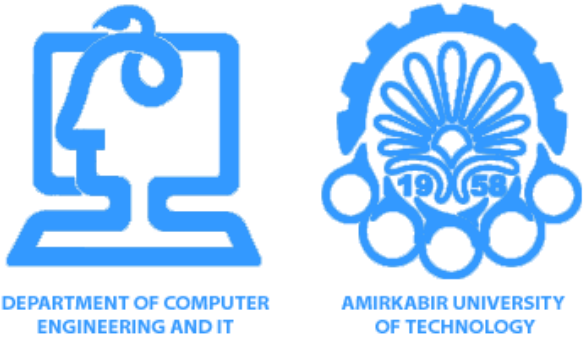


تمرین سری دوم درس یادگیری ماشین

دکتر ناظرفرد

پاییز ۹۸

- تمامی مستندات خود شامل گزارش و کدهای خود را در یک فایل فشرده با فرمت زیپ ذخیره کرده و با عنوان #studentId_HW.zip بارگزاری نمایید.
- مهلت انجام این تمرین تا ساعت ۲۳:۵۵ روز ۲۹ آبان می‌باشد و به هیچ وجه تمدید نمی‌شود.

تمرینات تشریحی:

بخش اول) در این بخش به بررسی درخت تصمیم و جنگل تصادفی پرداخته می‌شود. لطفا به سوالات زیر پاسخ دهید.

سوال اول) هرس درخت تصمیم چه تاثیری بر بیش برازش دارد؟ این هرس چه زمانی باید انجام شود؟ توضیح دهید.

سوال دوم) با توجه به جدول زیر به سوالات زیر پاسخ دهید.

	ویژگی A	ویژگی B	ویژگی C	کلاس
نمونه اول	۱	۱	۱	X
نمونه دوم	۱	۱	۰	X
نمونه سوم	۰	۰	۱	Y
نمونه چهارم	۱	۰	۰	Y

الف) در ابتدا با توجه به ویژگی‌های موجود در جدول بیان نمایید که کدام ویژگی برای کلاس‌بندی مناسب می‌باشد.

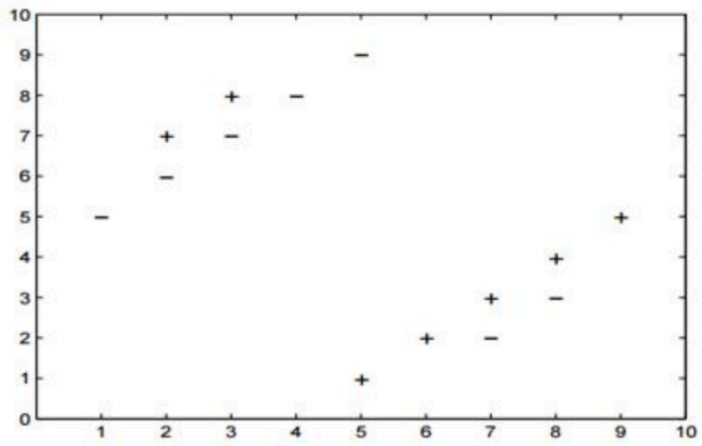
ب) سپس با استفاده از آنتروپی[1] و بهره اطلاعات[2] بهترین ویژگی را پیدا بکنید.

پ) درخت تصمیم را با توجه به ویژگی‌های بخش قبل رسم نمایید.

سوال سوم) در رابطه با جنگل تصادفی و نقطه قوت آن‌ها در مقایسه با درخت‌های تصمیم توضیح بدهید. منبع را نیز ذکر نمایید(منبع مورد استفاده لازم است که مقاله باشد).

بخش دوم) در این بخش به بررسی الگوریتم کلاس‌بندی KNN پرداخته می‌شود.

سوال اول) با توجه به شکل زیر به سوالات پاسخ دهید.



الف) با توجه به شکل زیر، بهترین مقدار k را برای الگوریتم KNN زمانی که از روش LOOCV[3] استفاده شود را پیدا کنید.

ب) دقت این الگوریتم را برای مقدار بدست آمده در قسمت الف بدست آورید.

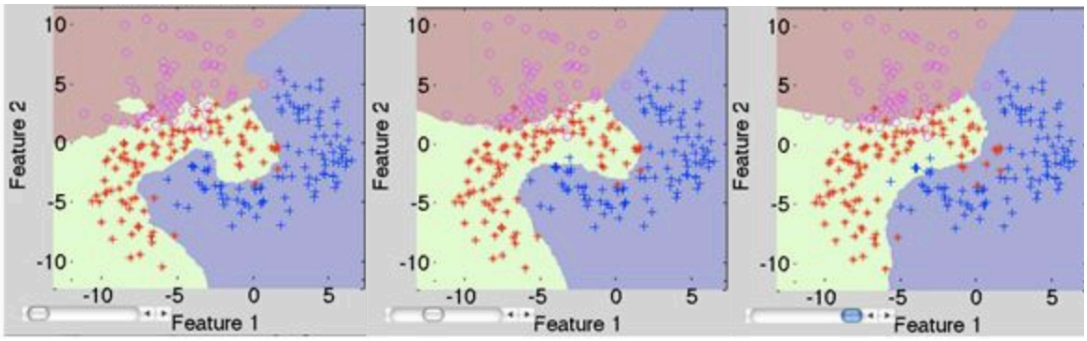
---

[1] Entropy
[2] Information gain
[3] Leave One Out Cross Validation

سوال دوم) افزایش مقدار k در الگوریتم KNN چه تاثیری بر برازش الگوریتم دارد. توضیح دهید.

