به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

شبکه های عصبی مصنوعی و یادگیری عمیق

تمرین شماره ۲

مهر ۹۹

فهرست سوالات

سوال ۱ ـ MLP (Regression)

هدف این تمرین پیش بینی هزینه اقامت در هتل به ازای ویژگی های(ستون‌های) مختلفی می باشد که مجموعه دادگان آن در فایل Hotel.zip آپلود شده است. از داده‌های فایل H1.csv به عنوان داده های train و validation و از داده‌های فایل H2.csv به عنوان داده های test استفاده نمایید.

الف) در این قسمت می بایست از تمامی ویژگی ها (numerical , categorical) به عنوان feature استفاده نمایید و شبکه عصبی چند لایه ای را طراحی کنید که هزینه اقامت در هتل را به ازای تمامی این feature ها پیش بینی نماید. نسبت داده های train به validation می بایست ۸۰ به ۲۰ باشد. نمودار loss را به ازای ۳۰، ۵۰ و ۱۰۰ ایپاک رسم نمایید.

نکته: می بایست در ابتدا ویژگی های numerical و categorical را از هم تشخیص داده و برای هر کدام نرمال سازی های لازم را انجام دهید تا بتوانید از این داده ها در شبکه بهره ببرید.

نکته: می‌توانید از کتابخانه pandas به منظور نرمال سازی ویژگی های categorical بهره ببرید.

نکته: می بایست ویژگی Average Daily Rates(ADR) را پیش بینی نمایید(هزینه اقامت در هتل) و به غیر از آن ۳۴ ویژگی دیگر وجود دارد که بر اساس آن ها ADR پیش بینی می شود.

نکته: اطلاعات دقیق dataset را می توانید از این لینک https://www.sciencedirect.com/science/article/pii/S2352340918315191 بررسی نمایید. ولی حتما از dataset ایی که در اختیار شما قرار گرفته است استفاده نمایید بخشی از این Dataset تغییر یافته است.

ب) ارزیابی را بر حسب معیار های MSE و MAE انجام داده و نمودار این معیار ها را به ازای تعداد ۳۰ و ۵۰ و ۱۰۰ ایپاک رسم نمایید. (۶ نمودار)

ج) به ازای داده های test مقدار هزینه اقامت در هتل(ADR) را پیش بینی نمایید و در یک فایل CSV جدید به نام Reults.csv مقدار واقعی داده ADR و مقدار تخمین زده شده ADR و اختلاف بین مقدار واقعی و مقدار تخمین زده شده را به دست آورید.(۳ ستون)

د) در این سوال اطلاعات feature های مختلفی (مانند IsCanceled، Meal و...) از یک اقامت مشتری هتل داریم. جمع آوری این اطلاعات هزینه بر است و هدف این است که تنها داده های مهم هر اقامت را نگهداری کنیم که تا حد خوبی با استفاده از آن ها بتوان به پیش بینی های دقیق تر هزینه اقامت واقعی بدون افزایش پیچیدگی شبکه دست یافت. چطور می توان کشف کرد که کدام feature ها (ستون) در داده های ما مهم تر است؟ دقت نمائید برای این سوال باید یک پاسخ کلی ارائه نمائید و راه حل ارائه شده نباید بر اساس پیش فرض های ذهنی از داده های آن مسئله باشد.

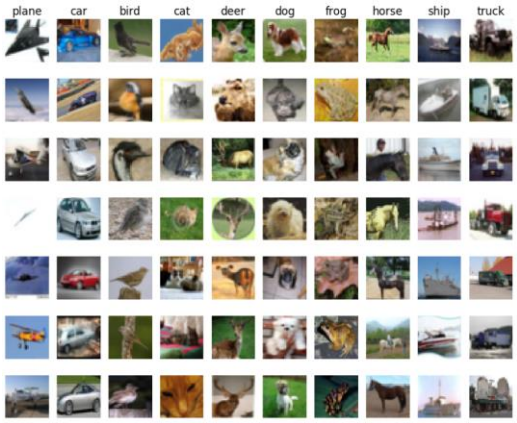
ه) راه حل پیشنهادی خود در قسمت (د) را پیاده سازی نمائید و با استفاده از آن اهمیت featureهای مهم را بدست آورده و آن ها را بیان نمایید. بار دیگر به ازای ۳۰ ایپاک شبکه ی خود را پیاده سازی کرده و مانند قسمت (ج) نتایج را در فایل SelectedResult.csv ذخیره نمایید.

