

مقدمه ای بر یادگیری ماشین ۲۵۷۳۷

دانشگاه صنعتی شریف

گروه ۲

دانشکده مهندسی برق

مدرس: سید جمال الدین گلستانی

نیمسال پاییز ۹۹-۰۰

تکلیف شماره ۶

موعد تحویل: جمعه ۱۷ بهمن ۹۹

توضیحات کلی

- در صورتی که برای عضو شدن در سایت‌های درس بر روی piazza.com و quera.ir یا برای آپلود کردن تکالیف خود دچار مشکل شدید، با آدرس ایمیل amirahosseinalmeli@gmail.com تماس بگیرید.
- هر دو بخش کامپیوتری و تئوری هر تکلیف را بر روی سایت آپلود نمایید. تحویل به صورت کاغذی لازم نیست.
- در مورد هر تکلیف، تمام فایل‌های مربوط به سوالات کامپیوتری را در یک فایل به نام HWCin.zip و تمام فایل‌های مربوط به سوالات تئوری را در فایلی به نام HWTin.zip قرار دهید که i شماره تکلیف و N شماره دانشجویی شماست.
- به دلیل قابلیت‌های سایت piazza.com، از این سایت برای مدیریت سوال‌های مطرح شده استفاده می‌گردد. سوالات خود را تنها از طریق این سایت بفرستید و از سایت quera.ir صرفاً برای آپلود تکالیف خود استفاده کنید. در صورت ایمیل کردن تکالیف به دستیاران آموزشی، نمره‌ای به آن تعلق نمی‌گیرد.

سوالات عملی

سوال C4:

در این سوال یادگیری توسط شبکه عصبی و بر روی مجموعه‌ای از داده مرتبط با وضعیت حرکتی یک فرد انسان صورت می‌گیرد و هدف از آن طبقه‌بندی وضعیت یا نوع حرکت فرد به یکی از ۱۲ حالت ممکن می‌باشد. این ۱۲ حالت، که توسط مقادیر ۱ تا ۱۲ برای برچسب Y مشخص می‌گردند شامل سه وضعیت حرکتی (راه رفتن، بالارفتن از پله و پایین رفتن از پله)، سه وضعیت غیر حرکتی (ایستادن، نشستن و دراز کشیدن) و شش وضعیت گذار بین دوتا از وضعیت‌های غیر حرکتی (مانند گذار از وضعیت دراز کشیدن به وضعیت ایستادن) می‌باشد. برای هر مثال (هر فرد)، داده X که با استفاده از آن عمل طبقه‌بندی صورت می‌گیرد، شامل ۵۶۱ مولفه (یا ۵۶۱ ویژگی) می‌باشد. این ویژگی‌ها از روی اطلاعات خام جمع‌آوری شده توسط یک تلفن همراه متصل به کمر فرد بدست می‌آید. اطلاعات خام جمع‌آوری شده توسط تلفن مربوط به سرعت زاویه‌ای و شتاب در سه بعد است که با فرکانس ۵۰ بار در ثانیه و طی ۲.۵ ثانیه نمونه برداری می‌شود. این اطلاعات پس از یک مرحله پردازش اولیه تبدیل به ۵۶۱ ویژگی مورد اشاره می‌گردد. در این مساله، نحوه جزییات مرحله پردازش اولیه مورد توجه ما نیست.

فایل داده موجود شامل ۵۶۱ ویژگی مورد اشاره برای ۷۷۶۷ مثال (یک سطر برای هر مثال) می‌باشد. ابتدا ده درصد از مثال‌ها را به صورت تصادفی (با `np.shuffle` جایگشت دهید و ۱۰ درصد ابتدایی را انتخاب کنید) برای تست جدا کنید و طی روند آموزش تنها از نود درصد باقیمانده داده استفاده کنید.

کتابخانه‌های آماده می‌تواند یک شبکه عصبی `fully connected` با تعداد T لایه و تعداد n گره در هر لایه را پیاده سازی کرده و عمل یادگیری را با استفاده از یک الگوریتم مشابه روش SGD طی i گام (iteration) انجام دهد که مقادیر T, n, i قابل تنظیم هستند. تابع فعال سازی (activation) به صورت default به فرم ReLU می‌باشد، هر چند انتخاب‌های دیگری نیز وجود دارد. برای پارامترهای دیگر مانند طول گام نیز انتخاب‌هایی به صورت default صورت گرفته است. انتخاب‌های default را تغییر ندهید.

