به نام خدا

دانشگاه صنعتی امیرکبیر
(پلی تکنیک تهران)
دانشکده مهندسی کامپیوتر و فناوری اطلاعات

تمرین سری دوم داده کاوی – بخش تئوری

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد و در صورت مشاهده هرگونه تقلب نمره صفر برای کل تمرین منظور خواهد شد.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- گزارش تمرین خود را در قالب یک فایل PDF با نام **«HW2_StudentNumber.pdf»** در سایت درس در مهلت معین بارگذاری نمایید.
- در صورت داشتن اشکال می‌توانید از طریق ایمیل **datamining.fall2020@gmail.com** با تدریس‌یاران درس در ارتباط باشید.
- همچنین لازم بذکر است که اگر مواردی در کلاس تدریس نشده انتظار می رود که خود دانشجویان جستجو کنند و انجام دهند.

نیم‌سال اول ۰۰-۹۹

سوال ۱- مجموعه داده زیر را برای یک مسئله دو کلاسه در نظر بگیرید.

شماره داده	A	B	برچسب کلاس
۱	T	F	+
۲	T	T	+
۳	T	T	+
۴	T	F	-
۵	T	T	+
۶	F	F	-
۷	F	F	-
۸	F	F	-
۹	T	T	-
۱۰	T	F	-

الف) با محاسبه بهره اطلاعاتی[۱] موقع جداسازی A و B ، کدام ویژگی باید توسط درخت تصمیم انتخاب شود؟

ب) با محاسبه Gini Index موقع جداسازی A و B ، کدام ویژگی باید توسط درخت تصمیم انتخاب شود؟

ج) همانطور که در درس دیده اید آنتروپی و Gini index هر دو در بازه [0, 0.5] بصورت پیوسته افزایش می یابند و در بازه [0.5, 1] کاهش پیدا می کنند. آیا ممکن است که بهره اطلاعاتی و Gini index ویژگی های مختلفی را ترجیح دهند؟ دلیل خود را توضیح دهید.

سوال ۲- مجموعه داده زیر را بگیرید.

شماره داده	A	B	C	برچسب کلاس

[۱] Information gain

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	+

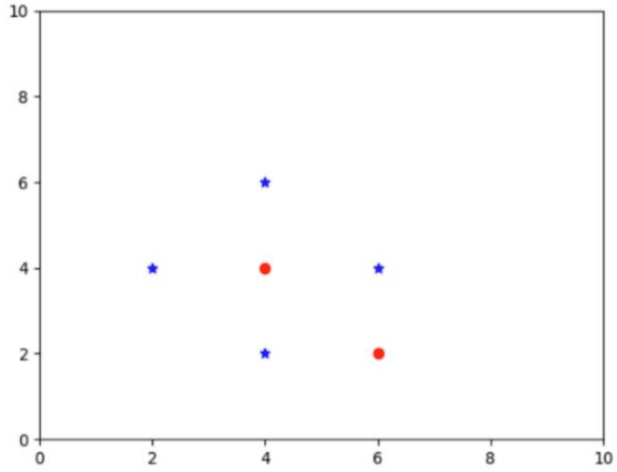
الف) احتمالات شرطی برای $P(A|+)$ ، $P(B|+)$، $P(C|+)$، $P(A|-)$، $P(B|-)$ و $P(C|-)$ را بصورت تخمینی محاسبه کنید.

ب) با استفاده از محاسبات قسمت قبل برچسب داده $(A = 0, B = 1, C = 0)$ را با استفاده از روش بیض ساده[۲] پیش بینی کنید.

سوال ۳- توضیح دهید که در چه مواقعی Accuracy معیار خوبی برای سنجش classifier نیست و معیار ارزیابی جایگزین در این مورد را پیشنهاد دهید.

سوال ۴- داده های نمایش داده شده ی زیر را در نظر گرفته و به سوالات پاسخ دهید.

[۲] Naïve bayes



الف) با استفاده از روش KNN و در حالت K = 1، مرزهای تصمیم گیری را برای این مجموعه داده مشخص نمایید. معیار فاصله را اقلیدسی در نظر گرفته و روش کار خود را توضیح دهید.

ب) با استفاده از روش NN۱ و با در نظر گرفتن معیار اقلیدسی، نقطه (۸, ۸) را به کدام کلاس نسبت میدهید؟

ج) آیا می توان از الگوریتم KNN برای مساله رگرسیون استفاده کرد؟ توضیح دهید.

د) در حالت کلی برای این الگوریتم مقدار K چگونه باید تعیین گردد؟

ه) آیا استفاده از KNN بر روی دیتاست های بزرگ پیشنهاد می شود؟ چرا؟

و) پیچیدگی زمانی KNN در حالت آموزش و آزمایش را با یکدیگر مقایسه نمایید.

ی) تفاوت معیارهای فاصله اقلیدسی و منهتن را بیان نمایید.

سوال ۵- Cross-validation چیست و چه زمانی استفاده می شود؟ سه دسته کلی آن را بیان نموده و تفاوتشان را توضیح دهید. بیان نمایید که مقادیر مختلف K در K-fold cross validation چه تاثیری بر بایاس/واریانس و پیچیدگی زمانی در حالت کلی خواهد داشت؟