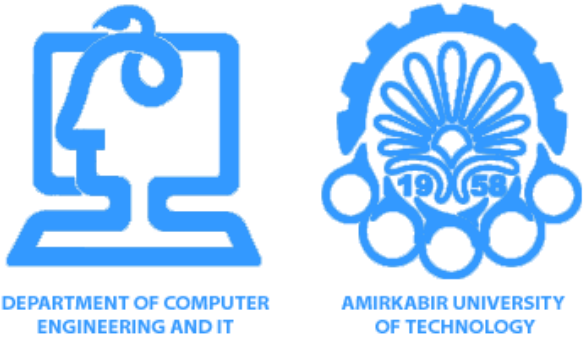


تمرین دوم درس یادگیری ماشین

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 ۳۰ آبان** می‌باشد **و به هیچ وجه تمدید نمی‌شود**.

تمرینات تشریحی: (۳۰٪)

بخش اول- در این بخش به بررسی درخت تصمیم[1] و جنگل تصادفی[2] پرداخته می‌شود. لطفا به سوالات زیر پاسخ دهید. (۸٪)

۱- هرس[3] درخت تصمیم چه تاثیری بر بیش برازش[4] دارد؟ این هرس چه زمانی باید انجام شود؟ توضیح دهید.

۲- 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 دسته‌بندی مناسب می‌باشد.

ب- سپس با استفاده از آنتروپی[5] و بهره اطلاعات[6] بهترین ویژگی را پیدا کنید.

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

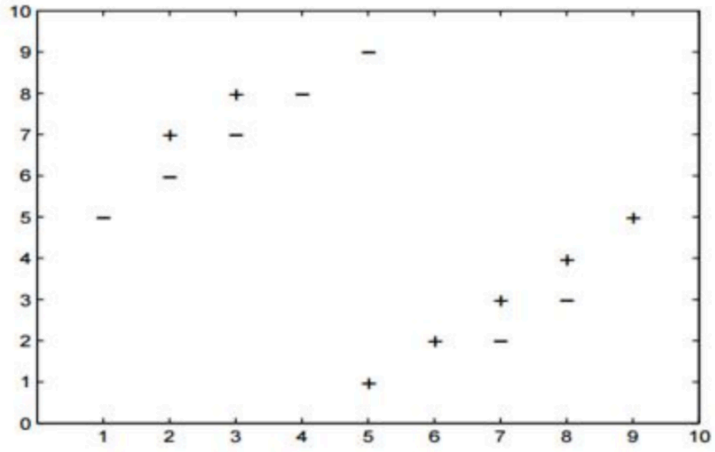
۳- در رابطه با جنگل تصادفی و نقاط قوت آن در مقایسه با درخت تصمیم توضیح دهید. منبع را نیز ذکر نمایید (منبع مورد استفاده لازم است که مقاله باشد).

---

[1] Decision tree
[2] Random forest
[3] Prunning
[4] Overfit
[5] Entropy
[6] Information gain

بخش دوم- در این بخش به بررسی الگوریتم دسته‌بندی[7] K نزدیک‌ترین همسایه[8] پرداخته می‌شود. (۸٪)

۱- با توجه به شکل زیر به سوالات پاسخ دهید.



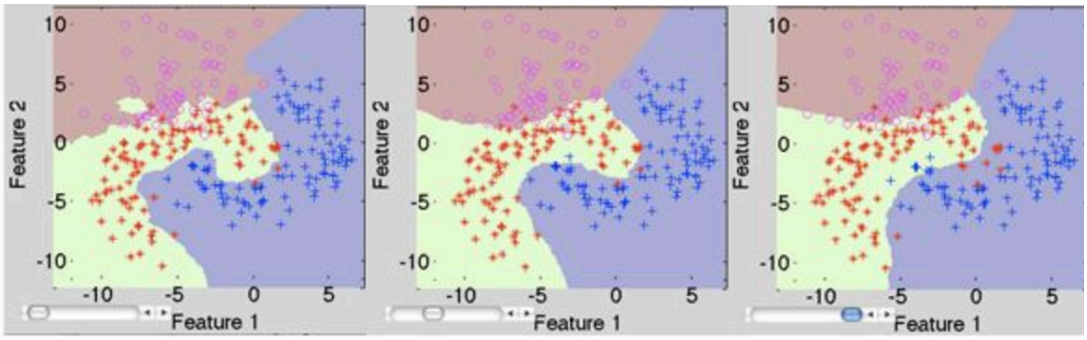
الف- بهترین مقدار K را برای الگوریتم K نزدیک‌ترین همسایه زمانی که از روش LOOCV[9] استفاده شود را پیدا کنید.