الف - عمق و تعداد گره در هر لایه را برابر $T = 8$ و $n = 8$ قرار دهید. آنگاه الگوریتم یادگیری را پنج بار، هر بار با تعداد گام $i = 10, 100, 200, 300, 400$ اجرا نمایید. پس از هر بار اجرا، خطای حاصله برای داده‌های آموزشی، همچنین خطای حاصله برای داده‌های تست را به دست آورید. این دو نوع خطا را به ترتیب L_T, L_S می‌نامیم. L_T, L_S را بر حسب i ترسیم نمایید. هر یک از دو خطا در چه مقدار i مینیمم می‌گردد. برداشت خود را از منحنی تغییرات L_T, L_S و مقایسه آنها را بیان کنید.

ب- تعداد گام‌های الگوریتم را برابر $i = 100$ و تعداد لایه‌ها را برابر $n = 8$ قرار دهید و این بار الگوریتم را هفت بار به ازای $n = 1, 2, 4, 8, 16, 32$ اجرا نمایید و بررسی‌های خواسته شده در بند الف را برای این حالت تکرار کنید.

ج- تعداد گام‌های الگوریتم را برابر $i = 100$ و تعداد گره در هر لایه را برابر $n = 8$ قرار دهید. الگوریتم را شش بار به ازای $T = 1, 2, 4, 8, 16$ اجرا نمایید و بررسی‌های خواسته شده در بند الف را برای این حالت تکرار کنید.

سوال C5:

در این سوال به بررسی و پیاده‌سازی multiclass classification با استفاده از روش‌های مختلف یادگیری و مقایسه این روش‌ها پرداخته می‌شود.

برای این سوال توصیه می‌شود از کتابخانه‌های Keras و scikit-learn زبان پایتون استفاده کنید اما استفاده از کتابخانه‌های متناظر pytorch و tensorflow بلامانع است. برای آشنایی و یادگیری این کتابخانه‌ها منابع زیر توصیه می‌شود:

- pytorch:

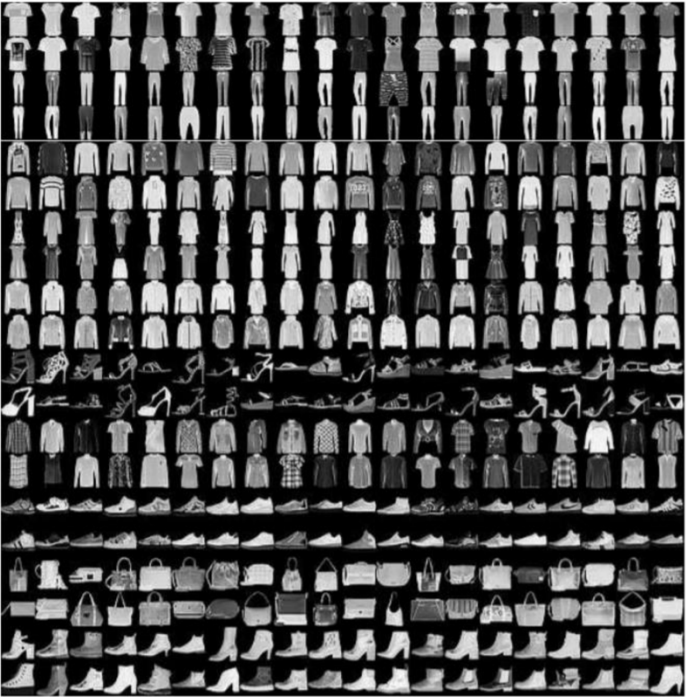
http://cs231n.stanford.edu/2019/notebooks/pytorch_tutorial.ipynb

- tensorflow:

http://cs231n.stanford.edu/2019/notebooks/CS231N_TensorFlow_Tutorial.ipynb

- keras: <https://keras.io/api/>

در این سوال از بخشی از دیتاست معروف fashion-mnist استفاده می‌کنیم که هدف آن تشخیص نوع لباس بر اساس تصویر آن است.

Label	Description	Examples
0	T-Shirt/Top	
1	Trouser	
2	Pullover	
3	Dress	
4	Coat	
5	Sandals	
6	Shirt	
7	Sneaker	
8	Bag	
9	Ankle boots	

داده‌ها در غالب یک فایل CSV با ۷۸۵ ستون و ۱۰۰۰۰ ردیف در اختیار شما قرار گرفته است. هر ردیف مربوط به یک عکس می‌باشد که ۷۸۴ ستون اول آن اعداد پیکسل‌های یک عکس ۲۸*۲۸ و ستون آخر class (نوع عکس) را مشخص می‌کند پس شما باید از ۷۸۴ ستون اول به عنوان ورودی های طبقه‌بندی مختلف استفاده کرده تا ستون آخر را به عنوان خروجی پیش‌بینی کنید .

