دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ریاضی و علوم کامپیوتر

تمرین سوم درس یادگیری ماشین

عنوان تمرین : Classification بخش دوم

استاد درس :

دکتر اکبری

تاریخ انتشار : 5 آذر

تاریخ برگزاری کلاس رفع اشکال : 11 آذر

مهلت تحویل : 17 آذر

تاریخ برگزاری کارگاه : 25 آذر

بخش اول : سؤالات تشریحی

۱) دسته‌بندی متن، کاربردی است که در آن برای یک متن ارائه شده دسته‌ای خاص از بین دسته‌های از پیش تعیین شده مشخص می‌شود. مدل naïve bayes ابزاری مناسب برای این مسئله است. متغیری که در این مسئله باید احتمالش مشخص شود دسته مربوط به متن است. متغیرهای مسئله که به‌واسطه آنها کلاس را مشخص می‌کنید حضور یا عدم حضور هر واژه در متن مربوطه است. فرض می‌شود که حضور یا عدم حضور هر کلمه در متن مستقل از دیگری است و فراوانی حضور هر کلمه نیز بستگی به دسته متن دارد.

الف) به طور دقیق تشریح کنید که این مدل برای این مسئله چگونه ساخته می‌شود. فرض کنید مجموعه‌ای از متن‌های مختلف داریم که دسته هر متن نیز مشخص شده است. (مجموعه آموزشی)
ب) به طور دقیق تشریح کنید که برای یک متن جدید چگونه می‌توان دسته آن را تشخیص داد.
ج) آیا فرض استقلال شرطی برای این مسئله فرض منطقی است؟ (توضیح دهید)

۲) با فرض معتبر بودن هسته‌های k_1 و k_2 اعتبار هسته‌های زیر را بررسی کنید.

1. $k_3(x_1, x_2) = k_1(x_1, x_2) + k_2(x_1, x_2)$
2. $k_4(x_1, x_2) = k_1(x_1, x_2) * k_2(x_1, x_2)$
3. $k_5(x_1, x_2) = e^{k_1(x_1, x_2)}$
4. $k_6(x_1, x_2) = (1 - x_1^T x_2)^{-1}$

۳) جدول زیر را به عنوان مجموعه داده در نظر بگیرید. در این مجموعه داده ستون های اول تا سوم را به عنوان ویژگی و ستون آخر را به عنوان Target در نظر بگیرید. به کمک مدل Naïve Bayes محاسبه کنید که هریک از نمونه های زیر به کدام برچسب تعلق دارد.

1. (Family Structure = Single Parent , AgeGroup = Middle-aged , Income Status = High)
2. (Family Structure = Childless, AgeGroup = Old , Income Status = Low)
3. (Family Structure = Extended , AgeGroup = Young , Income Status = Medium)
4. (Family Structure = Nuclear , AgeGroup = Young , Income Status = High)
5. (Family Structure = Nuclear , AgeGroup = Young , Income Status = Low)

Type of family structure	Age group	Income status	Will they buy a car?
Nuclear	Young	Low	Yes
Extended	Old	Low	No
Childless	Middle-aged	Low	No
Childless	Young	Medium	Yes
Single Parent	Middle-aged	Medium	Yes
Childless	Young	Low	No
Nuclear	Old	High	Yes
Nuclear	Middle-aged	Medium	Yes
Extended	Middle-aged	High	Yes
Single Parent	Old	Low	No

➢ سؤال‌های ۴ و ۵ امتیازی هستند.

۴) فرض کنید $X_1, X_2, X_3, \dots, X_N$ متغیر های تصادفی مستقل باشند و فرض کنید هر X_i در بازه $[m_i, M_i]$ قرار دارد. در آن صورت نامساوی زیر که به نامساوی هوفدینگ معروف است درست است.

$$P\left\{\sum_{i=1}^{N}(X_i - E(x_i) \geq t)\right\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^{N}(M_i - m_i)^2}\right)$$

الف) عبارت بالا را اگر $t \to \infty$ برود بررسی کنید و بگویید نشان دهنده چه چیزی است.

ب) به کمک نامساوی بالا مسئله زیر را ثابت کنید.

