Data mining
Homework #3

Dr Amirmazlaghani
Amirkabir University of Technology
(Tehran polytechnic)

توجه: پیش از شروع تمرین لطفا موارد زیر را با دقت مطالعه نمایید.

لطفا تمام فایل‌های تمرین را (از جمله فایل pdf گزارش و فایل‌های کد) در یک فایل zip/rar ذخیره کرده و نام آن را به HW3_XXXXXXXX.zip تغییر دهید. سپس آن را در مودل بارگذاری کنید.

سوال‌ها به دو بخش نظری و برنامه‌نویسی تقسیم شده‌اند. سوال‌های نظری را می‌توانید به جای فایل word در برگه کاغذ انجام داده و تصویر آن را در فایل word قرار دهید. دقت کنید که خوانایی تمرین شرط لازم برای دریافت نمره آن است. توجه کنید که فایل گزارش را حتما به صورت pdf شده تحویل دهید.

تمرین‌های برنامه‌نویسی را می‌توانید با یکی از زبان‌های Matlab یا Python انجام دهید.

برای هر سوال باید کدی جدا نوشته شود. برای مثال کدهای سوال ۳ بخش (a) را در فایل p3a.py ذخیره کنید.

مهلت تحویل ۲ سوال اول این تمرین تا ۱۱ دی است. موعد تحویل سوالات سوم و چهارم اعلام خواهد شد.

سوال اول - بخش اول

دیتاست زیر را در نظر بگیرید :

TID	ITEMS
t001	A, B, D, G
t002	B, D, E
t003	A, B, C, E, F
t004	B, D, E, G
t005	A, B, C, E, F
t006	B, E, G
t007	A, C, D, E
t008	B, E
t009	A, B, E, F
t010	A, C, D, E

الگوریتم a-priori را به کمک متلب یا پایتون، پیاده‌سازی کنید. تمام frequent itemsetها و همچنین تمام قوانین انجمنی آن‌ها را تولید کنید و سپس به سوالات زیر پاسخ دهید. (مقدار support threshold را ۴ درنظر بگیرید)

(a) What are the frequent 1-itemsets?

(b) What are the frequent 2-itemsets?

(c) What are the frequent 2-itemsets with support greater or equal to 7?

(d) What are the association rules generated by the A-priori algorithm with a confidence of 1?

(e) What is the confidence of the association rule {B}⇒{E} generated by the A-priori algorithm?

سوال اول - بخش دوم

دیتاست زیر را در نظر بگیرید :

TID	ITEMS
t001	A, C, D
t002	B, C, E
t003	A, B, C, E
t004	B, E

الگوریتم قبل را روی این دیتاست پیاده کرده و به سوالات زیر پاسخ دهید. (مقدار support threshold را ۲ درنظر بگیرید)

(a) What are the frequent itemsets?

(b) How many association rules can you generate from the frequent itemsets with confidence bigger than 0.65?

(c) What are the association rules that you can generate from the frequent itemsets with confidence bigger than 0.8?

(d) What is the confidence of the association rule {E}⇒{C}?

(e) What is the support value of the association rule {B}⇒{C}?

سوال دوم - بخش اول

در این تمرین قرار است یک مجموعه داده دو دسته ای را به روش SVM دسته‌بندی نمایید. در صورتی که در Python برنامه می نویسید می‌توانید از کتابخانه sklean.svm استفاده کنید. اگر در MATLAB کدنویسی می‌کنید، libsvm انتخاب مناسبی است.

معمولا در الگوریتم‌های دسته‌بندی که دارای پارامترهایی هستند که در فرایند آموزش تنظیم نمی‌شوند، علاوه بر مجموعه آموزشی (Train) و آزمایشی (Test)، مجموعه مجزای سومی از داده ها به نام مجموعه اعتبارسنجی (Validation) نیز تشکیل داده می شود. چنین پارامترهایی به نام فراپارامتر (Hyperparameter) شناخته می‌شوند و برای تنظیم مقدار بهینه آن ها، هر بار قبل از آموزش classifier بر روی مجموعه آموزشی، مقداری به آن‌ها تخصیص داده می‌شود و پس از انجام آموزش، دقت classifier بر روی داده اعتبارسنجی محاسبه می‌شود. پس از آموزش متوالی classifierc با مقادیر مختلف فراپارامترها، مقداری که منجر به بهترین نتیجه بر روی داده اعتبارسنجی شده است به عنوان مقدار بهینه فراپارامتر درنظر گرفته می شود.

