

بخش اول: پرسش‌های تشریحی

سوال (۱) الگوریتم k -نزدیک‌ترین همسایه

الف) باتوجه به نحوه قرارگیری داده‌ها در شکل زیر، در صورت استفاده از KNN برای دسته بندی و روش تخمین خطای $LOOCV^5$ ، با ذکر دلیل توضیح دهید که چه مقدار k برای این مسئله کم‌ترین میزان خطا را خواهد داشت.

ب) برای یافتن بهترین مقدار هایپرپارامتر k در الگوریتم KNN چه روشی را پیشنهاد می‌کنید؟

+	+	-	-
	-		-
+	+	-	-

سوال (۲) هرکدام از الگوریتم‌های KNN و درخت تصمیم را از لحاظ Generative و یا Discriminative بودن بررسی کنید.

سوال (۳) اگر σ تابع سیگموئید باشد،

الف) ثابت کنید

$$\frac{d\sigma(a)}{da} = \sigma(a)(1 - \sigma(a))$$

ب) با توجه به اینکه در رگرسیون لاجستیک احتمال درست نمایی به صورت $p(C_1|x) = \sigma(w^T x)$ فرض می‌شود، عبارت منفی لگاریتم درست‌نمایی را برای مجموعه داده‌ی

$$(x^{(1)}, y^{(1)}) \dots (x^{(n)}, y^{(n)})$$

بدست آورید.

ج) نشان دهید با گرادینان گیری از عبارت قبل نسبت به w به

$$\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)}) x^{(i)}$$

می‌رسیم که $\hat{y}^{(i)} = \sigma(w^T x^{(i)})$.

د) در صورتی که مجموعه داده های دو کلاس خطی جداپذیر باشند چگونه بیش برآزشی رخ می‌دهد؟

سوال (۴) داده‌های آموزشی که در ذیل آمده است مربوط به یک فروشگاه بزرگ می‌باشد، که بخش کوچکی از آن در قالب جدولی آورده شده است. در این جدول ستون آخر خرید یا عدم خرید جنس مورد نظر را توسط فرد مشخص می‌کند.

الف) با استفاده از دسته بند بیز ساده مشخص کنید افراد با مشخصات زیر جنس موردنظر را خریداری میکند یا خیر؟!

- X1 = (age = youth, income = high, student = yes, credit = fair)
- X2 = (age = senior, income = low, student = no, credit = excellent)
- X3 = (age = middle-aged, income = medium, student = no, credit = fair)

ب) با توجه به ویژگی آنالیز و بهره اطلاعات درخت تصمیم بهینه را برای این مجموعه داده بیابید.

age	income	student	credit	Buy
Youth	High	No	Fair	NO
Youth	High	No	Excellent	NO
Middle	High	No	Fair	YES
Senior	Medium	No	Fair	YES
Senior	Low	Yes	Fair	YES
Senior	Low	Yes	Excellent	NO
Middle	Low	Yes	Excellent	YES
Youth	Medium	No	Fair	NO
Youth	Low	Yes	Fair	YES
Senior	Medium	Yes	Fair	YES
Youth	Medium	Yes	Excellent	YES
Middle	Medium	No	Excellent	YES
Middle	High	Yes	Fair	YES
Senior	Medium	No	Excellent	NO

سوال ۵) Weka مجموعه ای از الگوریتم های یادگیری ماشین برای تسک های داده کاوی است. الگوریتم ها می توانند مستقیماً روی یک مجموعه داده اعمال شوند یا از کد جاوا خود فراخوانی شوند. Weka حاوی ابزارهایی برای پیش پردازش داده ها، طبقه بندی، رگرسیون، خوشه بندی و ... است.

در این سوال قسمت های زیر را با استفاده از ابزار وکا انجام دهید.

الف) از دیتاست labor استفاده کرده و آن را به کمک الگوریتم درخت تصمیم دسته بندی کنید.

ب) بعد از دسته بندی داده ها ماتریس درهم ریختگی آن را گزارش کنید.

ج) بار دیگر قسمت الف را در حالتی که از هرس درخت استفاده نمی شود مجدداً انجام دهید.

د) درخت تصمیم هر دو قسمت الف و ج را رسم کرده و با یکدیگر مقایسه کنید.

در صورت نیاز برای یادگیری و بررسی بیشتر این ابزار، می توانید از [ویدیوی](#) برندون واینبرگ استفاده کنید.