فرض کنید برای حل یک مسئله تصمیم‌گیری یک الگوریتم داریم (مثلاً آیا عدد p یک عدد اول است یا خیر). فرض کنید الگوریتم بصورت رندوم تصمیم میگیرد و جوابی برمیگرداند که به احتمال $\frac{1}{2} + \delta$ درست است که $\delta \geq 0$. در واقع مقدار کمی از حدس رندوم بهتر کار میکند. برای بهبود کارایی این الگوریتم را N بار اجرا میکنیم و در نهایت تصمیمی که بیشتر گرفته شود را در نظر میگیریم. نشان دهید برای $\varepsilon \in (0,1)$ احتمال اینکه جواب درست باشد برابر با $1 - \varepsilon$ اگر $N \geq \left(\frac{1}{2}\right)\delta^{-2}\ln(\varepsilon^{-1})$.

۵) فرض کنید مجموعه‌داده $\{(x_i, y_i): i = 1, \dots, m\}$ داده شده است که به وسیله یک تابع غیرخطی می‌توان آنها را جدا کرد و $y_i = \pm 1$. مسئله Soft Margin SVM با کرنل گاوسی را در نظر بگیرید ($k(x_1, x_2) = e^{-\gamma(\|x_1 - x_2\|^2)} = \emptyset(x_1)^T \emptyset(x_2)$). همین‌طور فرض کنید که C ضریب خطا حاشیه باشد و $f(x_i) = w^T \emptyset(x_i)$ باشد. همین‌طور فرض کنید ضرایب لاگرانژ مسئله مورد نظر α_i ها باشند. فرض کنید dataset داده شده دارای خصوصیاتی به شکل گفته شده باشد. اگر x_i و x_j در یک کلاس باشند انگاه $\|x_i - x_j\| \leq s_1$ و اگر x_i و x_j در دو کلاس متفاوت باشند انگاه $\|x_i - x_j\| \geq s_2$. همین‌طور می‌دانیم که $s_1 \geq s_2$. در این سؤال ما تلاش میکنیم که برای هایپرپارامترهای C و γ مقادیری پیدا کنیم که هیچ داده ای در کلاس اشتباه قرار نگیرد.

الف) فرض کنید برای هر i دو مجموعه $L_i = \{j | y_i = y_j\}$ و $D_i = \{j | y_i = -y_j\}$ تشکیل داده شده است. حال عبارت زیر را ثابت کنید.

$$y_i(f(x_i) - y_i) \geq \sum_{j \in L_i} \alpha_j e^{-\gamma s_1^2} - \sum_{j \in D_i} \alpha_j e^{-\gamma s_2^2} - 1$$

ب) سعی کنید نامساوی بالا را گسترش داده و عبارت زیر را ثابت کنید.

$$y_i(f(x_i) - y_i) \geq Cm\gamma(s_2^2 - s_1^2)$$

راهنمایی: برای این قسمت از نامساوی زیر که برای توابع محدب (convex) برقرار است استفاده کنید.

$$f(x) \geq f(y) + f'(y)(x - y)$$

ج) تحلیل خود را از رابطه بالا بنویسید. (سعی کنید برای $C\gamma$ کرانی قرار دهید تا هیچ داده ای به اشتباه دسته‌بندی نشود.)

بخش دوم : Reading Assignment

در این بخش باید درباره یکی از موضوعات زیر که در کلاس تدریس نشده، تحقیق و مطالعه کنید. هدف از این تمرین آشنایی با روند یادگیری مطالب جدید و همین‌طور ارائه این مطالب است. گزارش تحویلی شما باید بین دو الی سه صفحه باشد.

۱. **One Class SVM (link)**
۲. **K-D Tree (link)**
۳. **KNN with Locality Sensitive Hashing (link)**
۴. **Reproducing Kernel Hilbert Space (link)**

- نحوه تخصیص موضوعات به شکل تصادفی بوده و از طریق فرمول (4 - n % 4) به دست می‌آید که n دو رقم آخر شماره دانشجویی شما است.
- می‌توانید از هر منبع دیگری در کنار منابع کمکی استفاده کنید.