در یک مساله طبقه‌بندی چندتایی یا multiclass classification دقت کار و انواع خطاهایی که صورت گرفته با یک ماتریس به نام confusion matrix بیان می‌شود. درایه سطر i و ستون j این ماتریس، تعداد نمونه‌هایی را نشان می‌دهد که طبقه (یعنی برچسب واقعی) آنها i بوده و الگوریتم طبقه‌بندی برچسب j را برای آنها پیش‌بینی کرده است. به این ترتیب درایه‌های روی قطر این ماتریس تعداد نمونه‌هایی را نشان می‌دهد که درست طبقه‌بندی شده‌اند و دقت طبقه‌بندی برابر است با نسبت جمع درایه‌های روی قطر این ماتریس به جمع کل درایه‌های ماتریس .

در این سوال طبقه‌بندی را با هر یک از پنج روش زیر انجام می‌دهید و بعد از اجرای هر روش ماتریس confusion و دقت طبقه‌بندی را برای آن روش به‌دست آورید .

به موارد زیر دقت بفرمایید:

- در ابتدا نیمی از داده‌ها را به طور تصادفی مانند سوال قبل برای validation جدا کنید.
- در هر روش دو دسته پارامتر یا گزینه مطرح هستند. گزینه‌های معین شده (که در توضیح روش در زیر مشخص شده‌اند) و گزینه‌های قابل انتخاب. گزینه‌های قابل انتخاب را باید خود شما به گونه‌ای با سعی و خطا تعیین کنید که به دقیق‌ترین طبقه بندی بیانجامد.
- گزارشی شامل دقت هر یک روش ها confusion matrix و پارامتر(گزینه) های مورد استفاده در هر روش و مقایسه روش‌های مختلف را به همراه کد بارگذاری کنید .

روش اول: SVM (این روش را SVM با کرنل خطی نیز می‌نامند زیرا مثل این است که از نگاشت $\varphi(x) = x$ استفاده شده است.)

گزینه‌های معین شده: نوع کرنل linear

گزینه‌های قابل انتخاب: ندارد.

روش دوم: SVM با کرنل گوسی

گزینه‌های معین شده: نوع کرنل *Gaussian* یا *rbf*

گزینه‌های قابل انتخاب: پارامتر کرنل گوسی (γ)

روش سوم: K-nearest neighbor

گزینه‌های معین شده: استفاده از فاصله‌ی اقلیدسی

گزینه‌های قابل انتخاب: K

روش چهارم: درخت تصمیم‌گیری

در این روش از پارامترهای پیشفرض توابع آماده استفاده کنید و نیازی به سعی و خطا نیست .

روش پنجم: شبکه عصبی

گزینه معین‌شده: یک شبکه تمام متصل با عمق $T = 3$ (یعنی با دو لایه مخفی) تعداد نورون‌های هر لایه مخفی برابر ۱۰۰ و لایه خروجی با ده نورون از نوع *softmax*. لایه *softmax* به هر یک از برجسب‌ها یک احتمال نسبت می‌دهد و سپس بزرگترین احتمال را به عنوان برجسب پیشنهادی انتخاب می‌کند. برای بهینه سازی از الگوریتم SGD با تابع هزینه *cross entropy* استفاده کنید. برای سایر پارامترها از مقادیر پیشفرض استفاده کنید .

گزینه قابل انتخاب: نوع تابع فعالساز لایه‌های میانی

*نمودار تابع هزینه برحسب زمان یادگیری را در گزارش خود رسم کنید.

سوال C6:

الف - برنامه‌ای بنویسید که برای n بردار m مولفه‌ای، الگوریتم K-means را اجرا کند و n برچسب بین 0 تا $k - 1$ که خوشه‌بندی حاصل را مشخص می‌کنند، برگرداند. ورودی برنامه (تابع) یک ماتریس $n \times m$ و عدد k است.

- مراکز دسته‌ها را در ابتدا به صورت تصادفی از میان داده‌ها انتخاب می‌شوند.

- به عنوان تابع فاصله از فاصله‌ی اقلیدسی استفاده کنید.

ب - داده‌های iris چهار ویژگی از سه نوع گل را در اختیار ما قرار می‌دهد که در فایل iris.csv در اختیار شما قرار گرفته است .

- به وسیله‌ی تابع خودتان، داده‌های iris flower را با استفاده از ۴ ویژگی موجود، خوشه بندی کنید.

- داده‌های حاصل را در فضای دو بعدی به ازای هر دو ویژگی (جمعا ۶ نمودار) رسم کنید.

- رنگ هر نقطه باید متناسب با خوشه‌ی متناظر باشد.

ج - طبق شکل‌های به دست آمده؛ به نظر شما، آیا می‌توان یکی از این ۴ ویژگی را حذف کرد بدون آن که دقت خوشه‌بندی تغییر زیادی داشته باشد؟ دلایل خود را شرح دهید. (راهنمایی: می‌توانید با استفاده از ۳ ویژگی خوشه‌بندی کنید و نتایج را مقایسه کنید).