بخش اول: درخت تصمیم

دیتاست

دیتاست بخش اول Hayes-Roth نام دارد. این دیتاست شامل اطلاعات مختلف مانند نام، نوع سرگرمی، سن، مقطع تحصیلی، وضعیت ازدواج و نام کلاس ۱۳۲ نفر است. اطلاعات این دیتاست از حالت categorical به عددی تبدیل شدند. مجموعه‌ی آموزشی آن به شرح مقابل است:

	0	1	2	3	4	5
0	92	2	1	1	2	1
1	10	2	1	3	2	2
2	83	3	1	4	1	3
3	61	2	4	2	2	3
4	107	1	1	3	4	3
...	...	...	...	...	...	...
127	44	1	1	4	3	3
128	40	2	1	2	1	1
129	90	1	2	1	2	2
130	21	1	2	2	1	2
131	9	3	1	1	2	1

در توضیحات دیتاست ذکر شده است که هر ستون چه معنایی دارد:

Attribute Information:

-- 0. name: distinct for each instance and represented numerically
-- 1. hobby: nominal values ranging between 1 and 3
-- 2. age: nominal values ranging between 1 and 4
-- 3. educational level: nominal values ranging between 1 and 4
-- 4. marital status: nominal values ranging between 1 and 4
-- 5. class: nominal value between 1 and 3

بنابراین اگر برای مثال، نام ستون‌ها را در داده‌های آموزش تغییر دهید، مجموعه داده‌ی زیر حاصل می‌شود:

	name	hobby	age	educational level	marital status	class
0	92	2	1	1	2	1
1	10	2	1	3	2	2
2	83	3	1	4	1	3
3	61	2	4	2	2	3
4	107	1	1	3	4	3
...	...	...	...	...	...	...
127	44	1	1	4	3	3
128	40	2	1	2	1	1
129	90	1	2	1	2	2
130	21	1	2	2	1	2
131	9	3	1	1	2	1

132 rows × 6 columns

توجه: این مجموعه داده در دو گروه آموزش و آزمون قرار دارد، که ستون آخر هر دو گروه، متغیر y آن گروه را تشکیل می‌دهد. دیتاست در فایل تمرین شما قرار داده شده است. جهت دریافت اطلاعات بیشتر می‌توانید به این لینک مراجعه کنید.

عملیات

الف) الگوریتم درخت تصمیم را بر روی مجموعه داده‌ی ذکر شده اجرا کنید. دقت و خطای آموزش و تست را گزارش کنید.

ب) با استفاده از روش K-Fold Cross Validation، عمق بهینه‌ی درخت را محاسبه کنید.

ج) الگوریتم درخت تصمیم را این بار با عمق بهینه‌ی حاصل از قسمت ب بر روی دیتاست پیاده کنید. خروجی را با قسمت الف مقایسه کرده و نتایج را تفسیر کنید.

توجه: پیش‌پردازشی که باید بر روی دیتاست انجام گیرد، تغییر نوع ستون‌هایی مانند hobby، educational level و marital status به str است.

https://archive.ics.uci.edu/ml/datasets/Hayes-Roth

بخش دوم: KNN

دیتاست

دیتاست بخش دوم، بازماندگان Haberman نام دارد. مجموعه داده شامل مواردی از مطالعه ای است که بین سال های ۱۹۵۸ و ۱۹۷۰ در بیمارستان بیلینگز دانشگاه شیکاگو در مورد بقای بیمارانی که به دلیل سرطان سینه تحت عمل جراحی قرار گرفته بودند، انجام شد. در توضیحات دیتاست ذکر شده است که هر ستون چه معنایی دارد:

Attribute Information:

1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
-- 1 = the patient survived 5 years or longer
-- 2 = the patient died within 5 year

بنابراین اگر نام ستون‌ها را تغییر دهید، مجموعه داده‌ی زیر حاصل می‌شود:

	Age	Year of operation	Number of positive cases	Survival status
0	30	64	1	1
1	30	62	3	1
2	30	65	0	1
3	31	59	2	1
4	31	65	4	1
...	...	...	...	...
301	75	62	1	1
302	76	67	0	1
303	77	65	3	1
304	78	65	1	2
305	83	58	2	2

306 rows × 4 columns

توجه: این مجموعه داده به صورت یک فایل در اختیار شما قرار داده شده است. بنابراین لازم عملیات split کردن با نسبت ۸۰ به ۲۰ را انجام دهید. لازم به ذکر است، ستون آخر این مجموعه را نیز می‌توانید به عنوان y در نظر بگیرید. برای کسب اطلاعات بیشتر در رابطه با دیتاست، به این لینک مراجعه کنید.

https://archive.ics.uci.edu/ml/datasets/Haberman's+Survival

عملیات

الف) به ازای مقادیر از ۱ تا ۴۰ برای k ،الگوریتم KNN را روی دیتاست داده شده با استفاده از متر فاصله‌ی اقلیدسی اجرا کرده و نمودار دقت و خطای تست و آموزش را به ازای k های مختلف رسم کنید. (plot کنید)

ب) تغییرات حاصل از اجرای الگوریتم مقادیر کوچک k و مقادیر بزرگ برای k را توضیح دهید. به این منظور برای مقادیر k = [1, 5, 20, 40] نمودار scatter plot را برای دو کلاس age و Number of positive cases رسم کنید و نشان دهید با افزایش k، داده‌ها بیشتر تمایل پیوستن به کدام کلاس را دارند؟

راهنمایی: به ازای مجموعه kهای داده شده، Knn را بر روی داده‌های آموزش فیت کنید و y_pred را با استفاده از مدلی که ساختید بدست آورید. سپس scatter plot را با متغیرهای x1 = age و x2 = Number of positive cases و y = y_pred رسم کنید.

ج) با استفاده از k-fold cross validation ، مقدار مناسب برای k را بدست آورید.

توجه: تنها پردازشی که باید بر روی دیتاست انجام داد، نرمالایز کردن متغیر x است. می‌توانید از روش میانگین-واریانس استفاده کنید.

توجه: تنها پردازشی که باید بر روی دیتاست انجام داد، نرمالایز کردن متغیر x است. می‌توانید از روش میانگین-واریانس استفاده کنید.

$$x := \frac{x - \mu}{\sigma}$$

نکات مهم

- جهت انجام بخش پیاده سازی، از Jupyter استفاده کنید و فایل نهایی را با پسوند. ipynb آپلود کنید. همچنین شماره دانشجوی خود را به عنوان نام فایل در نظر بگیرید.
- نکته مهم در گزارش نویسی و سوال تشریحی روشن بودن پاسخ می‌باشد نه حجم زیاد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، و پاسخ نهایی را به صورت واضح بیان کنید.
- هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب می‌باشد و کل نمره تمرین صفر می‌باشد. (**می‌توانید از اینترنت به عنوان منبع کمکی هم در سوالات تشریحی و هم در سوالات پیاده سازی استفاده کنید، اما کپی برداری ممنوع می باشد و نمره‌ی صفر تعلق می‌گیرد)**