بخش سوم : پیاده‌سازی

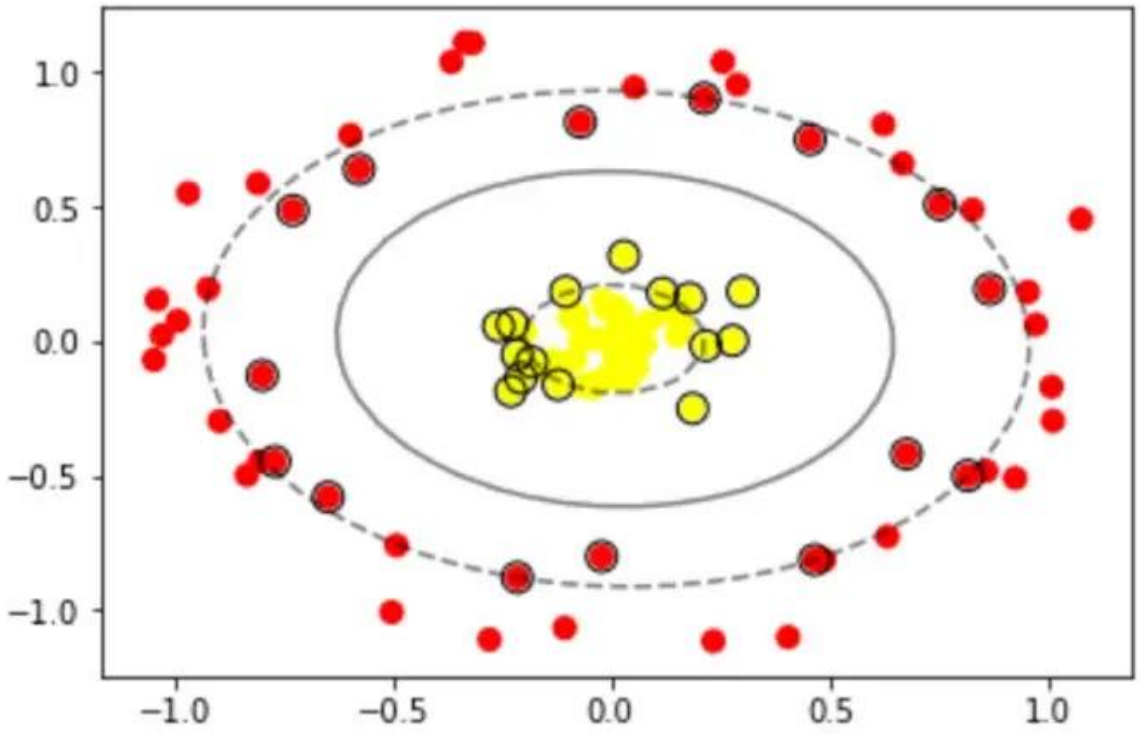
بخش پیاده‌سازی این تمرین از دو قسمت تشکیل شده که در ادامه به توضیح هر قسمت می‌پردازیم.

قسمت اول : مقایسه عملکرد مدل‌های کلاسیک Machine Learning

هدف از این قسمت پیاده‌سازی سه مدل KNN ، SVM و Naive Bayes و بررسی و ارزیابی عملکرد آن‌ها است. برای پیاده سازی این قسمت مجموعه‌داده Adult درنظر گرفته شده است که هدف از آن پیش بینی درآمد افراد با توجه به سایر ویژگی های مجموعه‌داده است. همین‌طور مجموعه‌داده های مربوط به این قسمت تمرین در پوشه Data قرار داده شده است. مجموعه‌داده Adult_TrainDataset برای آموزش مدل‌ها درنظر گرفته شده است. این مجموعه‌داده حاوی 32561 نمونه و در 2 کلاس مختلف طبقه بندی شده است. در این مجموعه‌داده ستون Income حاوی برچسب ها و سایر ستون‌ها ویژگی هایی هستند که شما می‌توانید برای پیاده سازی مدل از آن‌ها استفاده کنید. برای پیاده سازی این قسمت به موارد زیر توجه کنید.

- در این بخش می توانید از مدل های آماده کتابخانه SKlearn آماده استفاده کنید.
- در صورت نیاز می‌توانید قبل از آموزش مدل‌ها پیش‌پردازش‌های مختلفی مثل نرمال کردن مقدار ویژگی‌ها و ... روی مجموعه‌داده انجام دهید. همین‌طور با توجه به وجود مقادیر null در این مجموعه‌داده، باید مقادیر null با مقدار مناسب جایگزین کنید و یا آن ستون را حذف کنید.
- با توجه به این که برخی از ستون‌ها حاوی مقادیر categorical هستند و درصورتی‌که می‌خواهید از آن‌ها برای آموزش مدل استفاده کنید، باید مقادیر موجود را به کمک روش‌های موجود Encode و به مقادیر عددی تبدیل کنید.