سوال سوم) برای کاهش تاثیر نویز بر دقت الگوریتم چه تغییری در الگوریتم KNN می‌توان اعمال کرد؟ توضیح دهید.

سوال چهارم) در شکل زیر الگوریتم KNN با مقادیر مختلف k بر روی یک مجموعه داده تست شده است و ناحیه بندی شده است. با توجه به مرزهای ناحیه‌بندی در این اشکال، مقادیر K را مقایسه نمایید.



بخش سوم) آشنایی با ابزار وکا.

سوال اول) ابزار وکا را دانلود نمایید و با استفاده از مجموعه داده iris عملیات ریز زا انجام دهید.

الف) ابتدا مجموعه داده iris را از میان مجموعه داده‌های خود وکا بارگزاری نمایید.

ب) سپس از میان ویژگی‌های مختلف این مجموعه داده، sepalwidth و petalwidth را برای کلاس‌بندی انتخاب نمایید.

ج) این مجموعه داده را با استفاده از این دو ویژگی نمایش دهید.

د) الگوریتم KNN را در بخش کلاس‌بندی انتخاب نمایید و به ازای مقادیر ۱ تا ۵ دقت الگوریتم را بررسی و ثبت نمایید. لازم به ذکر است که این الگوریتم با عنوان IBK نام‌گذاری شده است.

ه) بار دیگر عملیات الف تا د را برای حالتی که بر روی مجموعه داده ۱۵ درصد نویز اضافه شده است، انجام دهید (اضافه کردن نویز در این ابزار در بخش پیش‌پردازش می‌باشد).

و) دقت الگوریتم به‌دست آمده برای هر دو حالت به ازای مقادیر مختلف k را با یکدیگر مقایسه نمایید و تاثیر پارامتر k در مقابله با نویز را توضیح دهید.

سوال دوم) مجموعه داده vote را بارگزاری نموده و به سوالات زیر پاسخ دهید.

الف) با استفاده از درخت الگوریتم درخت تصمیم، این مجموعه داده را کلاس‌بندی نمایید و ماتریس پریشانی آن را نشان دهید.

ب) بار دیگر قسمت الف را در حالتی که از هرس درخت استفاده نمی‌شود انجام دهید.

ج) در هر دو حالت درخت حاصل از کلاس‌بندی را نمایش دهید.

د) نتایج حاصل ازین دو بخش را با یکدیگر مقایسه نمایید و بیان کنید استفاده از هرس در درخت تصمیم چه سودی دارد.

تمرینات برنامه‌نویسی:

در انجام تمرینات زیر به نکته‌های زیر توجه نمایید:

- تمرین بدون گزارش فاقد ارزش می‌باشد و نمره‌ای به آن تعلق نمی‌یابد.
- گذاشتن کامنت برای کدها الزامی می‌باشد.
- گذاشتن عنوان برای نمودارها و برچسب‌گذاری محورهای نمودار الزامی می‌باشد.
- هر سوال باید در فایل جداگانه‌ی پایتون قرار داده شود.
- برای پیاده‌سازی اجازه استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را ندارید مگر در حالتی که در صورت سوال ذکر شده باشد.
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده نمایید. هم‌چنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده نمایید.
- برای بهبود سرعت برنامه توصیه می‌شود که تا حد ممکن از عملیات ماتریسی استفاده نمایید و به کار بردن حلقه فور بپرهیزید.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس پریشانی می‌توانید از کتابخانه استفاده نمایید ولی برای پیاده‌سازی آن‌ها نمره تشویقی تعلق خواهد گرفت.
- تا حد ممکن سعی کنید اصول لازم برای گزارش مهندسی را رعایت نمایید. در این راستا برای گزارش‌هایی با ظاهر عالی نمره تشویقی در نظر گرفته شده است.
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر دو طرف صفر در نظر گرفته می‌شود.

سوال اول) هدف ازین سوال آشنایی با الگوریتم KNN برای مسایل کلاس‌بندی می‌باشد. مجموعه داده مشخص شده[4] را دانلود نمایید و ازآن برای پاسخ دادن به سوالات زیر استفاده نمایید.