برای این آزمایش، مجموعه داده svmdata.csv را بارگذاری کرده و آن را به سه دسته آموزشی، اعتبارسنجی و آزمایشی تقسیم کنید. سعی کنید نسبت دسته ها به گونه ای باشد که در هر دسته مقدار داده کافی وجود داشته باشد.

(a) مجموعه آموزشی را در صفحه دوبعدی نمایش داده و هر دسته را با رنگ جداگانه‌ای نشان دهید.

(b) این مجموعه داده جداپذیر خطی نیست؛ برای جداسازی دو دسته می توان از Soft-margin SVM استفاده کرد. با استفاده از مقادیر مختلف پارامتر تنظیمی (C) SVM را آموزش دهید. پس از هر بار آموزش خطای داده اعتبارسنجی را اندازه بگیرید. میزان خطای اعتبارسنجی را بر اساس مقدار پارامتر C در یک نمودار نمایش دهید.

(c) داده های آزمایشی را مانند قسمت اول نمایش داده و خط جداکننده دو دسته را برای بهترین classifier که در مرحله قبل پیدا کردید نمایش دهید. خطای داده آزمایشی برای این classifier را گزارش کنید.

سوال دوم - بخش دوم

SVM classifier تنها می‌تواند داده را به صورت خطی جدا کند. برای جبران این کاستی، می‌توان بر روی داده تبدیلی را اعمال کرد تا در فضای جدید نمونه‌های داده به صورت خطی جداپذیر باشند. برای ادامه مجموعه داده svmdata2.csv را بارگذاری کرده و آن را به دو دسته آموزشی و آزمایشی تقسیم کنید.

(a) **مجموعه آموزشی داده‌های جدید را در صفحه دوبعدی نمایش داده و هر دسته را با رنگ جداگانه‌ای نشان دهید. می‌توان تبدیل های متنوعی را پیدا کرد که پس از اعمال، داده در مختصات جدید به صورت خطی جداپذیر باشد. رابطه یک تبدیل با این ویژگی را بنویسید. تابع هسته معادل ($K(u,v)$) را به دست آورید.**

(b) **مجموعه آزمایشی را در مختصات جدید نمایش دهید.**

(c) **داده ها در مختصات اولیه را با یک SVM (با پارامتر دلخواه) دسته‌بندی کرده و دقت دسته‌بندی را بر روی مجموعه آزمایشی گزارش کنید.**

(d) **داده های تبدیل یافته را با یک SVM دسته بندی کرده و دقت دسته بندی را بر روی مجموعه آزمایشی گزارش کنید.**

سوال سوم - بخش اول

هدف از این تمرین پیاده‌سازی شبکه‌ی عصبی سه‌لایه است.

صورت مساله بدین ترتیب است که در یک کارخانه ی تولید نوشیدنی، سه نوع نوشیدنی مختلف تولید می‌شود. به دلایلی تعدادی ازین نوشیدنی ها پس از تولید، بدون برچسب مانده اند و نمی دانیم متعلق به کدام دسته‌اند.

میخواهیم شبکه عصبی‌ای را به عنوان classifier برای این مساله طراحی کنیم که با در اختیار داشتن ۱۳ خصوصیت(فیچر) بتواند مشخص کند که محصول از کدام نوع است. به این منظور دیتاست Drinks شامل ۱۷۸ عدد از محصولات کارخانه در اختیار شما قرار گرفته است. سه ستون آخر نشان دهنده ی نوع محصول هستند و ستون‌های دیگر، فیچرها را نشان می‌دهند. در این بخش سوال مجاز به استفاده از کتابخانه و توابع آماده نیستید.

مراحل پیاده سازی:

معماری کلی شبکه عصبی مورد نیاز به این ترتیب است:

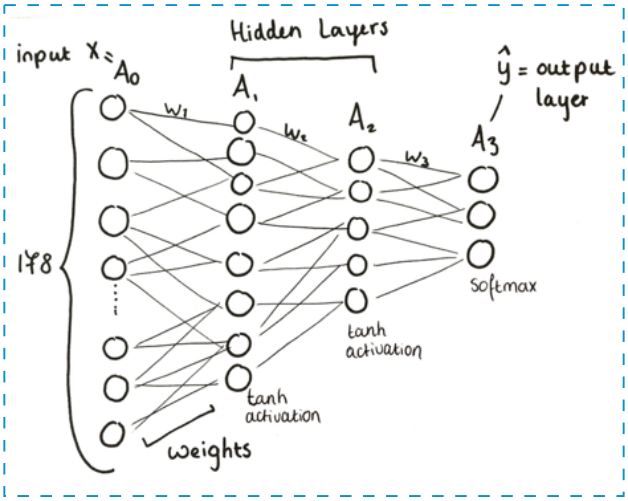
- **لایه ی ورودی شامل ۱۳ نورون(X)**
- **لایه ی اول شامل ۸ نورون(A1)**
- **لایه ی دوم شامل ۵ نورون(A2)**
- **لایه سوم شامل ۳ نورون(A3)**

