

مقدمه ای بر یادگیری ماشین ۲۵۷۳۷

دانشگاه صنعتی شریف

گروه ۱

دانشکده مهندسی برق

مدرس: سید جمال الدین گلستانی

نیمسال بهار ۹۸-۹۷

### تکلیف شماره ۸

موعد تحویل: دو شنبه ۹۸/۴/۱۰

### 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 attarisadegh@yahoo.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.
- برای سوال‌های کامپیوتری از زبان برنامه نویسی پایتون یا متلب استفاده کنید.

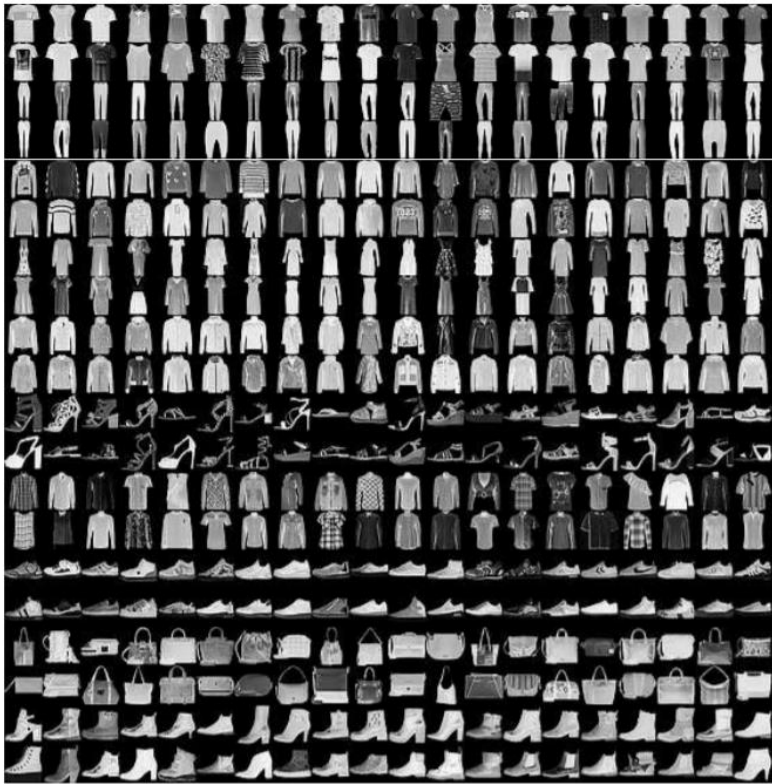
### سوالات عملی

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 توابع متناظر<sup>۱</sup> و toolbox های یادگیری عمیق متلب نیز بلامانع است.

<sup>۱</sup> توابعی مثل fitcecoc و fitctree و fitknn

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 طبقه‌بندی‌های 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 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 به دست آورید.

به موارد زیر دقت نمایید:

- در ابتدا نیمی از داده‌ها را بصورت تصادفی برای validation جدا کنید.
- 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.
- 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.
- کد مربوط به خواندن داده‌ها از دیتاست و نمایش چند عکس به عنوان نمونه در فایل‌های Question5.m و Question5.py قرار داده شده است. به دلخواه خود یکی از فایلها را تغییر داده و ارسال 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 ( $\sigma$ )

**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 معین شده: 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 فعال سازی لایه‌های میانی

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

## سوال C6

الف) تابعی بنویسید که برای  $n$  بردار  $m$  مؤلفه‌ای، الگوریتم K-Means را اجرا کند و  $n$  برچسب بین  $0$  تا  $k-1$ ، که خوشه بندی حاصل را مشخص می‌کنند، برگرداند. ورودی تابع یک ماتریس  $n \times m$  و عدد  $k$  است.

- مراکز دسته ها در ابتدا به صورت تصادفی انتخاب می‌شوند.

- به عنوان تابع فاصله از فاصله ی 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 وسیله ی تابع خودتان، داده‌های iris flower را با استفاده از هر ۴ ویژگی موجود، خوشه بندی کنید.

- داده‌های حاصل را در فضای دو بعدی رسم کنید. (۶ نمودار)

- رنگ هر نقطه باید متناسب با خوشه ی 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 خوشه بندی تغییر زیادی داشته باشد؟ دلایل خود را شرح دهید. (راهنمایی: می‌توانید با استفاده از ۳ ویژگی خوشه‌بندی کنید و نتایج را مقایسه کنید.)

د) کد خود را به گونه ای تغییر دهید که مراکز دسته ها در هر مرحله را نگهداری کند. کد را چندین بار بر اساس دو ویژگی petal width - petal length اجرا کنید و به ازای چند نقطه‌ی اولیه مسیر تغییر مراکز دسته‌ها را مشاهده کنید.