بسمه تعالی

مقدمه ای بر یادگیری ماشین

نیمسال اول ۹۸-۹۹

تمرین عملی سری اول

مدرس: صابر صالح

● مهلت تحویل تمرین ها: ۱۳۹۸/۰۷/۲۸ ●

تمرین‌های برنامه‌نویسی

○ مقدمه

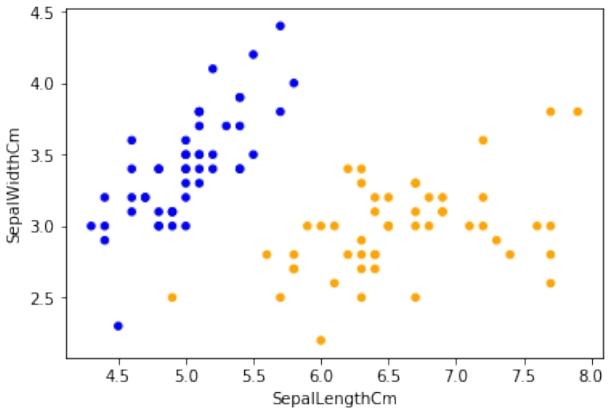
هدف از این تمرین آشنایی با زبان برنامه نویسی Python و استفاده از آن برای یادگیری ماشین است. پایتون یک زبان برنامه نویسی قوی است که به کمک کتابخانه های آن به یکی از پرطرفدارترین ابزارهای علوم داده (Data Science) تبدیل شده است. توضیحات کامل در رابطه با نصب و استفاده از پایتون را می توانید در در لینک های داده شده در نوتبوک بیابید.

۱. تمرین NumPy

این تمرین برای شروع کار با پایتون و کتابخانه ی NumPy است. در این قطعه کد، تابع main از قبل نوشته شده که تعدادی تابع را صدا می زند. هدف شما نوشتن کد این توابع می باشد. خروجی توابع نوشته شده توسط شما با خروجی مورد نظر چک شده و درستی آن به شما نشان داده خواهد شد. توضیحات لازم داخل هر تابع داده شده است.

۲. تمرین Visualization

یکی از کارهای مهم در فرآیند کار با داده، آنالیز و مشاهده گرافیکی داده موجود است. دانشمندان علوم داده از انواع نمودارها و آماره ها برای این کار استفاده می کنند. در این تمرین بخشی از دادگان Iris (گل زنبق) به شما داده می شود. این دادگان شامل اطلاعات دو نوع گل virginica و setosa می باشد. گل ها بر اساس ۴ ویژگی طول کاسبرگ، عرض کاسبرگ، طول گلبرگ و عرض گلبرگ بررسی شده اند. این ۴ ویژگی می توانند برای شناسایی و دسته بندی هر گل جدید استفاده شوند. هدف در این تمرین آن است که با نمایش گل ها بر اساس این ویژگی بیابید که چگونه ویژگی های مختلف با گونه ی گل مرتبط هستند. هر بار بر اساس ۲ ویژگی نمودار رسم کنید. به عنوان مثال تصویر حاصل از رسم گل ها بر اساس طول و عرض کاسبرگ در اینجا آمده است. شش نمودار به این صورت رسم کنید.



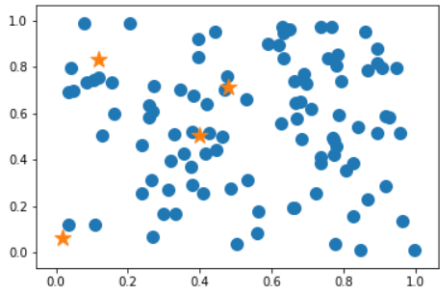
۳. پیش بینی توزیع داده

یکی از رویکردهای رایج در مسائل یادگیری ماشین تخصیص نمونه های آماری به یک توزیع با پارامترهای مشخص است. برای یک توزیع آماری با پارامترهای θ تابع درست نمایی (Likelihood function) برای x ، به صورت $p(x|\theta)$ تعریف می شود. در این تمرین فرض بر این است که چهار توزیع نرمال با میانگین و واریانس مشخص داریم. داده ها در فضای دو بعدی هستند. هدف آن است که برای تعدادی داده موجود میزان Likelihood تولید هر نقطه توسط هر توزیع را حساب کنیم. از مقایسه ی این مقادیر درمیابیم که تولید هر نقطه توسط کدام توزیع محتمل تر است. به این ترتیب هر نقطه را به منبعی اختصاص می دهیم که احتمال تولید آن بیشترین است.
راهنمایی: تابع Likelihood برای یک توزیع نرمال (گاوسی) به این صورت است:

$$P(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^n|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}(x-\mu)^T\mathbf{\Sigma}^{-1}(x-\mu)\right)$$

که μ میانگین و Σ ماتریس کواریانس (covariance matrix) می باشد. در این تمرین میانگین به صورت بردار دوبعدی μ و ماتریس کواریانس به صورت ماتریس 2×2 به نام sigma به شما داده می شود.
پس از پیش بینی، نتایج را رسم کنید. به این صورت که مشابه آن چه در بخش قبل انجام دادید هر نقطه را بر حسب مختصاتش در صفحه دو بعدی رسم کنید و با رنگ مشخص کنید که به کدام منبع نسبت داده شده است..

برای مثال به تصاویر زیر دقت کنید: در تصویر اول مجموعه داده های دسته بندی نشده را میبینیم. همچنین محل میانگین هر یک از چهار تابع توزیع احتمالی با ستاره مشخص شده اند.



حال در تصویر بعدی همان نقاط را پس از محاسبه احتمالات و پیش بینی می بینیم. رنگ ها نشانگر دسته بندی ما هستند. مشاهده می شود که داده های اطراف مرکز یک توزیع به احتمال زیاد از همان توزیع بوده اند.