لایه اول و دوم، لایه های پنهان شبکه ی عصبی هستند و از tanh به عنوان تابع فعالسازی(activation function) برای آن ها استفاده میشود. مقدار خروجی این تابعد عددی بین ۱ و -۱ خواهد بود.

برای لایه ی خروجی از تابع softmax استفاده میشود. خروجی آن بین ۰ و ۱ است.

برای اطلاعات بیشتر در خصوص این توابع می‌توانید به اینجا و اینجا مراجعه کنید.

دقت کنید که لازم است پیش از شروع پیاده سازی، وزن‌ها و پارامترهای دیگر را در صورت لزوم، به طور تصادفی مقدار دهی اولیه نمایید.



(a) پیاده سازی Forward propagation:

به منظور اینکار، در هر مرحله، توابع z1, z2, z3 را (که ورودی لایه‌های داخلی هستند) به کمک رابطه‌ی خطی بین w (وزنها) و مقادیر بایاس (b) بسازید. سپس به کمک آن، مقادیر A1, A2, A3 را محاسبه نمایید.

در هر مرحله، پارامترهای شبکه عصبی و مقادیر به دست آمده را ذخیره کنید.

(b) پیاده سازی backward propagation:

در این مرحله باید به کمک gradient descent میزان خطا را کاهش داده و وزنها را به روز رسانی کنید.

برای این کار، لازم است در هر مرحله، dz،dw،db را محاسبه کنید (توضیح: به عنوان مثال، نماد dt به معنای مشتق loss function نسبت به t است).

تعداد نمونه‌ها را m در نظر بگیرید و فرمول مشتقات را به دست آورید.

(c) فاز آموزش

در این مرحله با آموزش دادن شبکه‌ی عصبی، باید مقادیر بهینه‌ی وزن‌ها و بایاس‌ها را پیدا کنید. از accuracy به عنوان معیار تصمیم‌گیری استفاده کنید. برای آموزش دادن شبکه، لازم است یک نرخ آموزش برای آن تعریف کنید. تعریف این نرخ به صورت زیر است:

$$w := w - \eta \times \frac{dL(w)}{dw}$$

یادگیری شبکه‌های عصبی در چند مرحله صورت می‌گیرد. به هر یک از این مراحل epoch گفته می‌شود. لازم است تابعی برای آموزش شبکه بنویسید که در هر مرحله، back propagation و forward propagation را انجام داده، کارایی شبکه را با استفاده از accuracy بررسی کند، پارامترها را به روز رسانی کند و در نهایت شبکه عصبی را به حالت بهینه آموزش دهد. برای این سوال، learning rate را برابر با ۰.۷ قرار دهید.

سوال سوم - بخش دوم

سوال قبل را با یکی از کتاب‌خانه‌ها یا توابع آماده‌ی موجود پیاده‌سازی کنید.

سوال چهارم

توجه: در این تمرین لزومی به استفاده از فریم‌وورک تنسورفلو نیست. اگرچه ترجیح بر آن است که با این فریم وورک آشنایی پیدا کنید. بنابراین پیاده سازی با استفاده از فریم وورک تنسورفلو نمره اضافی دارد.

صورت مساله پیاده سازی یک الگوریتم K-Means جهت خوشه بندی تصاویر مجموعه داده MNIST یا دست نوشته است. جهت دسترسی به این مجموعه داده می توانید مستقیما از کتابخانه Keras استفاده کرده و مجموعه داده را بارگذاری کنید. برای این کار می توانید از این آدرس کمک بگیرید.

جهت پیاده سازی الگوریتم و آشنایی با فریم‌وورک تنسورفلو می توانید از سندهای گوگل به عنوان مرجع استفاده کنید که از این آدرس قابل دسترسی است.

جهت یادگیری دقیقتر می توانید پیاده‌سازی گوگل را مرجع قرار داده و از آن کمک بگیرید. توجه کنید که نمره بیشتر به افرادی تعلق خواهد گرفت که پیاده‌سازی را به همراه توضیحات دقیق مربوطه ارائه دهند.