به نام خدا

دانشکده‌ی مهندسی برق و کامپیوتر

تمرین سری چهارم یادگیری ماشین

دانشگاه تهران

پردیس دانشکده های فنی

سلام بر دانشجویان عزیز، چند نکته مهم:

۱. حجم گزارش به هیچ عنوان معیار نمره دهی نیست، در حد نیاز توضیح دهید.

۲. نکته‌ی مهم در گزارش نویسی روشن بودن پاسخ‌ها می‌باشد، اگر فرضی برای حل سوال استفاده می‌کنید حتما آن را ذکر کنید، اگر جواب نهایی عددی است به صورت واضح آن را بیان کنید.

۳. کدهای ارسال شده **بدون گزارش** فاقد نمره می‌باشند.

۴. برای سوالات شبیه سازی، فقط از دیتاست داده شده استفاده کنید.

۵. فایل نهایی خود را در یک فایل زیپ شامل، pdf گزارش و فایل کدها آپلود کنید. نام فایل زیپ ارسالی الگوی ML_HW4_StudentNumber داشته باشد.

۶. از بین سوالات **شبیه سازی** حتما به هر دو مورد پاسخ داده شود.

۷. نمره تمرین ۱۰۰ نمره می‌باشد و حداکثر تا نمره ۱۱۰ (**۱۰ نمره امتیازی**) می توانید کسب کنید.

۸. هرگونه شباهت در گزارش و کد مربوط به شبیه سازی، به منزله تقلب می‌باشد و کل تمرین برای طرفین **صفر** خواهد شد.

۹. در صورت داشتن سوال، از طریق ایمیل mohammadjavadranjbark@gmail.com سوال خود را مطرح کنید.

سوال ۱: (15 نمره)

جدوال زیر برای دو عملگر and و Xor را در نظر بگیرید را در نظر بگیرید. هدف این مسئله طراحی یک شبکه عصبی و توضیح مراحلی است که برای آموزش آن نیاز است تا پیش‌بینی کند که با توجه به ورودی خروجی کلاس ۱ یا صفر خواهد بود. (توجه داشته باشید که نیازی به اجرای این فرآیند روی یک ماشین ندارید، فقط باید به کلمات توصیف کنید که چگونه این فرآیند انجام می‌شود).

A	B	A **AND** B
0	0	0
0	1	0
1	0	0
1	1	1

A	B	A **XOR** B
0	0	0
0	1	1
1	0	1
1	1	0

به سوالات زیر برای هر دو عملگر پاسخ دهید و پاسخ‌های خود را توضیح دهید(دقت کنید معماری شبکه برای این دو عملگر متفاوت است.):

معماری شبکه عصبی:

- چند لایه استفاده می‌کنید؟ چرا؟
- چند گره در هر لایه استفاده می‌کنید؟ چرا؟
- شکل شبکه عصبی‌تان رسم کنید.

آموزش شبکه عصبی:

- الگوریتم به عقب انتشار خطا به دنبال چه چیزی می‌گردد؟
- چگونه وزن‌های اتصالات شبکه‌تان را مقداردهی اولیه می‌کنید؟
- مراحل انجام شده توسط الگوریتم به عقب انتشار خطا در یک دوره (epoch) را توضیح دهید.

سوال ۲: (15 نمره)

همانطور که در درس یاد گرفتید اگر ما داده‌ی گسسته‌ی X را داشته باشیم، entropy آن را می‌توانیم از فرمول زیر بدست آوریم:

$$H(X) = -\sum_{i} P(X = i) * \log P(X = i)$$

فرض کنید متغییر دیگر Y که دارای توزیع $P(Y = j)$ و توزیع توام $P(X = i, Y = j)$ داریم که mutual information این دو متغییر از رابطه‌ی زیر بدست می‌آید:

$$I(X;Y) = \sum_{ij} P(X = i, Y = j) * \log \frac{P(X = i, Y = j)}{P(X = i) * P\,Y = j)} = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

حال به دو تعریف زیر دقت کنید:

یک کد nary-K برای X، یک نگاشت از X به یک رشته (رمز شده) به نام $C(X)$ است که هر کاراکتر می‌تواند K مقدار داشته باشد.
یک کد پیشوندی (prefix)، کدی است که هیچ کلمه‌ای در این کد کد پیشوندی کلمه‌ی دیگری نیست.