- بعد از آموزش مدلها، از confusion matrix و معیارهای ارزیابی precision ، recall ، f1-score برای ارزیابی عملکرد مدل استفاده کنید. برای سادگی کار می‌توانید از کتابخانه sklearn کمک بگیرید. برای ارزیابی عملکرد مدلها از مجموعه داده Adult_TestDataset برای ارزیابی مدل استفاده کنید. در این بخش تحلیل خودتان را از بررسی ماتریس آشفتگی مدلی که پیاده‌سازی کردید و معیارهای ارزیابی به دست آمده، گزارش دهید.
- برای هریک از مدل‌ها برای رسیدن به نتیجه بهینه هایپرپارامترهای مدل را تنظیم کنید و عملکرد مدل را مجدداً و با استفاده از هایپرپارامترهای تنظیم شده بررسی کنید. همین‌طور برای مدل KNN و به‌ازای مقادیر از 1 تا 40 برای k ، نمودار دقت و خطای تست و آموزش را به‌ازای k های مختلف رسم کنید و نتایج خود را تحلیل کنید.
- برای دو مدل SVM و KNN به کمک کتابخانه های Seaborn و Matplotlib نحوه عملکرد مدل را قبل و بعد از تنظیم هایپرپارامتر ها visualize کنید. همین‌طور تحلیل خودتان را از نحوه عملکرد مدل و همین‌طور تاثیر تنظیم هایپرپارامترهای مدل گزارش دهید. برای انجام این بخش می‌توانید از لینک های مشخص شده کمک بگیرید.



- درنهایت در آخرین بخش تمرین از شما خواسته شده است که از مفهوم Ensemble Models استفاده کنید. تا این جای کار شما سه مدل KNN ، SVM و Naive Bayes را آموزش دادید و عملکرد آن‌ها را ارزیابی

کردید. در این بخش از شما انتظار می‌رود به کمک عملگرها منطقی مثل and / or / xor و یا رویکردهای دیگر مثل magority voting از پیش‌بینی‌های انجام شده توسط این مدل‌ها بهره بگیرید. لزوماً نیازی نیست از هر سه مدل استفاده کنید و می‌توانید دو مدل از بین آن‌ها را انتخاب کنید. رویکردی که استفاده کردید و همین‌طور بهترین نتایجی که به دست آوردید را گزارش دهید.

قسمت دوم : آشنایی با کتابخانه DESlib

Dynamic Selection به تکنیک‌هایی گفته می‌شود که در آن مدل‌های طبقه‌بندی پایه، در زمان آزمون به‌صورت پویا و بر اساس هر نمونه جدیدی که قرار است طبقه‌بندی شود، انتخاب می‌شوند. هدف از این تکنیک‌ها انتخاب بهترین و دقیق‌ترین مدل طبقه‌بندی برای پیش‌بینی یک نمونه test است. کتابخانه DESlib یکی از کتابخانه‌هایی است که مدل‌ها و تکنیک‌های مختلفی در این زمینه ارائه می‌دهد و تمامی مدل‌های موجود در این کتابخانه به فرم مدل‌های کتابخانه SKlearn پیاده‌سازی شده است. روند انجام این قسمت از تمرین به‌صورت زیر است :

- در جدول زیر 15 مورد از مدل‌های ارائه شده در این کتابخانه آورده شده است. با توجه به دو رقم آخر شماره دانشجویی شما و به کمک فرمول (n % 15 + 1) ابتدا یکی از این مدل‌ها به‌صورت تصادفی به شما اختصاص داده می‌شود.

1	k-Nearest Oracle-Eliminate (KNORA-E)	9	DES-Logarithmic
2	META-DES	10	DES-KNN
3	k-Nearest Oracle Union (KNORA-U)	11	DES Multiclass Imbalance (DES-MI)
4	Dynamic Ensemble Selection performance	12	Modified Rank
5	k-Nearest Output Profiles (KNOP)	13	Local Class Accuracy (LCA)
6	Randomized Reference Classifier (RRC)	14	Modified Local Accuracy (MLA)
7	DES-Kullback Leibler	15	Multiple Classifier Behaviour (MCB)
8	DES-Exponential		

- بعد از مشخص‌شدن مدل مربوط به خودتان، ابتدا حدود یک الی دو صفحه به معرفی مدل انتخاب شده بپردازید. برای معرفی مدل‌ها می‌توانید از لینک‌های مشخص شده و یا هر منبع دیگری استفاده کنید.
- سپس یک مجموعه‌داده به‌دلخواه انتخاب کنید. برای انتخاب مجموعه‌داده می‌توانید از سایت‌های Kaggle ، UCI و یا هر سایت دیگری استفاده کنید. در نهایت درصورتی‌که نتوانستید مجموعه‌داده مدنظرتان را پیدا

کنید می‌توانید از مجموعه‌داده قسمت قبلی تمرین استفاده کنید. بعد از انتخاب مجموعه‌داده لطفا آن را در یک یا چند پاراگراف معرفی کنید.

