دانشکده مهندسی سامانه‌های هوشمند

گروه علوم داده

# یادگیری ماشین

# تمرین ششم

استاد درس

دکتر سامان هراتی‌زاده

زمان تحویل: 1403/10/25

| یادگیری ماشین – تمرین «6» | | |
|---|---|---|
| دکتر سامان هراتی‌زاده<br>دانشگاه تهران - دانشکده سامانه‌های هوشمند<br>نیم‌سال اول ۱۴۰۳-۱۴۰۴ | **دستیاران آموزشی**<br>مجتبی شاعفی<br>سجاد دشتی | ددلاین: 1403/10/25 |

مقدمه و نکات

- لطفا گزارش را در قالب مشخص شده نوشته و با فرمت pdf تحویل دهید.
- تمامی اجزا کد و نتایج می بایست توسط مصححین، عیناً تکرار پذیر باشند (حتی برای بخش بندی دادگان). می توانید از دستورات مناسب (برای pytorch از manual_seed) برای تکرار پذیری استفاده کنید.
- تحلیل نتایج حائز اهمیت است. لذا در تمامی قسمت ها نتایج به‌دست‌آمده از تمرین باید حتماً در گزارش درج و تحلیل شوند.
- در حل این تمرین، شما مجاز به استفاده از مدل‌های زبانی بزرگ (LLM) برای کمک به نوشتن کد یا حل مسائل هستید. با این حال، شما باید به‌طور کامل به تمامی کدی که تحویل می‌دهید تسلط داشته باشید و قادر به توضیح عملکرد و نیز تغییر کد باشید. استفاده از ابزارها و مدل‌های کمکی به این معنا نیست که بتوانید بدون درک کافی از کد آن را ارائه دهید؛ هدف این است که دانش و درک عمیقی از مفاهیم و راه‌حل‌ها داشته باشید.
- توضیح مختصری از نحوه عملکرد اجزای کلیدی کد ضروری است.
- برای سوالات خود می توانید از طریق ایمیل های زیر اقدام بفرمایید.

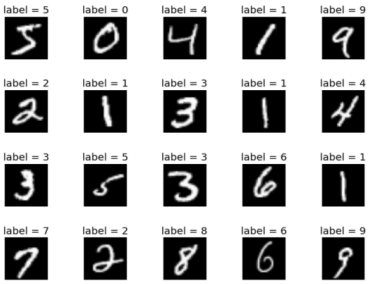
  سوال اول: shaefi@ut.ac.ir

  سوال دوم: s.dashti.k@gmail.com

| یادگیری ماشین – تمرین «6» | | |
|---|---|---|
| دکتر سامان هراتی‌زاده<br>دانشگاه تهران - دانشکده سامانه‌های هوشمند<br>نیم‌سال اول ۱۴۰۳-۱۴۰۴ | **دستیاران آموزشی**<br>مجتبی شاعفی<br>سجاد دشتی | ددلاین: 1403/10/25 |

1. هدف این تمرین طراحی یک طبقه‌بند شبکه عصبی بدون استفاده از کتابخانه آماده با استفاده از numpy برای تشخیص اعداد دستنویس 0 تا 9 می‌باشد. در لایه خروجی برای هر یک از ده عدد یک نورون در نظر گرفته می‌شود (one-hot encoding)

- **دیتاست:** مجموعه داده TinyMNIST شامل بردار ویژگی تصویر اعداد دستنویس 14*14 می‌باشد. این مجموعه شامل 5000 داده آموزش و 2500 داده آزمون می‌باشد. بردار ویژگی به طول 196 و مقادیر حقیقی صفر تا یک می‌باشد.
- **الف:** برای هر یک موارد زیر شبکه مورد نظر را آموزش داده و نمودار accuracy برای 10 ایپاک را رسم نمایید. وزن‌های اولیه را به صورت تصادفی و نرخ یادگیری را برابر 0.005 در نظر بگیرید. تابع فعالسازی لایه پنهان را relu ولایه خروجی را softmax لحاظ نمایید. تحلیل کنید از نمودار هر قسمت و مقایسه آن چه نتیجه‌ای می‌گیرید.
  - شبکه عصبی شامل یک لایه پنهان با فعالساز relu و لایه خروجی با فعالساز softmax را نظر گرفته و تعداد نورون‌های لایه پنهان را با مقادیر 16، 32، 64 تغییر دهید و نمودار خواسته شده رسم نمایید.
  - شبکه عصبی شامل دو لایه پنهان یکسان، مشابه اندازه های قسمت قبل در نظر بگیرید.
  - شبکه عصبی شامل دو لایه پنهان با ترکیب های (32،64) و (64،32)در نظر بگیرید: تنها ترتیب لایه ها عوض می‌شود.
  - شبکه عصبی با سه لایه پنهان با ترکیب های (16،16،16) و (32،32،32) و (32،16،32) و (32،64،32) و (64،64،32) در نظر بگیرید.
- **ب:** برای شبکه های با سه لایه پنهان تابع فعالساز لایه پنهان را به tanh تغییر دهید و با استفاده از نمودار خروجی شرح دهید چه تاثیر در عملکرد مدل می‌گذارند.
  - با مقایسه نتایج ساختار شبکه بهینه را گزارش دهید و تحلیل کنید با تغییر تعداد نورون ها و تابع فعالساز در لایه های پنهان در خروجی مدل چه تغییری ایجاد می شود.
  -