یک مثال شایع از یک کد پیشوندی، نمایش دودویی اعداد است. در کد دودویی، هیچ نمایش دودویی یک عدد، پیشوندی دیگر نمی‌باشد. به عنوان مثال:
- نمایش دودویی عدد ۲ برابر "۱۰" است.
- نمایش دودویی عدد ۳ برابر "۱۱" است.
- نمایش دودویی عدد ۴ برابر "۱۰۰" است.
- نمایش دودویی عدد ۵ برابر "۱۰۱" است.

در این مثال، می‌توانید ببینید که هیچ نمایش دودویی یک عدد، پیشوندی دیگر نمی‌باشد که با تعریف یک کد پیشوندی مطابقت دارد.

میانگین طول کد به صورت $L(C) = \sum_i P(X = i) * l_i$ تعریف می‌شود که در آن l_i طول $C(i)$ است. می‌توان ثابت کرد که $H(X)$ حداقل میانگین طول کد مورد نیاز برای رمزگذاری X است، و این حداقل فقط و فقط زمانی بدست می‌آید که $P(X = i) = k^{-l_i}$

حال با توجه به موارد ذکر شده به سوالات زیر پاسخ دهید:

الف) تعداد ۹ توپ فلزی وجود دارد. یکی از آن‌ها سنگین‌تر از دیگران است. لطفاً یک استراتژی با استفاده از ترازو برای پیدا کردن توپ سنگین‌تر با کمترین تعداد تست ممکن طراحی کنید. سپس نشان دهید که از نظر اطلاعاتی بهینه است.

راهنمایی: ارتباطی بین طول میانگین کد با تعداد مورد انتظار آزمایش‌ها را برقرار کنید(به موارد گفته شده در بالا دقت کنید).

ب) اگر الگوریتم ID3 را بر روی مسئله ۹ توپ فلزی اجرا کنیم، آیا این الگوریتم درخت تصمیم بهینه را ایجاد خواهد کرد؟ نکته: نتیجه مقایسه هر دو مجموعه توپ را به عنوان یک ویژگی در نظر بگیرید و اطلاعات به دست آمده از مقایسه آن‌ها را مقایسه کنید.

سوال ۳: (15 نمره)

الف) درخت تصمیمی که از آموزش با ۴ داده‌ی زیر به وجود می‌آید را بدست آورید و رسم کنید(فرض کنید ویژگی‌ها همین‌هایی هستند که در جدول وجود دارند.):

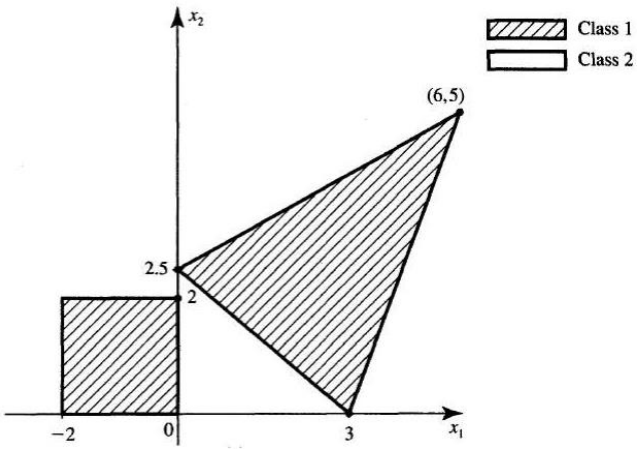
Movie	Genre	Age rating	Language	Source	Film Location	Studio	Enjoyed
1	Drama	PG-13	French	Book Adaptation	Canda	Warner Bros	Yes
2	Drama	PG-13	English	Book Adaptation	Canda	Warner Bros	Yes
3	Horror	G	English	Book Adaptation	Canda	A24	No
4	Drama	PG-13	English	Book Adaptation	USA	A24	Yes

