به نام ایزد منان

تمرین سری دوم
داده‌کاوی

توضیحات:

- پاسخ به تمرین‌ها باید به صورت انفرادی صورت گیرد. حداقل برخورد با پاسخ‌های مشابه، تخصیص نمره کامل منفی به طرفین خواهد بود.
- پاسخ‌های خود را به زبان فارسی و به صورت مرتب، در قالب یک فایل فشرده (.zip) با الگوی زیر در صحفه‌ی درس بارگذاری کنید:

DM_HW[No]_[Student_number].pdf

- لطفاً نظم، ساختار و توالی سوالات را در پاسخ‌ها رعایت کنید.
- تمیزی و خوانایی گزارش تمرین از اهمیت بالایی برخوردار است.
- برای تمرین‌های عملی، علاوه بر کد گزارش کتبی نیز ارسال کنید.
- در صورتی که درمورد این تمرین سوال یا ابهامی داشتید با ایمیل dm.1401.spring@gmail.com با تدریس‌یاران درس در ارتباط باشید.
- مهلت ارسال تمرین تا ساعت 11:55 روز ۱۹ اردیبهشت است.

نیم‌سال دوم ۱۴۰۰ - ۱۴۰۱

سوالات تشریحی

سوال ۱:

همانطور که می‌دانید دسته‌بندی [1]و رگرسیون[2]، دو نوع از مسائل هستند که در یادگیری‌ماشین با آن‌ها بر می‌خوریم. با توجه به مفهومی که از آن‌ها در ذهن دارید، مثال‌هایی بیاورید و بگویید جزء کدام دسته هستند.

برای مثال اگر قیمت تعدادی خانه را داشته باشیم و بخواهیم مدلی بسیازیم که با مشاهده این داده‌ها می‌تواند قیمت خانه‌های جدید را پیش‌بینی کند، این از نوع مسائل رگرسیون است.

با تفکر روی موضوعات مختلف از هر کدام ۵ مثال کاربردی بزنید. (کاربردی بودن مثال‌ها و این که حاصل تفکر شما در دنیای اطراف باشد، نمره دارد.)

سوال ۲:

بخش اول:

مفاهیم زیر را به طور کامل تعریف کرده و فرمول آنها را بنویسید.

۱- Accuracy

۲- Recall

۳- Precision

۴- F1-Score

بخش دوم:

همانطور که متوجه شده‌اید همه معیارهای بالا برای اندازه‌گیری دقت و میزان خوب‌بودن یک مدل است. ولی چرا فقط از همان مورد اول همه‌جا استفاده نمی‌کنیم؟ آیا کاربردهای آن‌ها در جاهای مختلف متفاوت است؟

[1] Classification
[2] Regression

برای مثال در بررسی یک مدل که قرار است سالم بودن پکت‌های یک شبکه را تشخیص دهد و هدف آن در واقع شناسایی حمله‌هایی است که روی شبکه اتفاق میافتد، شاید معیار Accuracy خیلی جالب نباشد. چراکه اکثر پکت‌ها سالم بوده و تعداد پکت‌های خطرناک بسیار کمتر است. تصور کنید یک مدل که می‌تواند ۹۰ درصد پکت‌های ناسالم را تشخیص دهد در صورتی که Accuracy اش را برای داده‌های یک روز حساب کنیم به یک عدد بالا مثلا ۹۸ درصد می‌رسد و در مقابل مدلی که فقط می‌تواند ۱۰ درصد از پکت‌های ناسالم را تشخیص دهد و بقیه را به اشتباه سالم تشخیص می‌دهد باز هم دقتی با عدد بالا، مثلا ۹۷ درصد می‌رسد. این به این دلیل است که تعداد پکت‌ها در روز شاید در محدوده میلیونی باشد و تعداد پکت‌های خراب چند ده تا باشد. پس معیار دقت هیچ درک درستی به ما نمی‌دهد. در مقابل اگر از معیارهای دیگر استفاده کنیم یک درک کاملا درست و با اختلاف بالا برای بررسی صحت این دو مدل به ما می‌دهد.

با جستجو در اینترنت و خلاقیت خودتان ۳ سناریو مطرح کنید و عدم کاربرد و یا خوبی معیارهای مختلف را با هم مقایسه کنید.

سوال ۳:

با توجه به مطالب تدرس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آن‌ها لازم است.)

بیماری قلبی دارد	عروق خونی بسته	گردش خون مناسب	درد سینه
خیر	خیر	خیر	خیر
بله	بله	بله	بله
خیر	خیر	بله	بله
بله	بله	خیر	بله

سوال ۴:

با توجه به مطالب تدرس شده در کلاس، برای داده‌های زیر یک درخت تصمیم درست کنید. (ذکر تمام مراحل و توضیح آن‌ها لازم است.)

