دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

هوش مصنوعی
بهار ۱۴۰۰
استاد: محمدحسین رهبان
گردآورندگان: مهدی عرفانیان، صبرینه مختاری

تمرین عملی چهارم | دسته‌بندی خطی | مهلت ارسال: ۲۳ خرداد ۱۴۰۰

۱. (۱۵ نمره)

EDA

در این بخش باید کارهای زیر بر روی دیتاست انجام شود.

- نمایی کلی از دیتاست (تعداد سطرها، درصد خالی بودن سطرها، نوع ستون‌ها و ...)
- انجام عملیات پاک‌سازی
  (آ) حذف ستون‌هایی که مورد نیاز نیستند
  (ب) نرمالایز کردن ستون‌ها
  (ج) جایگزین کردن ستون‌های متنی با اعداد
  (د) انتخاب سیاست مناسب برای عملیات روی سطرهایی که مقدار خالی دارند
- توصیف دیتاست به همراه نمودار
  دقت کنید که نوع نمودارهای مورد نیاز بسته به انتخاب شماست. این بخش تا بی‌نهایت می‌تواند نمرهٔ امتیازی داشته باشد.
- افراز دیتاست به دو دستهٔ `train` و `test`
  هشتاد درصد دیتاست اصلی را به `train` اختصاص دهید.

۲. (۲۵ نمره)

Naive Bayes

در این قسمت باید از ابتدا و بدون از استفاده کتاب‌خانه‌ها الگوریتم دسته‌بند Naive Bayes پیاده‌سازی شود.
فقط استفاده از `numpy` و `pandas` و دستورات پایتون مجاز است.
پس از پیاده‌سازی مدل نحوهٔ عمل‌کرد آن را با استفاده از معیارهای معرفی‌شده در نوت‌بوک بررسی کند.
استفاده از کتاب‌خانهٔ `sklearn` برای محاسبهٔ نحوهٔ عمل‌کرد مجاز است.

۳. (۱۵ + ۵ نمره)

Logistic Regression

در این قسمت ابتدا باید تفاوت‌های Logistic Regression و Regression خطی که در کلاس تدریس شده را در پاراگرافی شرح دهید. سپس با استفاده از کتاب‌خانهٔ `sklearn` مدل را آموزش دهید و معیارهای خواسته‌شدهٔ ارزیابی را نیز گزارش کنید.
بخش امتیازی این قسمت مربوط به توضیحات دربارهٔ Logistic Regression می‌باشد.

۴. (۵ + ۱۵ نمره)

Random Forest

- **بخش اجباری:** در این قسمت باید ابتدا یک درخت تصمیم را با عمق کم به انتخاب خودتان روی دادهٔ آموزش تولید کنید. معیارهای ارزیابی را برای این درخت گزارش کنید. استفاده از کتاب‌خانه در این قسمت مجاز است ولی دقت کنید که حداکثر عمق درخت ۳ می‌تواند باشد. یعنی می‌توانید فقط روی سه ویژگی یاد بگیرید.
  ویژگی‌هایی که کتاب‌خانه برای یاد گرفتن روی داده‌ها انتخاب کرده گزارش کنید و بنویسید که با چه معیاری این ویژگی‌ها بر بقیه ترجیح داده شده‌اند.
- **بخش امتیازی:** در این بخش می‌خواهیم یک جنگل بسازیم. باید با پیدا کردن پارامترهای مناسب در کتاب‌خانه ویژگی‌های مختلف را برای یادگیری به او بدهید و سپس با رای‌گیری بین درخت‌های مختلف نتیجهٔ دسته‌بند کلی را اعلام کنید. برای این که با مفهوم Random Forest بیشتر آشنا شوید حتما لینک‌های داخل نوت‌بوک را مطالعه کنید.
  تعداد درخت‌هایی که ساخته می‌شوند و هم‌چنین عمق درخت‌ها هایپرپارامترهایی هستند که باید شما تعیین کنید. در این قسمت باید رای‌گیری را From scratch بزنید و استفاده از Random Forest کتاب‌خانه‌ها مجاز نیست.
  عمل‌کرد جنگل خود را گزارش کنید و در پاراگرافی کوتاه دلیل استفاده از مجموعه‌های ویژگی انتخاب‌شده خود را بنویسید. این بخش ۱۵ نمره امتیازی دارد.