ب- دقت[10] این الگوریتم را برای مقدار K به‌دست‌آمده در قسمت الف محاسبه کنید.

۲- افزایش مقدار K در الگوریتم K نزدیک‌ترین همسایه چه تاثیری بر برازش الگوریتم دارد. توضیح دهید.

۳- برای کاهش تاثیر نویز بر دقت الگوریتم چه تغییری در الگوریتم K نزدیک‌ترین همسایه می‌توان اعمال کرد؟ توضیح دهید.

۴- در شکل زیر الگوریتم K نزدیک‌ترین همسایه با مقادیر مختلف K بر روی یک مجموعه داده تست و ناحیه‌بندی شده است. با توجه به مرزهای ناحیه‌بندی در این اشکال، مقادیر K را نسبت به یکدیگر مقایسه نمایید.



[7] Classification
[8] K-Nearest Neighbors (KNN-
[9] Leave One Out Cross Validation
[10] Accuracy

بخش سوم- آشنایی با ابزار وکا (۱۴٪)

۱- ابزار وکا را دانلود نمایید و با استفاده از مجموعه داده iris عملیات زیر را 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 دسته‌بندی انتخاب نمایید.

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

د- الگوریتم K نزدیک‌ترین همسایه را در بخش دسته‌بندی انتخاب نمایید و به ازای مقادیر ۱ تا ۵ برای K، دقت الگوریتم را بررسی و گزارش نمایید. لازم به ذکر است که این الگوریتم با عنوان IBK نام‌گذاری شده است.

هـ - 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.

۲- مجموعه داده vote را بارگزاری نموده و به سوالات زیر پاسخ دهید.

الف- با استفاده از الگوریتم درخت تصمیم، این مجموعه داده را دسته‌بندی نمایید و ماتریس پریشانی[11] 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 دسته‌بندی را نمایش دهید.

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

[11] Confusion matrix

تمرینات برنامه‌نویسی: (۱۵+۷۰٪)

در انجام تمرینات زیر به نکات زیر توجه نمایید:

- تمرین بدون گزارش فاقد ارزش می‌باشد و نمره‌ای به آن تعلق **نمی‌یابد**.
- کامنت گذاری کدها در حد **لازم و کافی** الزامی می‌باشد.
- گذاشتن عنوان[۱۲] برای نمودارها و برچسب‌گذاری[۱۳] محورهای نمودار **الزامی** می‌باشد.
- در پیاده‌سازی بخش‌های مختلف، امکان استفاده از کتابخانه‌های آماده مرتبط با الگوریتم‌های یادگیری ماشین را به طور کلی ندارید. **موارد مجاز در صورت سوال بخش‌ها ذکر شده است.**
- برای نمایش نمودارها و عملیات ماتریسی می‌توانید از کتابخانه‌های numpy و matplotlib استفاده کنید. هم‌چنین برای خواندن داده‌ها به عنوان ورودی می‌توانید از pandas استفاده کنید.
- برای بهبود سرعت برنامه توصیه می‌شود که تا حد ممکن از عملیات ماتریسی استفاده شود.
- برای محاسبه معیارهای ارزیابی مانند دقت و ماتریس پریشانی می‌توانید از کتابخانه استفاده نمایید. **درصورت پیاده‌سازی آن‌ها نمره تشویقی تعلق خواهد گرفت. (+۵)**
- 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. (+۵)**
- مطابق قوانین دانشگاه هرگونه کپی‌برداری ممنوع می‌باشد و در صورت مشاهده نمره هر **دو طرف صفر** در نظر گرفته می‌شود.

