به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر

یادگیری ماشین

پروژه پایانی

دی 99

فهرست مطالب

مقدمه و توضیحات کلی

در طول درس با روش ها و مدل های متفاوتی برای کاربرد های طبقه بندی و خوشه بندی آشنا شده اید. در این مرحله همواره چالش انتخاب مدل بهینه برای حل یک مسئله با روش های یادگیری ماشین وجود دارد. در این تمرین برای یک مسئله طبقه بندی روش های مختلفی که در درس وجود داشته است (شامل قسمت های اختیاری درس) را با یکدیگر مقایسه خواهید کرد و مناسب ترین روش را انتخاب می‌کنید. علاوه بر این یک مقاله که در مورد همین مسئله است را مطالعه کرده و مطابق با خواسته های بخش مربوطه مطالبی را ارائه می‌دهید.

آشنایی با مقالات مربوط

در حل یک مسئله، مطالعه تحقیقات انجام شده در آن موضوع اهمیت بسیاری دارد. مطالعه مقاله های مربوط از چند جهت در پیشبرد تحقیق شما مفید خواهد بود. در درجه اول با اطلاع ار تحقیقات پیشین مطمئن خواهید شد که روش/ایده/مدل شما قبلا توسط محقق دیگری امتحان و بررسی نشده باشد. علاوه بر آن در انتهای مطالب علمی گاها نویسندگان مسیری برای ادامه و پیشبرد تحقیقات ارائه می‌دهند که این پیشنهادات می‌تواند در شکل دهی ایده های شما و مسیری که برای تحقیقات خود انتخاب می‌کنید بسیار مفید باشد.

در این قسمت از تمرین، از شما خواسته می‌شود تا یکی از مقالاتی که در اختیار شما قرار داده شده است را مطالعه کنید و خلاصه ای از آن تهیه کنید. در این خلاصه باید چند قسمت اصلی متداول را حتما ذکر کنید. لذا خلاصه شما باید پاسخگو و شامل قسمت های زیر باشد.

۱. خلاصه ای از مقدمه

۲. در تحقیق از چه مجموعه داده‌ای استفاده شده است؟

۳. آیا داده های توسط خود نویسندگان جمع آوری شده است یا خیر؟

۴. از چه روش هایی برای پیش‌پردازش داده ها و انتخاب ویژگی ها استفاده شده است؟

۵. از چه مدل هایی برای طبقه بندی/خوشه بندی/درون یابی استفاده شده است؟

۶. عملکرد مدل ها با چه اطلاعاتی گزارش شده است؟ آیا این گزارش دقیق است یا خیر؟ در صورتی که پاسخ منفی است، شیوه بهتری برای گزارش عملکرد مدل پیشنهاد دهید.

۷. نتیجه گیری و دست‌آورد های پژوهش

در انتخاب مقالات تلاش شده است تا موضوعات تا جای ممکن به قسمت پیاده سازی تمرین شباهت داشته باشد تا پیش از پیاده سازی دید بهتری نسبت به مسئله داشته باشید. با این حال اگر تمایل به بررسی مقاله دیگری با موضوع مشابهی را دارید، می‌توانید درخواست خود را به دستیاران آموزشی(به همراه مقاله مورد نظر) اطلاع دهید.

معرفی مجموعه داده

در سال‌های اخیر استفاده از سیستم های تشخیص و کنترل از راه دور برای تشخیص بیماری پارکینسون بر اساس اندازه‌گیری اختلالات در عملکرد قسمت موتور کورتکس مغز افزایش یافته است. در حدود 90 درصد از بیماران، نشانه‌های اختلال در صحبت کردن در همان مراحل ابتدایی بیماری آشکار می‌شود. بنابراین تحلیل و پردازش صوت یک روش مناسب برای تشخیص زود هنگام بیماری می تواند باشد.