و) (سوال امتیازی) لطفا فایل پیوست شده تمرین را مطالعه نمائید تا با بحث شاخص Smoothness آشنایی پیدا کنید ، سپس با استفاده از روش Forward selection مجموعه ورودی هایی که Smoothness Index بیشتری دارند را انتخاب کنید.

راهنما : در روش Forward selection از بین تمام feature ها ابتدا هر feature را بررسی می کنیم که کدام یک بالاترین Smoothness Index را به ما می دهد و آن را کنار می گذاریم ، در ادامه n-1 تا feature باقی مانده را بررسی می کنیم و feature ای که در کنار feature قبلی ، بالاترین Smoothness Index را به ما می دهد را پیدا می کنیم ، در مرحله سوم بین n-2 تا feature باقی مانده این جست و جو را انجام می دهیم ... . این کار را تا زمانی ادامه می دهیم که Smoothness Index جمعی به بالاترین مقدار برسد و دیگر اضافه کردن یکی feature باعث بهتر شدن آن به طرز قابل توجهی نشود.

سوال ۲ ـ MLP (Classification)

هدف در این تمرین ایجاد یک طبقه بند برای طبقه بندی مجموعه داده CIFAR-10 با استفاده از شبکه های MLP[1] است . این مجموعه داده (نمونه ای از آن را تصویر زیر مشاهده می نمائید) شامل ۶۰ هزار تصویر رنگی است که در ۱۰ کلاس دسته بندی شده است .



تصویر (۲) – نمونه مجموعه داده CIFAR-10

به طور معمول از ۵۰ هزار تصویر آن به عنوان مجموعه داده آموزشی استفاده می کنند و ۱۰ هزار تصویر را به عنوان مجموعه داده تست استفاده می کنند. شما این مجموعه داده را دانلود و ۱۰ تصویر ابتدای آن را همراه با نام آن شی نمایش داده و سپس داده ها را به سه بخش آموزش[2]،تست[3] و ارزیابی[4] تقسیم کنید.

از طریق این لینک می توانید این مجموعه داده را دانلود نمائید.

با استفاده از کتابخانه Keras نیز می توانید مجموعه داده را دانلود نمائید.

```
from keras.datasets import cifar10
(x_train, y_train), (x_test, y_test) = cifar10.load_data()
```

در ادامه پیش پردازش های لازم را انجام دهید تا داده ها برای آموزش شبکه عصبی آماده بشوند.

حال با توجه به آموخته های کلاس درس ، یک شبکه MLP طراحی نمائید. هدف این سوال پیاده سازی شبکه عصبی ، بررسی تاثیر تغییرات هایپرپارامترها و حل چالش مجموعه داده نامتوازن است.

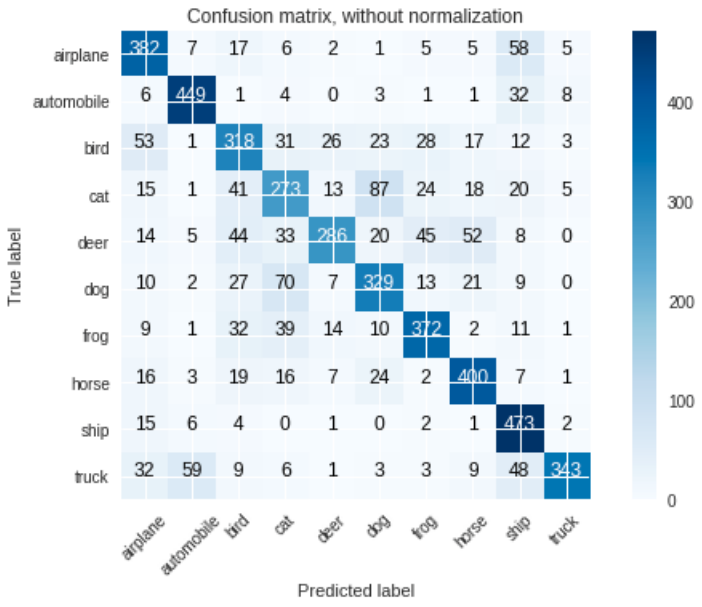
[1] Multilayer perceptron
[2] Train
[3] Test
[4] Validation

فرضیات مسئله :

- تعداد لایه های مخفی را برابر ۲ در نظر بگیرید.
- از روش Stochastic mini batch based استفاده نمائید.
- هایپرپارامترها مانند تابع خطا ، نرخ یادگیری و ...را نیز خودتان انتخاب نمائید برای این کار می توانید از Grid Search یا آزمون و خطا استفاده نمائید. پارامترهای انتخاب شده را در گزارش بیاورید.