۱- هدف از این سوال آشنایی با الگوریتم K نزدیک‌ترین همسایه برای مسایل دسته‌بندی می‌باشد. مجموعه داده مشخص شده[۱۴]را دانلود نمایید و از آن برای پاسخ دادن به سوالات زیر استفاده نمایید. **(۳۰٪)**

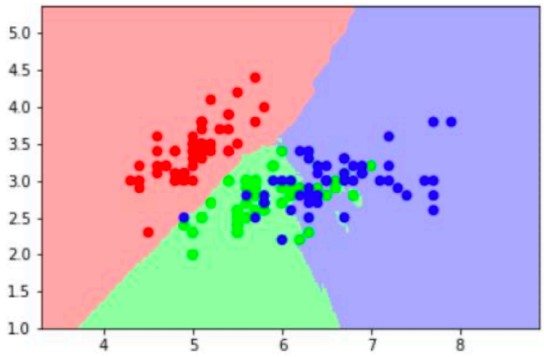
الف- الگوریتم K نزدیک‌ترین همسایه را با استفاده از معیار فاصله اقلیدسی[۱۵] پیاده‌سازی نمایید. در این پیاده‌سازی از تمامی ویژگی‌های این مجموعه داده که شامل ۱۳ ویژگی است، استفاده نمایید. این مجموعه داده شامل ۳۰۳ شخص می‌باشد که به دو کلاس بیمار و سالم تقسیم شده‌اند. مقادیر ۱ تا ۷، ۱۰ و ۱۵ را برای این الگوریتم استفاده نمایید و بهترین مقدار K را بدست آورید. لازم به ذکر است که در این بخش داده‌ها را به دو بخش یادگیری[۱۶] و آزمون[۱۷] با نسبت ۲ به ۱ تقسیم کنید. در نهایت دقت الگوریتم و ماتریس پریشانی را برای مجموعه داده آموزش و آزمون 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 دسته‌بندی می‌تواند داشته باشد؟ توضیح دهید.

---

[12] Title
[13] Labeling
[14] https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
[15] Euclidean distance
[16] Train
[17] Test

ج- نمودار ویژگی‌های سن[18] بر بیشترین ضربان[19] برای کلاس‌های مختلف رسم نمایید و مرزهای ناحیه‌ها که الگوریتم K نزدیک‌ترین همسایه مشخص می‌نماید را با رنگ‌های مختلف مانند شکل زیر نمایش دهید. در این دسته‌بندی فقط از این دو ویژگی استفاده نمایید (برای این تقسیم‌بندی از مجموعه داده یادگیری استفاده نمایید که با نسبت ۱ به ۲ تقسیم شده‌اند)[20].



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

1. Manhattan 2. Chebyshev

هـ - در این قسمت مجموعه داده یادگیری را با استفاده از روش K-fold cross validation دسته‌بندی 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.

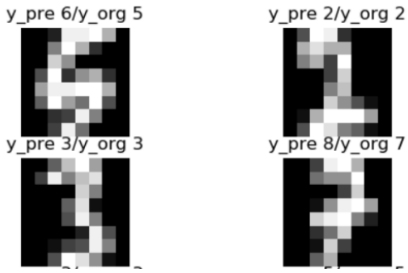
۲- مجموعه داده mnist را دانلود[21] و یا اینکه از طریق دستورات زیر آن را بارگزاری نمایید. (**۲۵٪**)

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

الف- این مجموعه داده را به سه دسته آموزش، ارزیابی[22] و آزمون با نسبت ۳، ۱، ۱ تقسیم نمایید. پس از تقسیم‌بندی از الگوریتم K نزدیک‌ترین همسایه برای دسته‌بندی 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 آزمون و ماتریس پریشانی را برای داده‌های آزمون گزارش کنید. در نهایت ۱۵ نمونه از داده‌های آزمون را به همراه برچسب صحیح و پیش‌بینی شده مانند شکل زیر نمایش دهید.

---

[18] Age
[19] Maximum heart rate
[20] https://sciKit-learn.org/stable/auto_examples/neighbors/plot_classification.html
[21] http://archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits
[22] Evaluation



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

۳- در این تمرین از الگوریتم K نزدیک‌ترین همسایه برای رگرسیون استفاده نمایید. مجموعه داده regression که در کنار این فایل قرار داده شده است را به دو بخش آموزش و آزمون 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 آزمون و آموزش گزارش کنید. مجموعه داده آموزش و آزمون را با رنگ‌های متفاوت در نموداری نمایش دهید **(کمترین خطای MSE برای مجموعه داده آزمون شامل نمره تشویقی خواهد بود)**. **(۵+۱۵٪)**

موفق باشید