مجموعه داده این پروژه ویژگی‌های استخراج شده از صدای 188 بیمار پارکینسونی (107 مرد و 81 زن) و 64 فرد سالم (23 مرد و 41 زن) را در بر دارد. الگوریتم‌های استخراج ویژگی متفاوتی بر روی صوت‌های جمع آوری شده اعمال شده تا اطلاعات بالینی مناسبی برای تشخیص بیماری پارکینسون بدست آید. در جدول صفحه بعد توضیحات کلی درباره ویژگی‌های این مجموعه داده قابل مشاهده است.

Feature Set	Measure	Explanation	# of features
Baseline Features	Jitter variants	Jitter variants are employed to capture the instabilities that occurred in the oscillating pattern of the vocal folds and this feature subset quantifies the cycle-to-cycle changes in the fundamental frequency.	5
	Shimmer variants	Shimmer variants are also employed to capture instabilities of the oscillating pattern of the vocal folds, but this time this feature subset quantifies the cycle-to-cycle changes in the amplitude	6
	Fundamental frequency parameters	The frequency of vocal fold vibration. Mean, median, standard deviation, minimum and maximum values were used.	5
	Harmonicity parameters	Due to incomplete vocal fold closure, increased noise components occur in speech pathologies. Harmonics to Noise Ratio and Noise to Harmonics Ratio parameters, which quantify the ratio of signal information over noise, were used as features.	2
	Recurrence Period Density Entropy (RPDE)	RPDE gives information about the ability of the vocal folds to sustain stable vocal fold oscillations and it quantifies the deviations from F0.	1
	Detrended Fluctuation Analysis (DFA)	DFA quantifies the stochastic self-similarity of the turbulent noise.	1
	Pitch Period Entropy (PPE)	PPE measures the impaired control of fundamental frequency F0 by using a logarithmic scale.	1
Time-Frequency Features	Intensity Parameters	Intensity is related to the power of speech signal in dB. Mean, minimum, and maximum intensity values were used.	3
	Formant Frequencies	Frequencies amplified by the vocal tract, the first four formants were used as features.	4
	Bandwidth	The frequency range between the formant frequencies, the first four bandwidths were employed as features.	4
Mel Frequency Cepstral Coefficients (MFCCs)	MFCCs	MFCCs are employed to catch the PD effects in the vocal tract separately from the vocal folds	84
Wavelet Transform based Features	Wavelet transform (WT) features related to F0	WT features quantify the deviations in F0	182
Vocal Fold Features	Glottis Quotient (GQ)	GQ gives information about the opening and closing durations of the glottis. It is a measure of periodicity in glottis movements.	3
	Glottal to Noise Excitation (GNE)	GNE quantifies the extent of turbulent noise, which is caused by incomplete vocal fold closure, in the speech signal.	6
	Vocal Fold Excitation Ratio (VFER)	VFER quantifies the amount of noise produced due to the pathological vocal fold vibration by using nonlinear energy and entropy concepts.	7
	Empirical Mode Decomposition (EMD)	EMD decomposes a speech signal into elementary signal components by using adaptive basis functions and energy/entropy values obtained from these components are used to quantify noise.	6

معرفی مسئله مورد بررسی

در ادامه با استفاده از مجموعه داده معرفی شده در قسمت قبل، مدل‌هایی برای تشخیص بیماری ارائه و یک مسئله طبقه بندی در دنیای واقعی را حل کنید. با توجه به این که کارایی مدل نهایی ارائه شده اهمیت به سزایی در تشخیص دقیق دارد، روش های مختلف پیش پردازش، انتخاب ویژگی و طبقه بندی را مقایسه کنید و در نهایت بهترین پروسه(روش کاهش بعد و طبقه‌بندی) پردازش اطلاعات برای مجموعه داده را پیشنهاد دهید.

دقت کنید که هدف اصلی آن است که میان گزینه های موجود روشی را بیابید که عملکرد بهتری را ارائه دهد. بنابراین باید با کوشش و خطا(همراه با استدلال و شهودی که در مورد مدل ها دارید) این مدل را پیدا کنید. برای سنجش عملکرد مدل می‌توانید از معیاری های متفاوتی که در زیر به آن‌ها اشاره شده است استفاده کنید.

