دانشگاه تهران، دانشکده ریاضی، آمار و علوم کامپیوتر
درس یادگیری ماشین
استاد درس: دکتر باباعلی

تمرین سری هفتم

مهلت تحویل: سه شنبه ۱۶ بهمن

طراح تمرین: امیررضا جلیلی

نکات ضروری

۱. لازم است کد تمرین به همراه گزارش تحویل داده شود که شامل پاسخ سوالات نظری، توضیحات و تحلیل کد و تصاویر و نمودارهای لازم می باشد.

۲. کد تمرین باید در زبان پایتون، به صورت notebook و شامل نتایج اجرا باشد.

۳. فایل گزارش باید به فرمت pdf باشد. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، لذا لطفا در حد نیاز توضیح دهید.

۴. استفاده از کتابخانه های از پیش آماده مجاز است، مگر اینکه نقیض آن گفته شود. در گزارش کتابخانه ها ذکر شوند.

۵. نمره هر تمرین از ۱۰۰ است (مگر آنکه نمره امتیازی در تمرین مشخص شده باشد).

۶. مشاهده شباهت نامتعارف در گزارش و کد به منزله تقلب می باشد و طرفین مشمول کسر نمره خواهند شد.

۷. برای این تمرین، می‌توانید گزارش و پاسخ‌هایتان را به زبان‌های فارسی و یا انگلیسی تحویل دهید.

سناریو

کشتی‌ای بزرگ که تعداد زیادی از مسافران را حمل می‌کرده، دچار سانحه شده و بخشی از مسافران نجات یافته‌اند، اما تعداد زیادی همچنان مفقود هستند. با استفاده از داده‌های موجود، مدلی بسازید که بتواند احتمال زنده ماندن افراد را پیش‌بینی کند.پیشبینی افراد زنده برای ما از اهمیت بیشتری برخوردار میباشد زیرا این مدل می‌تواند در عملیات جستجو و نجات برای اولویت‌بندی تلاش‌ها و تمرکز بر افرادی که احتمال بیشتری برای زنده بودن دارند، موثر باشد. همچنین می‌توان از این مدل برای پیش‌بینی و مدیریت خطرات در آینده بهره برد.

مسئله ۱. Analysis Data Exploratory (EDA)

با استفاده از مجموعه داده تایتانیک[۱]، موارد زیر را انجام دهید:

(آ) ویژگی‌های داده را بررسی کرده و آماری کلی از ستون‌های موجود ارائه دهید. مشخص کنید کدام ستون‌ها دارای مقادیر گمشده هستند.

(ب) با استفاده از نمودارها، توزیع جنسیت، سن، و کلاس سفر[۲] مسافران را تحلیل کنید.

(ج) بررسی کنید چه نسبتی از مسافران بر اساس جنسیت و کلاس سفر زنده مانده‌اند.

(د) میزان وابستگی[۳] هر کدام از متغییر های مستقل را با متغییر وابسته با نمودار مناسب مشخص کنید.

۱

[۲۰ نمره]

مسئله ۲. Preprocessing Data

(آ) مقادیر گمشده در ستون‌های داده را با روش مناسب پر کنید (سه مورد از روش های رایج برای پرکردن داده های گمشده را بیان کرده و دلیل انتخاب روش خود را توضیح دهید).

(ب) ستون‌های غیرعددی (مانند Name و Sex) را به داده‌های عددی تبدیل کنید.(پنج مورد از روش‌های رایج برای کدگذاری[۴] را توضیح داده و دلیل انتخاب روش خود را براساس نوع متغیر و برداشت خود شرح دهید.

(ج) داده‌های یادگیری را با نسبت ۸۰٪ به ۲۰٪ جدا کنید، به‌طوری‌که داده‌های یادگیری و ارزیابی نسبت یکسانی از متغیر هدف را داشته باشند.

(د) آیا متغییر هدف دارای توضیع نامتوازن است؟ اگر چنین است چه راه‌حلی برای حل این مشکل پیشنهاد می‌کنید؟(۳ مورد از روش های رایج را توضیح داده و دلیل انتخاب روش خود را شرح دهید.)

[۲۰ نمره]

مسئله ۳. Implementation Model

(آ) یک شبکه عصبی ساده بدون استفاده از کتابخانه‌های آماده (مانند PyTorch ،Keras و غیره) بنویسید .

(ب) از چه تابع فعالسازی برای مدل خود استفاده می‌کنید؟ درباره توابع فعال‌سازی[۵] که در اسلاید ۱۹ آمده‌اند[۶]، تحقیق کنید. نقاط قوت، نارسایی‌ها و موارد استفاده هر یک را شرح دهید و دلیل انتخاب متد خود را بیان کنید.

(ج) با توجه به نوع تسک و سناریو چه معیارهایی[۷] برای ارزیابی مدل خود پیشنهاد میدهید؟

[۴۰ نمره]

---

[۱] https://www.kaggle.com/competitions/titanic/data
[۲] Pclass
[۳] Correlation
[۴] Encoding
[۵] Activation function
[۶] Sigmoid، tanh، ReLU، LeakyReLU، Maxout، ELU
[۷] Metrics

مسئله ۴. Tuning Hyperparameter

(آ) عملکرد شبکه عصبی خود را با تغییر تعداد نرون‌ها، تعداد لایه‌ها، و مقدار نرخ یادگیری [۸] بهبود ببخشید. نمودار عملکرد مدل را بر اساس تغییرات این پارامترها ترسیم کرده و مقادیری که مدل در آن بالاترین دقت را دارد بیان کنید.

(ب) مدل خود را با تعداد پارامتر کمتر و بیشتر از مقدار پارامتر ایده‌آل بدست آمده در بخش قبل اجرا کنید و خروجی هر کدام را از نظر *underfit* و *overfit* تحلیل کنید.

(ج) برای حالات *overfit* و *underfit*، در شبکه‌های عصبی به جز تغییر سایز مدل، چه راه‌حل‌های دیگری پیشنهاد میدهید؟ سه مورد برای هرکدام را شرح دهید.

(د) اگر پس از اعمال تکنیک‌های مناسب، دقت هر سه مدل برابر با $n\%$ باشد، کدام گزینه را ترجیح می‌دهید:
۱. مدلی با تعداد پارامترهای بالا که با اعمال تکنیک‌های رفع *overfitting*، به دقت $n\%$ رسیده است.
۲. مدلی با تعداد پارامترهای کم که با اعمال تکنیک‌های رفع *underfitting*، به دقت $n\%$ رسیده است.
۳. مدلی با اندازه متوسط که بدون نیاز به اعمال تکنیک‌های رفع *overfitting* یا *underfitting*، به دقت $n\%$ رسیده است.
دلیل انتخاب خود را توضیح داده و تحلیل کنید.

[۲۰ نمره]

[بارم کل: ۱۰۰ نمره]

با آرزوی موفقیت

[۸] Learning Rate