موارد زیر را در گزارش برای قسمت های ب و ج بیاورید :

- در دو نمودار جداگانه تغییرات دقت [1] و خطای [2] مدل در هر دور [3] را برای داده ی ارزیابی و آموزش حالت های خواسته شده نشان دهید.
- همچنین خطا، دقت و ماتریس آشفتگی [4] را برای داده ی تست محاسبه کنید.



تصویر (۳)- نمونه ای از ماتریس آشفتگی

[1] Accuracy
[2] Loss
[3] Epoch
[4] Confusion Matrix

الف) لازم است برای حل این مسئله از روش Stochastic mini batch based استفاده شود ، از سه دسته با اندازه های ۳۲ ، ۶۴ و ۲۵۶ استفاده نمائید و تاثیر تفاوت اندازه دسته ها را در دقت و زمان آموزش شبکه بررسی نمائید.

نکته : دقت نمائید در این آزمایش ها بقیه هایپرپارامترها ثابت هستند.

ب) توابع فعالساز هر لایه را تغییر دهید و تاثیر توابع فعالساز را در دقت آموزش شبکه بررسی نمائید، مجموعا ۳ مرتبه توابع فعالساز را در لایه های ماقبل آخر تغییر دهید و نتایج آن را در گزارش بیاورید. مزایا و معایب این توابع فعالساز را نسبت به دیگری بررسی نمائید.

نکته : از توابع فعالساز ReLU ، TanH و Sigmoid استفاده کنید.

نکته : دقت نمائید در این آزمایش ها بقیه هایپرپارامترها ثابت هستند و از بهترین مدل قسمت (الف) استفاده نمائید.

ج) تابع خطا شبکه را تغییر دهید و تاثیر تابع خطاهای متفاوت را در دقت آموزش شبکه بررسی نمائید، مجموعا ۲ مرتبه تابع خطا را تغییر دهید و نتایج آن را در گزارش بیاورید. دلیل این تفاوت را از منظر ریاضی بررسی نمائید.

نکته : تابع خطا را از دو خانواده متفاوت انتخاب کنید ، به طور مثال تابع خطای Cross Entropy را با تابع خطای MSE می توانید مقایسه کنید.

نکته : دقت نمائید در این آزمایش ها بقیه هایپرپارامترها ثابت هستند و از بهترین مدل قسمت (ب) استفاده نمائید.

د) بهینه ساز شبکه را تغییر دهید و تاثیر بهینه سازهای متفاوت را در دقت آموزش شبکه بررسی نمائید، مجموعا ۲ مرتبه بهینه ساز شبکه را تغییر دهید و نتایج آن را در گزارش بیاورید.

نکته : از بهینه سازهای SGD و Adam استفاده کنید.

نکته : دقت نمائید در این آزمایش ها بقیه هایپرپارامترها ثابت هستند و از بهترین مدل قسمت (ج) استفاده نمائید.

ه) با توجه به ارزیابی های انجام شده ، انتخاب کدام پارامترها بهترین نتیجه را می دهد؟ با توجه به نتیجه بدست آمده از این شبکه، جدول (۱) را پر نمائید.

برای این مدل علاوه بر خطا و دقت ، معیارهای ارزیابی دیگر شامل Precision ، Recall و F-Score را نیز گزارش نمائید.

و) اگر در یک مجموعه داده ، تعداد داده های دسته های ما با هم برابر نباشند چه مشکلی هنگام آموزش شبکه رخ می دهد؟ یک راه حل برای حل این مشکل ارائه دهید.

ز) از دسته های Airplane و Bird هر کدام نصف تصاویر داده آموزشی را انتخاب کنید و از بقیه دسته ها تمامی تصاویر داده اموزشی را انتخاب کنید و با استفاده از راه حل گفته شده در قسمت (و) مشکل نابرابری دسته ها را حل کنید و شبکه را آموزش دهید .