- accuracy
- confusion matrix
- ROC curve
- AUC
- F1 Score
- Log Loss/Binary Cross Entropy
- Categorical Cross Entropy

دقت شود که لزومی به استفاده از تمامی این متریک ها نیست! بلکه لازم است تعدادی را انتخاب کنید و مدل های مختلف را بر اساس این متریک ها با یکدیگر مقایسه کنید. برای اطلاعات بیشتر در مورد هر یک از متریک ها می‌توانید به این منبع مراجعه کنید.

روش های پیش‌پردازش و انتخاب ویژگی

۱. مجموعه داده را یکه سازی کنید.

۲. با توجه به این که در این مجموعه داده تعداد نمونه‌های دو کلاس نامتوازن است، چه روشی را برای مواجه با این مشکل انتخاب می‌کنید؟ (روش‌های متفاوتی مانند استفاده از متریک مناسب، ایجاد سمپل جدید برای داده یادگیری و ... را می‌توان استفاده کرد و یا با توجه به جنس مساله از عدم توازن چشم پوشی کرد.) به دلخواه روشی را انتخاب کنید و توجه داشته باشید انتخاب شما، مقایسه عملکرد طبقه بندها را در قسمت بعد تحت تاثیر قرار می‌دهد.

۳. از ابعاد مجموعه داده مشخص است که با نحسی ابعاد مواجه هستیم. برای رفع این مشکل از روش‌های زیر برای کاهش ابعاد استفاده کنید.

- LDA
- ICA
- PCA with Whitening
- PCA without Whitening
- Sequential Backward Feature Elimination
- Autoencoders (امتیازی)

توجه کنید که می‌توانید از ترکیبی از روش ها برای کاهش بعد استفاده کنید.

روش های طبقه بندی

روش های طبقه‌بندی که در درس مطرح شده است را می‌توان به دو دسته کلی Generative و Discriminative تقسیم کرد. در هر دسته گزینه هاي متعددي وجود دارد که به صورت خلاصه در ادامه به آنها اشاره خواهد شد.

طبقه بندهای Generative :

الگوریتم‌های Generative سعی بر مدل کردن کلاس، بر اساس ویژگی های کلاس دارند. به طور خلاصه این الگوریتم ها نحوه تولید داده ورودی توسط کلاس را مدل می کنند و که هنگامی که نمونه جدیدی مشاهده می شود سعی در پیش بینی محتمل ترین کلاس برای این نمونه را دارد. در واقع با یادگیری محیط و ایجاد مدل احتمالاتی طبقه بندی انجام می شود.

۱. یک طبقه بند بهینه ي بیز با روش Parzen Window براي تخمین چگالی احتمال طراحی کنید.

۲. سوال قبل را با استفاده از روش KNN براي تخمین چگالی احتمال تکرار کنید.

۳. با استفاده از GMM طبقه بندی برای مجموعه داده طراحی کنید.

طبقه بندهای Discriminative :

الگوریتم های Discriminative سعی بر پیدا کردن مرزهای کلاس ها با استفاده از داده‌های یادگیری دارند و بر اساس این که نمونه جدید مشاهده شده در کدام طرف مرز قرار گرفته است طبقه بندی صورت می گیرد. در زیر چند نمونه از این طبقه بند ها معرفی شده است که باید میان آن ها مقایسه انجام شود و مدلی که بهترین عملکرد را دارد معین شود.

- Logistic regression
- SVM
- Decision Tree
- KNN
- MLP (امتیازی)
- RBF (امتیازی)

همانطور که پیش از این بیان شد، باید عملکرد این مدل ها را با استفاده از متریکی که در قسمت های پیش انتخاب کرده اید را بسنجید و در گزارش بیاورید. همچنین عملکرد مدل های Generative و Discriminative را مقایسه کنید و بیان کنید که از کدام دسته برای طبقه بندی نهایی استفاده خواهید کرد.

نکاتی در مورد عملکرد قابل قبول برای مدل نهایی:

با توجه به انتخاب های متعددی که در قسمت های پیشین وجود دارد، ارائه یک دقت(متریک عملکرد) ثابت مناسب نخواهد بود. مهم ترین قسمت آن است که به درستی مدل های مختلف را آموزش دهید و با یکدیگر مقایسه کنید و در نهایت بهترین مدلی که بدست آورده اید را ارائه دهید. دقت این مدل باید بیشتر از ۸۰ درصد روی داده های آموزش باشد.