سریال کلاه‌قرمزی را دوست دارد؟	سن	آب گازدار دوست دارد؟	پاپ‌کُرن دوست دارد؟
خیر	۷	بله	بله
خیر	۱۲	خیر	بله
بله	۱۸	بله	خیر
بله	۳۵	بله	خیر
بله	۳۸	بله	بله
خیر	۵۰	خیر	بله
خیر	۸۳	خیر	خیر

سوال ۵:

بررسی کنید شاخص جینی [3]چیست و برای چه از آن استفاده می‌کنیم؟

بالا و پایین بودن عدد آن به چه معنایی است؟

همچنین بررسی کنید آیا شاخص‌های دیگری نیز وجود دارد که کارایی مشابهی با آن داشته باشند.

با اعداد دلخواه خود، چند نود تعریف کنید و این شاخص را برای آن حساب کنید.

[3] GINI index

سوال ۶:

مفهوم بیش‌برازش [4]چیست و کی اتفاق می‌افتد؟

برای رفع کردن آن چه کارهایی لازم است انجام دهیم؟ (به دلخواه خود چند مثال بزنید)

سوالات برنامه‌نویسی

برای انجام این قسمت پیشنهاد می‌شود در ابتدا زمانی را در سایت tensorflow playground بگذرانید و با تغییر پارامترها و بررسی انواع مسائل، شرایط مختلف را بررسی کنید. (برای به دست آوردن درک بهتر در بخش‌های بعدی می‌تواند به شما کمک کند.)

استفاده از jupyter notebook پیشنهاد می‌شود و می‌توانید همه موارد ذکر شده را در آن درج کنید و توضیحات لازم را بنویسید و گزارش کتبی جدا نفرستید. ولی در صورتی که کد پایتون معمولی می‌فرستید می‌توانید موارد ذکر شده را در یک فایل گزارش مربوط به کد بیاورید.

نکته دیگر اینکه شما باید به کمک کتابخانه TensorFlow این تمرین را پیاده سازی کنید و کتابخانه‌های دیگر به جز برای موارد گرفتن مجموعه‌داده مورد قبول نیست.

بخش اول:

ابتدا با دستور `from sklearn.datasets import make_circles` این کتابخانه را اضافه کنید. حال به کمک این کتاب‌خانه تعدادی دایره با مقداری نویز به دلخواه خودتان بکشید. هدف از بخش اول دسته‌بندی این دایره‌ها است. شما باید مرحله به مرحله متناسب با دستور کار جلو بروید و برای هر بخش در حد چند خط توضیح دهید دلیل نتایج به دست آمده چیست و چگونه می‌توان آن را بهینه کرد؟ (همچنین برای هر بخش اسکرین شات از نتایج و نمودار تغییرات دقت [5]و خطا [6]را در گزارش درج کنید).

[4] Overfitting
[5] Accuracy
[6] Loss

ابتدا یک شبکه عصبی بسازید و برای لایه‌های آن از تابع فعال‌ساز استفاده نکنید. آیا می‌توان داده‌ها را به صورت مناسب دسته‌بندی کرد؟

حال شبکه‌عصبی جدیدی ساخته و این‌بار برای دسته‌بندی داده‌ها از یک تابع فعال‌ساز خطی استفاده کنید. آیا دسته‌بندی به صورت صحیح انجام می‌شود؟

با توجه به این که مسئله ما از نوع دسته‌بندی است، این‌بار از یک خطای مناسب برای مسئله رگرسیون استفاده کنید. آیا می‌توانید نتیجه مناسب بگیرید؟

در این مرحله یک شبکه‌عصبی با فقط یک لایه با تعداد دلخواه نرون تعریف کنید. آیا بازهم نمی‌توانید داده‌ها را دسته‌بندی کنید ؟

برای تست بعدی مقدار نرخ‌یادگیری را به صورت دستی تنظیم کنید. با امتحان کردن مقادیر مختلف سعی کنید به بهترین مقدار آن برسید. استدلال خود را برای این انتخاب توضیح دهید و همچنین بگویید دیگر مقادیر به چه دلیل خوب عمل نمی‌کنند.

در آخر یک شبکه عصبی به انتخاب خودتان ایجاد کرده و سعی کنید بهترین نتیجه را بگیرید. دلیل عملکرد مناسب و دلیل انتخاب ابرپارامتر[7] را توضیح دهید.

بخش دوم:

در این بخش ابتدا مجموعه‌داده fashion_mnist را با دستور `from tensorflow.keras.datasets import fashion_mnist` به دست آورید. سپس با توجه به مطالبی که از بخش قبل آموختید و با ساختاری که مد نظر دارید مجموعه‌داده را دسته‌بندی کنید و برای آن یک ماتریس درهم ریختگی رسم کنید.

[7] Hyperparameter