مبانی یادگیری ماشین - تکلیف سری اول

مدرس: دکتر حامد ملک

پاییز ۱۴۰۲

ددلاین: ۱۲ آبان ماه - ساعت ۲۳:۵۹

سوالات تحلیلی

۱. استفاده از الگوریتم جنگل تصادفی چگونه می‌تواند مشکل overfit شدن را حل کند؟

۲. با توجه به مقادیر موجود در جدول زیر:

الف) مقدار entropy شکار شدن گاومیش را در شرط کروکودیل بودن یا نبودن شی نامعلوم بیابید.

ب) اگر بدانیم شی مورد نظر یک کروکودیل است، چه میزان اطلاعات درباره شکار شدن گاومیش پیدا می‌کنیم؟

ج) مقدار entropy کروکودیل نبودن شی را در صورت شکار نشدن گاومیش بدست آورید.

	Be eaten	Not to be eaten
Crocodile	38/100	26/100
Not a crocodile	14/100	22/100

مسائل کدی

۱. در این تمرین به دنبال پیاده‌سازی از صفر درخت تصمیم به هدف طبقه‌بندی چندکلاسه هستیم.

الف) ابتدا نوت‌بوکی که در اختیارتان قرار گرفته است را کامل کنید.

ب) این مدل را بر روی مجموعه داده داده شده آموزش دهید و نتایج مدل‌های آموزش داده شده را با ابر پارامترهای مختلف گزارش کنید. بایستی برای انتخاب هر یک از ابر پارامترهای مدل با یک نمودار و تحلیل، بهترین مقدار را انتخاب کنید.

ج) با استفاده از کتابخانه‌های موجود، مدل‌های جنگل تصادفی و تقویت گرادیان را بر روی این مجموعه داده آموزش دهید.

سپس مدل‌های خود را به ترتیب با ۲۵٪، ۵۰٪ ، ۷۵٪ و ۱۰۰٪ داده‌ها آموزش داده و نمودار Learning Curve در Scikit-learn را به صورت جداگانه برای هر یک از آن‌ها رسم کنید. در انتها تحلیل خود را از نتایج بدست آمده بنویسید.

نکات:

۱. دقت کنید که در کنار پیاده سازی موجود در نوت‌بوک، می‌توانید هر گونه پیاده‌سازی داشته باشید؛ اما در چنین حالتی، بایستی مستندسازی کامل نیز برای آن فراهم کنید. در غیر این صورت نمره‌ای به شما تعلق نخواهد گرفت. یک تغییر کوچک اما مفید می‌تواند تغییر ورودی تابع از داده (X, Y) به اندیس آن‌ها در مجموعه داده اصلی (indexes) باشد.

۲. دقت کنید که برای ارزیابی مدل خود از دسته‌های آموزش، اعتبارسنجی و آزمون استفاده کنید. در غیر این صورت نمره‌ای به قسمت تحلیلی شما تعلق نخواهد گرفت.

۲. مجموعه دادهای که در اختیار شما قرار گرفته است، شامل ویژگی‌های زیر است.

Disease: The name of the disease or medical condition

(Fever: Indicates whether the patient has a fever (Yes/No

(Cough: Indicates whether the patient has a cough (Yes/No

(Fatigue: Indicates whether the patient experiences fatigue (Yes/No

(Difficulty Breathing: Indicates whether the patient has difficulty breathing (Yes/No

Age: The age of the patient in years

(Gender: The gender of the patient (Male/Female

(Blood Pressure: The blood pressure level of the patient (Normal/High

(Cholesterol Level: The cholesterol level of the patient (Normal/High

Outcome Variable: The outcome variable indicating the result of the diagnosis or assessment for (the specific disease (Positive/Negative

در این مجموعه داده، درست یا اشتباه بودن بیماری تشخیص داده شده (Disease) بر اساس علائم پزشکی افراد (سایر ویژگی‌های موجود) در ستون Outcome Variable آمده است.

لازم به ذکر است با توجه به مقادیر برخی ویژگی‌ها که به صورت اسم مشخص شده‌اند. (مثال: Male/Female)، ابتدا پیش‌پردازش لازم را روی مجموعه داده انجام دهید.

در مرحله بعد، بر روی مجموعه داده موجود:
الف) پیاده سازی K-Nearest neighbors را از پایه (from scratch) انجام دهید.
سپس میزان دقت مدل خود را با استفاده از توابع کتابخانه Scikit-learn بر روی داده های آزمون گزارش کنید.

ب) منحنی ROC به چه معناست و چه کاربردی دارد؟
آن را برای نتایج مدل KNN رسم کرده و تحلیل خود نسبت به عملکرد آن را بنویسید.

نکات تمرین

- در صورت هرگونه تقلب نمره صفر برای شما لحاظ می‌گردد.

- استفاده از زبان غیر از پایتون مجاز نیست.
- فایل تکلیف خود را به صورت خواسته شده در سامانه کوئرا بارگذاری نمایید.

موفق باشید