نکته : در این سوال حق ندارید از حالت های اتوماتیک کتابخانه ها برای حل این مسئله استفاده کنید.

سوال ۳ ـ Dimensionality Reduction

الف) در مجموعه داده CIFAR-10 با کمک یک Autoencoder ابتدا فضای ویژگی را کاهش داده و سپس با استفاده از بهترین مدل شبکه عصبی بدست آمده در سوال ۳ ، به آموزش یک شبکه MLP بپردازید.

ب) در ابتدا بیان کنید روش تحلیل مولفه ی اصلی (PCA) چگونه ابعاد یک داده را کاهش می دهد و سپس در مجموعه داده CIFAR-10 با کمک یک PCA ابتدا فضای ویژگی را کاهش داده و سپس با استفاده از بهترین مدل شبکه عصبی بدست آمده در سوال ۳ ، به آموزش یک شبکه MLP بپردازید.

نکته : در این سوال حق استفاده از کتابخانه های یادگیری ماشین را برای قسمت PCA ندارید و PCA را باید خودتان پیاده سازی نمائید.

ج) در نهایت با توجه به نتایج بدست آمده از داده های تست قسمت های الف و ب جدول زیر را پر کنید و مقایسه کنید این دو روش کاهش بعد در مقایسه با هم چطوری کار می کنند.

جدول (۱) – مقایسه دقت شبکه های مختلف

مورد	دقت داده های تست	خطا داده های تست
بهترین شبکه بدست آمده در سوال ۳		
Autoencoder		
PCA		

دقت نمائید مجموعا ۲ آزمایش در این سوال لازم است.

نکته : برای اینکه آزمایش هایتان قابل اتکا باشد ، لازم است ساختار و پارامترهای شبکه عصبی مورد استفاده در هر ۲ قسمت سوال مطابق با بهترین شبکه بدست آمده در سوال ۳ (قسمت ه) باشد. همچنین ابعاد را در PCA و Autoencoder به یک اندازه کاهش دهید.

نکات:

- مهلت تحویل این تمرین تا ۱۴ آبان است.
- گزارش را در قالب تهیه شده که روی صفحه درس در Elearn بارگذاری شده، بنویسید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است. لطفاً تمامی نکات و فرض‌هایی که برای پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید را در گزارش ذکر کنید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست. اما باید نتایج بدست آمده را گزارش و تحلیل کنید.
- دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند. بنابراین هرگونه نتیجه و یا تحلیلی که در شرح سوال از شما خواسته شده است را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- در صورت مشاهدهٔ تقلب امتیاز تمامی افراد شرکت‌کننده در آن، ۱۰۰- لحاظ می‌شود.
- برای انجام تمرین ها و مینی پروژه ها، تنها زبان برنامه نویسی مجاز Python است. در این تمرین به جز قسمت های گفته شده در بقیه قسمت ها می توانید از کتابخانه های یادگیری عمیق استفاده نمائید.
- استفاده از کدهای آماده برای تمرین‌ها به‌هیچ‌وجه مجاز نیست. اما برای مینی‌پروژه‌ها فقط برای قسمت‌هایی از کد و به عنوان راهنمایی برای پیاده‌سازی، می‌توانید از کدهای آماده استفاده کنید.
- نحوهٔ محاسبه تاخیر به این شکل است: مهلت ارسال بدون جریمه تا تاریخ اعلام شده و پس از آن به مدت هفت روز تا ۲۱ آبان بارگذاری ممکن است و در نهایت، پس از بازه تاخیر نمره تکلیف صفر خواهد شد.
- لطفاً گزارش، فایل کدها و سایر ضمایم مورد نیاز را با فرمت زیر در سامانه مدیریت دروس بارگذاری نمایید.

HW2 _[Lastname]_[StudentNumber].zip

- در صورت وجود هرگونه ابهام یا مشکل می‌توانید از طریق رایانامه‌های زیر با دستیاران آموزشی مربوطه خانم رومینا اوجی (سوال ۱) و آقای علی کریمی (سوال ۲ و ۳) در تماس باشید:

Alikarimi120@gmail.com

Romina.oji @ut.ac.ir