دانشگاه صنعتی شریف	**مقدمه ای بر یادگیری ماشین ۲۵۷۳۷**
دانشکده مهندسی برق	**گروه ۱**
نیمسال پاییز ۱۴۰۱-۱۴۰۰	**مدرس: سید جمال الدین گلستانی**

تکلیف کامپیوتری شماره 3

موعد تحویل: جمعه 10 دی 1400

توضیحات کلی

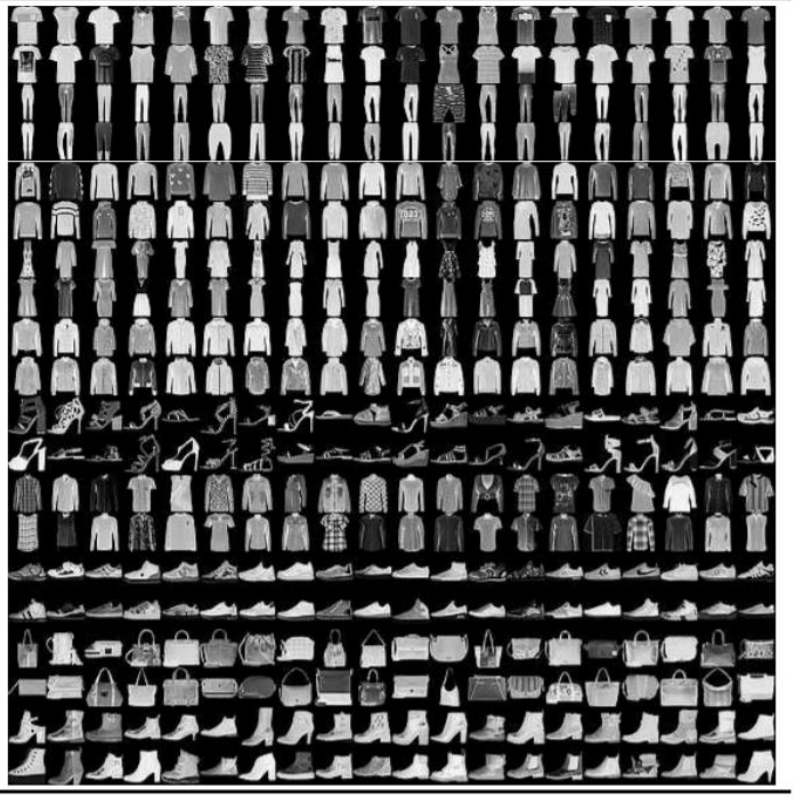
- تمام فایل‌های مربوط به سوالات کامپیوتری را در یک فایل به نام CHW3N.zip قرار دهید که N شماره دانشجویی شماست.
- سوالات خود را در مورد این تکلیف با دستیار آموزشی آقای آرمین عزیزی در آدرس ایمیل armin.az77@yahoo.com مطرح کنید.

در پایان فایل های نوت بوک به فرمت ipynb. را که هم شامل کدها و نتایج و هم شامل گزارش هست بفرستید. سعی کنید تمام چیزهایی که خواسته شده را داخل نوت بوک ها بنویسید اما اگر راحت تر بودید که بعضی سوالات تشریحی را به دلیل نیاز به فرمول نویسی یا موارد دیگر در Word یا ... بنویسید، می توانید این کار را انجام دهید اما در همان فایل نوت بوک بگویید که در کجا پاسخ این قسمت داده شده است.

سوال C5:

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

برای این سوال توصیه می‌شود یا از کتابخانه‌های Keras وscikit-learn زبان پایتون و یا از کتابخانه pytorch استفاده کنید. 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 785 ستون و 10000 ردیف در اختیار شما قرار گرفته است. هر ردیف مربوط به یک عکس میباشد که 784 ستون اول آن اعداد پیکسل‌های یک عکس 28*28 و ستون آخر class (نوع عکس) را مشخص می‌کند. پس شما باید از 784 ستون اول به عنوان ورودی های طبقه‌بندهای مختلف استفاده کرده تا ستون آخر را به عنوان خروجی پیشبینی کنید.

در یک مسأله 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 طبقه بندی برچسب j را برای آن‌ها پیش‌بینی کرده است. به این ترتیب درایه‌های روی قطر این ماتریس تعداد نمونه‌هایی را نشان می‌دهد که درست طبقه بندی شده‌اند و دقت طبقه‌بندی (A) 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 نمایید:

- در ابتدا به صورت تصادفی داده ها را به سه مجموعه آموزشی S (50 درصد داده ها)، اعتبار سنجی یا *Validation* (35 درصد) و تست T (15 درصد) تفکیک کنید. در هر روش به کار رفته در این مساله، خطای بدست آمده بر اساس این سه مجموعه را به ترتیب با L_S ، L_V و L_T نشان میدهیم.
- کد شما باید به گونه ای باشد که پس از اجرا تمام مراحل انجام شود و نتایج حاصل نمایش داده شود.
- نام و پسوند فایل دیتا را تغییر ندهید، زیرا کد شما با فایلی با نام مشابه و دیتایی که در اختیار شما قرار نگرفته است چک خواهد شد.
- 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.
- در هر روش از داده آموزشی S برای یادگیری و از داده اعتبارسنجی V برای محاسبه ماتریس Confusion و تعیین دقت روش (A) استفاده کنید و بر اساس آن روش با بهترین دقت را تعیین نمایید.
- در خاتمه برای روش برگزیده خود، با استفاده از داده تست، ماتریس Confusion و دقت A را محاسبه نمایید.
- گزارشی شامل الف – پارامتر (گزینه به کار رفته) در هر روش، ب – ماتریس Confusion و دقت هر روش بر اساس مجموعه دیتای V و ج – ماتریس Confusion و دقت روش منتخب خود بر اساس مجموعه دیتای T *تهیه و به همراه کد بارگزاری نمایید.*

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

گزینه معین شده: نوع کرنل Gaussian یا rbf

گزینه قابل انتخاب: پارامتر کرنل گوسی (σ)

روش سوم: K-nearest-neighbor

گزینه معین شده: استفاده از فاصله اقلیدسی

گزینه قابل انتخاب: K

روش چهارم: درخت تصمیم گیری

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 100 و لایه خروجی با ده نورون از نوع softmax. لایه softmax به هر یک از برچسب‌ها یک احتمال نسبت میدهد و سپس بزرگترین احتمال را به عنوان برچسب 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 فعال‌سازی لایه‌های میانی

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