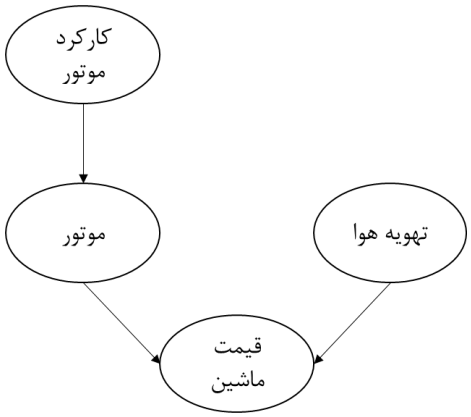
در ارتباط با ماشین های بردار پشتیبان به سوالات زیر پاسخ دهید.

الف) با توجه به SVM دو کلاسه شرح داده شده در کلاس، یک سناریو برای SVM چند کلاسه(مثالْ با m کلاس) ارائه دهید.

ب) منظور از مدل با حاشیه سخت چیست؟

سوال ۶.

شکل زیر شبکه بیزی مربوط به مجموعه داده درون جدول را نشان می دهد.



کارکرد موتور	موتور	تهویه هوا	تعداد نمونه ها با قیمت ماشین = زیاد	تعداد نمونه ها با قیمت ماشین = کم
زیاد	خوب	سالم	۳	۴
زیاد	خوب	خراب	۱	۲
زیاد	بد	سالم	۱	۵
زیاد	بد	خراب	۰	۴
کم	خوب	سالم	۹	۰
کم	خوب	خراب	۵	۱
کم	بد	سالم	۱	۲
کم	بد	خراب	۰	۲

الف) جدول احتمال را برای هر گره درون شبکه رسم کنید.

ب) از شبکه بیزی برای محاسبه عبارت زیر استفاده نمایید.

P (بد = موتور، خراب = تهویه هوا)

سوال ۷.

در یک انتخابات، Nنامزد در حال رقابت با یکدیگر هستند و مردم به هر یک از نامزدها رای می‌دهند. رای دهندگان در هنگام رای دادن با یکدیگر ارتباط برقرار نمی‌کنند.
کدام یک از روش‌های ensemble بیشتر شبیه انتخاباتی که در بالا مطرح شد عمل می‌کند؟ توضیح دهید.