| یادگیری ماشین – تمرین «6» | | |
|---|---|---|
| دکتر سامان هراتی‌زاده<br>دانشگاه تهران - دانشکده سامانه‌های هوشمند<br>نیم‌سال اول ۱۴۰۳-۱۴۰۴ | **دستیاران آموزشی**<br>مجتبی شاعفی<br>سجاد دشتی | ددلاین: 1403/10/25 |

2. هدف از این سوال آشنایی با کتابخانه pytorch و به کار گیری یک مدل MLP است. در این سوال تمامی پارامترها و هایپرپارامترها به اختیار شما می‌باشد. نیاز است تمامی فرضیات انتخابی خود را ذکر نمایید. پیشنهاد می‌شود از محیط Google Colab استفاده گردد.

- **دیتاست**: از دیتاست Adult Income استفاده کنید. این دیتاست را از لینک زیر لود کنید:

https://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data

هدف آن است که بر اساس ویژگی‌های جمعیت‌شناسی و شغلی، پیش بینی کنیم که درآمد یک شخص در سال کمتر از 50 هزار دلار است یا بیشتر مساوی این مقدار است.

این دیتاست شامل ویژگی های عددی و دسته‌ای (categorical) است.

**ویژگی‌های عددی**: سن، ساعت‌های کاری در هفته، تعداد سال‌های تحصیل، سود سرمایه، ضرر سرمایه و وزن نهایی

**ویژگی‌های دسته‌ای**: نوع شغل، سطح تحصیلات، وضعیت تاهل، شغل، رابطه خانوادگی، نژاد، جنسیت و کشور محل تولد

ویژگی وزن نهایی (fnlwgt یا final weight) نشان می دهد که هر نمونه، نماینده چند فرد با ویژگی های مشابه است. می توانید در این سوال از این ویژگی صرف نظر نمایید.

- **الف**: این موارد را در گزارش خود ذکر کنید:
  - دیتاست را به سه بخش آموزش، تست و اعتبار سنجی تقسیم کنید. درصد اختصاص داده به هر یک از این موارد را ذکر نمایید.
  - روش مناسبی برای مدیریت ویژگی های categorical استفاده کرده، پیش پردازش های لازم را انجام دهید و آن ها را گزارش کنید.
  - معماری شبکه‌ای که به کار گرفته‌اید را در یک جدول ارائه دهید (تعداد لایه‌ها، تعداد نورون‌ها در هر لایه، تابع فعال ساز و ...).
  - تابع هزینه، تعداد نورون و تابع فعال ساز لایه آخر را ذکر کنید. چرا این موارد را انتخاب کردید؟
  - نرخ یادگیری و بهینه ساز را ذکر نمایید.
- **ب**: شبکه را آموزش دهید.
  - نمودار loss و accuracy را بر حسب ایپاک برای داده‌های آموزش و اعتبار سنجی رسم کنید. تحلیل خود را ارائه کنید.
  - شماره ایپاکی که بهترین عملکرد روی داده های اعتبار سنجی دارد را گزارش کنید و پارامتر های شبکه مربوط به آن ایپاک را لود کنید (بهتر است در حین آموزش آن را ذخیره کنید).
  - داده‌های تست را به شبکه دهید و ماتریس آشفتگی را برای پیش‌بینی شبکه رسم کنید و گزارش طبقه بندی را اعلام کنید.