ب) مثال آموزشی زیر را به مجموعه‌ی بالا اضافه کنید و دوباره درخت تصمیم را بدست آورید.

5	Drama	PG-13	French	Original Screenplay	Canda	Warner Bros	No

سوال ۴: (15 نمره)

یک شبکه عصبی چند لایه‌ای طراحی کنید که طبقه‌بندی شکل (۱) را انجام دهد:
توجه: از کمترین تعداد لایه‌ها و نورون‌ها استفاده کنید.



سوال ۵: (شبیه سازی، 30 نمره)

دیتاست MNIST یکی از دیتاست‌های معروف در زمینه تشخیص اعداد دست‌نویس است. این دیتاست شامل تصاویری از اعداد دست‌نویس (از ۰ تا ۹) در ابعاد ۲۸*۲۸ پیکسل می‌شود. به طور کلی، MNIST شامل مشخصات زیر است:

1. تعداد داده‌ها: MNIST شامل ۶۰,۰۰۰ تصویر برای آموزش و ۱۰,۰۰۰ تصویر برای تست می‌شود.
2. اندازه تصاویر: تمام تصاویر از نوع سیاه و سفید (grayscale) هستند و ابعاد ۲۸*۲۸ پیکسل دارند.
3. برچسب‌ها: هر تصویر متناظر با یک عدد از ۰ تا ۹ است و دارای یک برچسب است.
4. مقدار داده آموزش و تست: داده‌های آموزش برای آموزش مدل‌ها استفاده می‌شوند و داده‌های تست برای ارزیابی عملکرد مدل‌ها در مرحله آخر استفاده می‌شوند.

5. مقادیر پیش‌پردازش: معمولاً تصاویر به مقادیر پیکسل‌ها تقسیم ۲۵۵ می‌شوند تا به مقادیر بین ۰ و ۱ تبدیل شوند. این مقادیر معمولاً به عنوان ورودی‌های شبکه عصبی استفاده می‌شوند.

MNIST به عنوان یک بنچمارک معروف در زمینه یادگیری ماشین و بخصوص در برنامه‌های تشخیص تصویر و شبکه‌های عصبی کوچک و متوسط استفاده می‌شود. بسیاری از افراد از این دیتاست به عنوان نقطه شروع خود در زمینه یادگیری عمیق و شبکه‌های عصبی استفاده می‌کنند.

بخش ۱:

در ابتدا، دادگان MNIST را با استفاده از یک شبکه عصبی MLP (Multilayer Perceptron) برای مسئله طبقه بندی تصاویر اعداد دستنویس را آموزش دهید. ۲۰ درصد از دادگان را به عنوان دادگان ولیدیشن کنید و بقیه‌ی دادگان برای آموزش استفاده شوند.

- پارامترهای شبکه و تعداد لایه‌های مختلف را شرح دهید.
- Learning curve (نمودار خطا و دقت شبکه در طول آموزش) بر روی داده‌ی آموزش و ولیدیشن رسم کنید.
- دقت طبقه‌بند روی دادگان تست را نشان دهید و همینطور ماتریس درهمریختگی آن را رسم کنید.
- به نظر شما بهترین زمان برای متوقف کردن آموزش شبکه کجاست؟ چرا؟
- بخش ۲(امتیازی):

حال مدلی با استفاده از یک شبکه Convolutional neural network آموزش دهید:

- پارامترهای شبکه و تعداد لایه‌های مختلف را شرح دهید.

