

به نام خالق عقل و خرد



جلسه اول

یادگیری ماشین

نیمسال دوم ۱۴۰۴ - ۱۴۰۳

دانشکده ریاضی و آمار

دانشگاه اصفهان

اسفند ماه ۱۴۰۳

استاد درس:

دکتر فاطمه منصوری

عنوان تمرین : پیاده‌سازی k-نزدیکترین همسایه

هدف: درک سازوکار الگوریتم KNN با پیاده‌سازی آن از ابتدا، ارزیابی عملکرد آن و مقایسه با پیاده‌سازی کتابخانه.

دستورالعمل :

۱. دیتاست بیماری دیابت (diabetes.csv) را از گیت‌هاب درس دانلود کنید.
۲. مجموعه داده را در یک Dataframe پانداس بارگزاری کرده و چند داده اول آنرا مشاهده نمایید.
۳. داده‌ها را به دو دسته آموزش و تست تقسیم کنید
۴. الگوریتم KNN را پیاده‌سازی نمایید.
- برای پیاده‌سازی این الگوریتم یک کلاس به نام *my_KNN* ایجاد کنید. مقدار *k* را به عنوان attribute این کلاس قرار دهید. و متدهای زیر را به آن اضافه کنید:
 - (آ) متد *fit* را به نحوی تعریف کنید که داده‌های آموزش را به عنوان ورودی گرفته و مدل را با استفاده از آنها آموزش دهد.
 - (ب) یک متد به نام *euclidean_distance* برای محاسبه فاصله اقلیدسی بین دو داده
 - (ج) یک متد به نام *get_neighbors* برای پیدا کردن *k* نزدیک‌ترین همسایه برای یک نمونه آزمایشی داده شده
 - (د) یک متد به نام *predict* برای پیش‌بینی بر اساس رای اکثریت *k* نزدیک‌ترین همسایه داده تست داده شده
۵. دقت مدل خود را روی مجموعه داده تست بررسی کنید. (از معیار *accuracy* استفاده کنید.)
۶. معیار *F1* را برای مدل خود محاسبه کنید.
۷. بهترین مقدار *k* را برای داده‌های خود پیدا کنید و با قرار دادن *k* برابر مقدار جدید، مقدار دقت مدل را محاسبه کنید.
۸. داده‌ها را قبل از آموزش مدل نرمال کنید و نتایج حاصل از اجرای مدل روی داده‌های نرمال شده را با داده‌های نرمال نشده مقایسه کنید.
۹. در ویژگی‌هایی که مقدار ۰ برای آنها معنی ندارد مثل فشار خون، مقادیر صفر را میانگین *k* نزدیک‌ترین همسایه جایگزین کنید. و نتایج مدل را روی مجموعه داده جدید با نتایج مدل قبلی مقایسه نمایید.

۱۰. متد فاصله اقلیدسی وزن دار را پیاده سازی کرده و با استفاده از آن نمودار ROC را رسم کنید.

۱۱. گزارشی از نتایج مقایسه های خود تهیه و ارائه نمایید.