بسمه تعالی

مقدمه‌ای بر یادگیری ماشین

نیمسال اول ۹۷-۹۸

تمرین عملی سری اول

مدرس: صابر صالح

● مهلت تحویل تمرین کامپیوتری: ۱۳۹۷/۰۷/۱۷ ●

۱ تمرین عملی

۱.۱ مقدمه

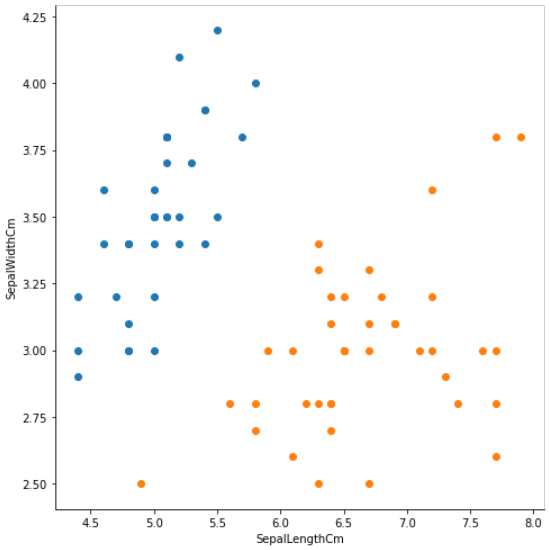
مجموعه داده‌ی Iris (گل زنبق) یکی از معروف‌ترین نمونه‌های داده در حوزه‌ی یادگیری ماشین است. این داده اولین بار توسط رونالد فیشر در یکی از مقاله‌های وی منتشر شد. در نمونه‌ی اصلی این داده، چهار ویژگی برای ۱۵۰ نمونه از ۳ نوع گل در یک مکان و زمان مشخص و با ابزار یکسان اندازه‌گیری شده که از آن به عنوان معیاری برای مقایسه‌ی بسیاری از الگوریتم‌های کلاس‌بندی در یادگیری ماشین استفاده می‌شود. ویژگی‌های اندازه‌گیری شده عبارتند از: طول کاسبرگ، عرض کاسبرگ، طول گلبرگ و عرض گلبرگ. همچنین خروجی داده، نام یکی از سه نوع گل زنبق setosa، virginica و versicolor است.

۲.۱ صورت مسئله

در این مسئله برای راحتی کار و تبدیل مسئله به طبقه بندی باینری، داده‌های مربوط به گل versicolor حذف شده است. پس داده اصلی به داده‌ای با چهار ویژگی برای ۱۰۰ نمونه از ۲ نوع گل تبدیل شده است. به طور مثال در جدول زیر، ۳ نمونه از این داده را مشاهده می‌کنید.

	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	6.2	2.8	4.8	1.8	Iris-virginica
1	4.4	2.9	1.4	0.2	Iris-setosa
2	5.0	3.6	1.4	0.2	Iris-setosa

در مسائل یادگیری با ناظر، داده‌ساختار به دو بخش (نه لزوما مساوی) train و test به منظور آموزش و تست مدل تقسیم می‌شوند. مدل شما از نمونه‌های train که دارایی ۴ ویژگی نام برده شده در کنار نام نوع گل است استفاده کرده و هنگامی که یک نمونه از داده‌های test که تنها دارای چهار ویژگی است (نه نوع گل) به این مدل داده شود، مدل شما باید بتواند نوع گل را تشخیص دهد. در این تمرین نیز داده‌ها را به دو بخش train و test تقسیم کرده، از داده‌های train استفاده کرده و مدل خود را آموزش دهید و سپس نوع گل را فقط با استفاده از ویژگی‌های داده در بخش تست تخمین بزنید. هر قدر تعداد تخمین صحیح شما روی داده‌ی تست بیشتر باشد، روش شما مناسب‌تر است. در این مسئله از ۷۰ درصد داده‌ها به عنوان train و ۳۰ درصد داده‌ها به عنوان test استفاده کنید. پس از تقسیم داده‌ها، نیاز است که ابتدا درک کلی از داده‌ها به دست آورید. از آنجا که تعداد ویژگی‌ها ۴ عدد است، ۶ شکل مانند تصویر زیر رسم کنید. این اشکال از انتخاب ۲تایی‌های ممکن از ۴ ویژگی موجود به دست خواهند آمد.



برای حل این مسئله، دو مدل را آموزش خواهیم داد. ساختار مدل اول به این ترتیب است که، برای هر نمونه‌ی test، ابتدا فاصله‌ی اقلیدسی این نمونه را از تمام نمونه‌های train اندازه گرفته و k تا از کوچکترین فاصله‌های به دست آمده را انتخاب می‌کنید. سپس داده‌های train مربوط به این k کوچک‌ترین فاصله را در نظر گرفته و بین نوع گل این k داده، آن نوع که دارای بیشترین تکرار است را انتخاب کنید. یک انتخاب خوب برای k می‌تواند ۵ باشد.

برای مدل دوم می‌توانید از هر روشی که خودتان مد نظر دارید استفاده کنید. به طور مثال می‌توانید با توجه به اشکالی که رسم کرده‌اید، با ناحیه‌بندی کردن فضای تشکیل شده، مثلا توسط دو ویژگی، نوع گل داده‌ی تست را مشخص کنید.

۳.۱ اهداف مسئله و خروجی مورد انتظار

- کدهای مربوط به رسم شکل داده‌ها را در یک فایل LoadVisual.ipynb بنویسید و همراه با ۶ تصویر خروجی ذخیره کنید.
- کد مربوط به مدل‌ها را در یک main.py ذخیره کنید. هر دو کد (فایل های LoadVisual.ipynb و main.py) باید به صورت مجزا و بدون خطا قابل اجرا باشند.
- در گزارش تمرین، به طور کامل فرآیند کار، منطق طراحی مدل، نحوه‌ی محاسبه دقت مدل و کدهای استفاده شده را توضیح دهید.
- دقت مورد انتظار برای مدل‌های ذکر شده در این تمرین 100% است. دقت کمتر باعث کسر نمره خواهد شد.
- کد ها و گزارش تمرین را در یک فایل zip با نام ML_HW1Prog ذخیره کرده و در Quera آپلود کنید.

هدف از این تمرین آشنایی اولیه با یادگیری‌ماشین و زبان برنامه‌نویسی Python و محیط گسترش آن است. پس برای آموزش مدل‌ها و استفاده از آن، اجازه‌ی استفاده از توابع آماده را ندارید و تنها برای رسم شکل و خواندن فایل‌ها، مجاز به استفاده از توابع آماده می‌باشید. کد باید تمیز، همراه با متغیرهای با معنی و دارای کامنت باشد. تمامی مدل‌های گفته شده، در آینده در کلاس، تدریس خواهد شد.