- Learning curve (نمودار خطا و دقت شبکه در طول آموزش) بر روی داده‌ی آموزش و ولیدیشن رسم کنید.
- دقت طبقه‌بند روی دادگان تست را نشان دهید و همینطور ماتریس درهمریختگی آن را رسم کنید.

بخش ۳(امتیازی):

از مدل MobileNet برای Transfer Learning استفاده کنید و شبکه‌ای برای دیتاست MNIST آموزش دهید.

- Learning curve (نمودار خطا و دقت شبکه در طول آموزش) بر روی داده‌ی آموزش و ولیدیشن رسم کنید.
- دقت طبقه‌بند روی دادگان تست را نشان دهید و همینطور ماتریس درهمریختگی آن را رسم کنید.

برای ساخت مدل‌ها از هرکدام از کتابخانه‌های Pytorch یا Tensorflow می‌توانید استفاده کنید.

سوال ۶: (شبیه سازی، 20 نمره)

برای این سوال از کتابخانه‌هایی مانند Sklearn استفاده نکنید و باید کد را از پایه بزنید.

مجموعه داده "Adult" از مشهورترین و پراستفاده‌ترین مجموعه‌های داده در زمینه یادگیری ماشین و انجام وظایف طبقه‌بندی (Classification)و تحلیل داده‌های اجتماعی است. این مجموعه داده UCI Machine Learning Repository (مرکز اطلاعاتی دانشگاه کالیفرنیا، آیرواین) برای اهداف تحقیقاتی و آموزشی به عنوان یک مجموعه داده آموزشی برای مسائل طبقه‌بندی انسان‌هایی که در آمریکا اقامت دارند و آیا درآمد سالیانه آنها کمتر یا بیشتر از ۵۰ هزار دلار است، تشکیل شده است.

مسئله اصلی در این مجموعه داده، پیش‌بینی این است که آیا یک فرد درآمد سالیانه‌اش بیشتر یا کمتر از ۵۰ هزار دلار دارد. در واقع، این یک مسئله طبقه‌بندی دودویی است که دو کلاس اصلی دارد ">50K" : (بیشتر از ۵۰ هزار دلار درآمد سالیانه) و"<=50K" (کمتر یا مساوی ۵۰ هزار دلار درآمد سالیانه).

ویژگی‌ها: این مجموعه داده شامل چندین ویژگی (ویژگی‌های پیوسته و categorical) مانند سن، جنسیت، میزان تحصیلات، وضعیت تاهل، شغل، و ... می‌شود. این مجموعه داده دارای ۱۴ ویژگی است که برای ساده‌سازی تعدادی از این ویژگی‌ها را حذف کرده‌ایم و فقط ۶ ویژگی categorical مانده است. همچنین داده‌هایی با مقادیر missing نیز از این مجموعه داده حذف شده است. ویژگی‌های مجموعه داده‌ی ما به صورت زیر است:

workclass (8 values)

education (16 values)

marital-status (7 values)

occupation (14 values)

relationship (6 values)

race (5 values)

sex (2 values)

native-country (41 values)

حال دو مجموعه داده‌ی test و train که هرکدام شامل ۱۰۰۰۰ داده می‌شوند داریم. اولین ستون برچسب هر داده را مشخص می‌کند و ۸ ستون دیگر ویژگی‌هاش مشخص شده‌ی بالا هستند.

با استفاده از الگوریتم ID3 روی مجموعه دادگان train یک درخت تصمیم بسازید، همانطور که اشاره شد در هر مرحله باید از Information gain برای انتخاب بهترین ویژگی استفاده کنید. دقت کنید در صورتی که پس از استفاده از تمام ویژگی‌ها کلاس مشخص نشد باید از رای گیری بین داده‌ها برای برچسب زنی برگ‌ها استفاده کنید.

سپس میزان دقت و صحت مدل برای دادگان test و train اعلام کنید.