بخش دوم: پیاده سازی

سوال اول:

در این سوال به دنبال پیش بینی بیمارانی قلبی در بیماران با استفاده از دیتاستی که لینک آن در زیر قرار داده شده است می باشیم. ابتدا اطلاعات مربوط به ویژگی های مختلف را بررسی کرده و سپس با استفاده از دیتاست بخش های زیر را انجام دهید.

الف) با فرض استقلال تمامی ویژگی های گسسته، و فرض توزیع نرمال چندمتغیری برای ویژگی های پیوسته یه دیتاست یک دسته بند بیز را آموزش داده و دقت آن را برای هربخش از دیتاست گزارش کنید.

ب) با فرض استقلال تمامی ویژگی ها یک دسته بند Naïve Bayes آموزش دهید و دقت آن را برای هربخش از دیتاست گزارش کنید.

ج) دسته بند قسمت قبل را یک بار با حذف ویژگی chol و یک بار با حذف ویژگی oldpeak آموزش داده و دقت آن ها را مقایسه کنید.

Dataset link: <https://www.kaggle.com/johnsmith88/heart-disease-dataset>

سوال دوم:

DecisionTreeClassifier یک کلاس است که قادر به انجام طبقه بندی چند کلاسه بر روی یک مجموعه داده است. همانند سایر طبقه بندی کننده ها، DecisionTreeClassifier دو آرایه ورودی می گیرد: یک آرایه X ، پراکنده یا متراکم، به شکل ($n_samples$ ، $n_features$) که نمونه های آموزشی را نگه می دارد، و یک آرایه Y از مقادیر صحیح، به شکل ($n_samples$) که برچسب های کلاس را برای نمونه های آموزشی نگه می دارد.

الف) مجموعه داده مورد نیاز برای این سوال در فایل تمرین پیوست شده است. در این دیتاست برخی از ویژگی های مسافران کشتی تایتانیک و در نهایت این که آیا از حادثه تایتانیک جان سالم به در برده اند یا خیر، گردآوری شده است. در این دیتاست مقادیر نامعلومی موجود می باشد. در ابتدا برای پیاده سازی روشی را برای حل این مشکل بیابید و شرح دهید.

ب) درخت تصمیم بهینه را با استفاده از این دیتاست آموزش دهید و سپس آن ها را دسته بندی کنید.

ج) و نتایج دسته بندی را در قالب یک فایل گزارش کنید.

سوال سوم:

در این سوال باید روش k نزدیک ترین همسایه (KNN) را بر روی دیتاست Iris بکار بگیرید.

الف) بهترین مقدار k را برای این دیتاست گزارش کنید. برای یافتن بهترین مقدار k و همچنین کمترین خطا از کراس ولیدیشن با ۱۰ فولد استفاده کنید.

ب) خطای RMSE، MSE و MAE را برای مجموعه داده ی تست و آموزش گزارش کنید.

دیتاست این سوال را می توانید از لینک زیر دانلود کنید.

<https://archive.ics.uci.edu/ml/datasets/iris>

پایدار باشید

نکات مربوط به تحویل تمرین

- کدهای خود را ترجیحا به زبان پایتون و در محیط jupyter پیاده‌سازی کنید.
- نظم و خوانایی در نوشتن گزارش و کدها از اهمیت بالایی برخوردار است. کدهای خود را تا حد امکان کامنت‌گذاری کنید.
- در پیاده‌سازی بخش‌های مختلف امکان استفاده از کتابخانه‌های آماده مربوط به الگوریتم‌های یادگیری ماشین را به طور کلی ندرید. مگر در مواردی که در صورت سوال ذکر شده باشد.
- برای خواندن داده‌ها می‌توانید از کتابخانه pandas و برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. برای محاسبه معیارهای ارزیابی مانند دقت، ماتریس درهم‌ریختگی و تقسیم داده‌ها به مجموعه‌های آموزش و آزمون نیز استفاده از کتابخانه آماده مجاز است.
- در صورتی که داده‌ها را به دسته‌های آموزشی، تست (ویا validation) تقسیم می‌کنید، درصد هر کدام را در گزارش ذکر کنید.
- فایل‌های کد و گزارش خود را در قالب یک فایل فشرده با فرمت HW02_StdNumber.zip که StdNumber شماره دانشجویی شماست، در سامانه بارگذاری کنید.
- سوالات ستاره دار(*) دارای نمره اضافی بر تمرین است.
- مطابق قوانین دانشگاه هر گونه کپی‌برداری ممنوع است و در صورت مشاهده، نمره هر دو طرف صفر داده می‌شود.
- در صورت وجود هر گونه سوال یا ابهامی با ایمیل درس در تماس باشید:

ml.ce.aut@gmail.com