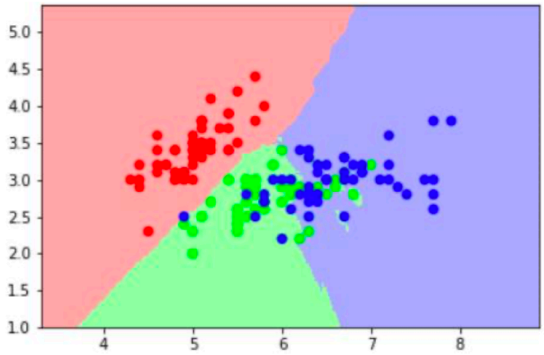
الف) الگوریتم KNN را با استفاده از معیار فاصله اقلیدسی پیاده‌سازی نمایید. در این پیاده‌سازی از تمامی ویژگی‌های این مجموعه داده که شامل ۱۳ ویژگی است، استفاده نمایید. این مجموعه داده شامل ۳۰۳ شخص می‌باشد که به دو کلاس بیمار و سالم تقسیم شده‌اند. مقادیر ۱ تا ۷، ۱۰ و ۱۵ را برای این الگوریتم استفاده نمایید و بهترین مقدار k را بدست آورید. لازم به ذکر است که در این بخش داده‌ها را به دو بخش یادگیری و تست با نسبت ۲ به ۱ تقسیم کنید. در نهایت دقت الگوریتم و ماتریس پریشانی را برای مجموعه داده آموزش و تست گزارش نمایید.

ب) قسمت الف را فقط برای بهترین مقدار K ولی بدون نرمال‌سازی ویژگی‌ها انجام دهید و نتیجه به‌دست‌آمده را با قسمت قبلی مقایسه نمایید. به صورت کلی استفاده از نرمال‌سازی چه تاثیری بر روی کلاس‌بندی می‌تواند داشته باشد؟ توضیح دهید.

---

[4] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data

ج) نمودار ویژگی‌های سن[5] بر بیشترین ضربان[6] برای کلاس‌های مختلف رسم نمایید و مرزهای ناحیه‌ها که الگوریتم KNN مشخص می‌نماید را با رنگ‌های مختلف مانند شکل زیر نمایش دهید. در این کلاس‌بندی فقط از این دو ویژگی استفاده نمایید (برای این تقسیم‌بندی از مجموعه داده یادگیری استفاده نمایید که با نسبت ۱ به ۲ تقسیم شده‌اند)[7].



د) قسمت الف را بار دیگر با معیارهای فاصله زیر انجام دهید.

1. Manhattan 2. Chebyshev

ه) در این قسمت مجموعه داده یادگیری را با استفاده از روش k_fold cross validation کلاس‌بندی نمایید و با انتخاب دلخواه k بهترین پارامترهای ممکن (اندازه همسایه‌ها و معیار فاصله) را برای این مجموعه داده بیابید. برای مدل به‌دست‌آمده ماتریس پریشانی و دقت داده‌های تست را گزارش نمایید.

سوال دوم) مجموعه داده mnist را در پایتون بارگزاری نمایید. می‌توانید از لینک زیر دانلود نمایید:

http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits

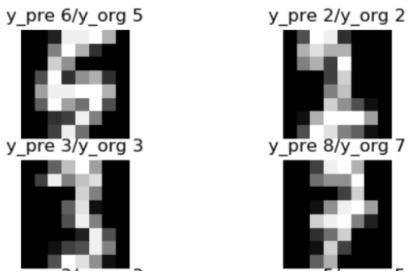
یا اینکه از طریق دستورات زیر آن را بارگزاری نمایید:

```
from sklearn import datasets
mnist = datasets.load_digits()
```

الف) این مجموعه داده را به سه دسته آموزش، ارزیابی و تست با نسبت ۳، ۱، ۱ تقسیم نمایید. پس از تقسیم‌بندی از الگوریتم KNN برای کلاس‌بندی استفاده نمایید. از مجموعه داده ارزیابی برای یافتن بهترین مقادیر k و بهترین معیار فاصله استفاده نمایید. پس از یافتن بهترین مدل دقت آموزش، ارزیابی و تست و ماتریس پریشانی را برای داده‌های تست گزارش نمایید. در نهایت ۱۵ نمونه از داده‌های تست را به همراه برچسب صحیح و پیش‌بینی شده مانند شکل زیر نمایش دهید.

---

[5] Age
[6] Maximum heart rate
7 https://scikit-learn.org/stable/auto_examples/neighbors/plot_classification.html



ب) بار دیگر عملیات قسمت الف را با استفاده از کتابخانه آماده انجام دهید. توصیه می‌شود در این بخش از کتابخانه sklearn استفاده شود.

سوال سوم) در این تمرین از الگوریتم KNN برای رگرسیون استفاده نمایید. مجموعه داده regression که در کنار این فایل قرار داده شده است را به دو بخش آموزش و تست با نسبت ۲ به ۱ تقسیم نمایید. سپس بهترین مقدار k را با آزمون خطا به‌دست آورید. خطای MSE را برای این مدل برای هر دو مجموعه داده تست و یادگیری گزارش کنید. مجموعه داده آموزش و تست را با رنگ‌های متفاوت در نموداری نشان دهید (کمترین خطای MSE برای مجموعه داده تست شامل نمره تشویقی خواهد بود).

با آرزوی موفقیت