- درنهایت فرایند آموزش و ارزیابی مدل را روی مجموعه‌داده مشخص شده انجام دهید. درصورتی‌که تعداد نمونه‌های مجموعه‌داده انتخابی شما زیاد نیست از رویکرد k-fold-cross-validation و در غیر این از 20 درصد مجموعه‌داده برای ارزیابی مدل استفاده کنید. طبیعتا باتوجه‌به مجموعه‌داده انتخابی روند پیش‌پردازش مجموعه‌داده متفاوت خواهد بود و در گزارش ارسالی روند پیش‌پردازش مجموعه‌داده را توضیح دهید.
- همین‌طور برای رسیدن به نتیجه مطلوب و بهینه، هایپر پارامترهای مدل را با یکی از روش‌های موجود مثل GridSearch تنظیم کنید.
- درنهایت به کمک معیارهای ارزیابی عملکرد مدل را ارزیابی کنید و نظر خودتان را از کارکردن با این نوع مدل‌ها بیان کنید. به نظر شما عملکرد این مدل‌ها نسبت به مدل‌های کلاسیک چه تفاوتی دارد ؟! همین‌طور نقاط ضعف و قوت این مدل‌ها نسبت به مدل‌های کلاسیک به چه شکل است ؟!

معیار ارزیابی شما

بخش اول سؤالات تشریحی (۳۰ نمره) :

- سؤال اول (۱۰ نمره)
- سؤال دوم (۱۰ نمره)
- سؤال سوم (۱۰ نمره)
- سؤال چهارم (۱۵ نمره امتیازی)
- سؤال پنجم (۱۵ نمره امتیازی)

بخش دوم Reading Assignment (۲۰ نمره)

بخش سوم پیاده‌سازی (۱۵۰ نمره) :

- قسمت اول (۸۰ نمره)
- قسمت دوم (۷۰ نمره)

این تمرین حدود ۲ نمره از ۲۰ نمره نهایی شما را شامل می‌شود.

نکات تکمیلی

- ✓ انجام این تمرین بسته به تسلط شما به مطالب درس و زبان پایتون حداقل بین ۴ الی ۷ روز از وقت مفید شما را خواهد گرفت. به همین علت انجام این تمرین را به‌روزهای پایانی موکول نکنید. همین‌طور باتوجه‌به برنامه فشرده کلاس و حجم زیاد مطالب، مهلت تحویل این تمرین تمدید نخواهد شد.
- ✓ **ارسال گزارش اجباری است.** نکته مهم در گزارش‌نویسی و سؤال تشریحی روشن بودن پاسخ است نه حجم زیاد، اگر فرضی برای حل سؤال استفاده می‌کنید حتماً آن را ذکر کنید، و پاسخ نهایی را به‌صورت واضح بیان کنید. گزارش کد و پاسخ سؤال تشریحی به‌صورت فایل pdf باشد.
- ✓ هرگونه شباهت در گزارش و پاسخ تشریحی به منزله تقلب است و کل نمره تمرین را نخواهید گرفت.
- ✓ فایل pdf مربوط به بخش اول و دوم تمرین و همین‌طور گزارش مربوط به بخش پیاده‌سازی را به همراه کدها را به‌صورت یک جا در قالب یک فایل zip در سامانه کورسز آپلود کنید (نام فایل = شماره دانشجویی)
- ✓ در صورت هرگونه ابهام درباره این تمرین می‌توانید در کلاس رفع اشکال سؤالات خودتان رو بپرسید و یا از طریق ایمیل‌های زیر با ما در ارتباط باشید.

مریم نظرلو : maryamnazarloo@aut.ac.ir

رئوف زارع : Raoofmoayedi@gmail.com

محمدعلی سفیدی اصفهانی : mohammadali.esfahani@aut.ac.ir

با آرزوی سلامتی و موفقیت برای شما عزیزان