روش های یادگیری تجمیعی(امتیازی)

در قسمت های پیشین مدلی برای حل مسئله‌ی ارائه شده یافتید. در تمام روش های پیشین ار یک نوع طبقه‌بند برای تصمیم گیری نهایی استفاده شده است. با این حال لزومی به استفاده از یک مدل برای تصمیم گیری وجود ندارد. این ایده ساده مدخل روش های یادگیری تجمیعی است. در این روش‌ها با این پیش‌فرض که چندین طبقه بند با عملکرد های متوسط بهتر از یک طبقه بند قوی عمل خواهند کرد، بر روی قسمت های مختلف داده، مدل های متفاوتی آموزش داده می‌شود.

۱. با مطالعه مطالبی که در اینجا قرار دارد، توضیح دهید که چرا استفاده از یادگیری تجمیعی میتواند مفید باشد.

۲. با مطالعه مطالبی که در اینجا قرار دارد، در مورد BAGGING توضیح دهید.

۳. تلاش کنید با استفاده از روش های یادگیری تجمیعی عملکرد مدلی که در قسمت های پیش استفاده کرده اید را بهبود دهید. در این قسمت باید گزارش کنید که از چه روش هایی استفاده کرده اید و این روش ها تا چه اندازه بر عملکرد مدل تاثیر داشته است.

نکات پایانی

- مهلت تحویل این پروژه ۲۶ دی است اما شما فرصت دارید تا ۱۷ بهن پروژه را تحویل دهید. در صورت تحویل تمرین تا تاریخ ۲۶ دی، نمره شما شامل ۲۰ درصد نمره امتیازی خواهد بود. این مقدار نمره امتیازی به صورت خطی تا ۱۷ بهمن کاهش پیدا می‌کند و به ۰ میرسد، لذا در صورتی که پروژه را ۱۷ بهمن نیز تحویل دهید، می‌توانید نمره کامل کسب کنید.
- پروژه را می‌توانید در گروه های دو نفره یا تک نفره انجام دهید. امکان انجام تمرین در گروه های سه نفره و بیشتر وجود ندارد.
- با توجه به این که در این تمرین انتخاب های متعددی در هر قسمت دارید، معیار نمره دهی درک و توضیحات شما از مطالب ارائه شده در درس می‌باشد. بنابراین با دقت کامل گزارش خود را تهیه کنید.
- گزارش شما در فرآیند تصحیح از اهمیت ویژه‌اي برخوردار است. لطفاً تمامی نکات و فرض هایی که برای پیاده سازی ها و محاسبات خود در نظر میگیرید را در گزارش ذکر کنید.
- در گزارش خود برای تصاویر زیرنویس و برای جداول هم بالانویس اضافه کنید.
- الزامی به ارائه توضیح جزئیات پیادەسازی در گزارش نیست. اما باید نتایج بدست آمده را گزارش و تحلیل کنید.
- برای انجام پروژه استفاده از کتاب‌خانه ها منعی وجود ندارد.
- لطفاً گزارش، فایل کدها و سایر ضمائم مورد نیاز را با الگو PROJECT_[StudentNumber].zip زیر در سامانه مدیریت دروس بارگذاری نمایید.

- در صورت وجود هرگونه ابهام یا مشکل میتوانید از طریق رایانامه های زیر با دستیاران آموزشی مربوطه سجاد پاکدامن و نیوشا میرحکیمی در تماس باشید. همچنین برای کیفیت ارتباطی بهتر دستیاران آموزشی با شما، متنی در اینجا قابل دسترس است که سوال های پرتکرار از پروژه همراه پاسخ های آن‌ها به مرور زمان اضافه خواهد شد. بنابراین حتماً پیش از ارسال پیام آن را بازبینی کنید.

sj.pakdaman@ut.ac.ir و mirhakimi@ut.ac.ir

-- موفق باشید