



دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده‌ی مهندسی برق و
کامپیوتر
گروه کنترل



تمرین سری ۵

موضوع تمرین

ترکیب داده

استاد درس: دکتر مشیری

تاریخ تحویل: خرداد ۱۴۰۳

تمرین ۱

روش‌های $Adaboosting$ و $Bagging$ ، $Boosting$ ، $stacking$ را مختصراً با ذکر مراحل توضیح داده و تفاوت‌های آن‌ها را با یکدیگر بیان کنید.

تمرین ۲

در دیتاست JIGSAWS، می‌خواهیم کارآموزان جراحی را بر اساس میزان مهارت آن‌ها در تمرین‌هایی از قبیل: عبور دادن نخ از سوراخ، بخیه زدن و گره زدن دسته بندی نمائیم. در این دسته بندی سه گروه بر اساس میزان مهارت مشخص شده اند:

۱. مبتدی

۲. متوسط

۳. خبره

برای تصمیم‌گیری درباره اینکه هر فرد دارای چه سطحی از مهارت در این تمرینات است از ۳ تصمیم‌گیر مختلف استفاده کرده ایم. اگر با رویکرد روش قالب تصمیم^۱ به این مسئله نگاه کنیم، برای هر یک از دسته‌ها قالب تصمیم‌های زیر به دست آمده است.

$$DT_1 = \begin{bmatrix} 0.3 & 0.2 & 0.5 \\ 0.6 & 0.3 & 0.1 \\ 0.2 & 0.4 & 0.4 \end{bmatrix}, DT_2 = \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.5 & 0.4 & 0.1 \\ 0.6 & 0.1 & 0.3 \end{bmatrix}, DT_3 = \begin{bmatrix} 0.2 & 0.3 & 0.3 \\ 0.1 & 0.5 & 0.4 \\ 0.4 & 0.4 & 0.2 \end{bmatrix} \quad (۱)$$

و ماتریس تصمیم مربوط به یکی از شرکت‌کنندگان به صورت زیر است://

$$DP(x) = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.6 & 0.2 & 0.2 \\ 0.5 & 0.1 & 0.4 \end{bmatrix} \quad (۲)$$

با استفاده از معیار شباهت $Squared Euclidean distance$ تعیین کنید این فرد احتمالاً جزو کدامیک از گروه‌های تعیین شده است؟

¹Decision Template

تمرین ۳

مطابق مطالب گفته شده در مبحث ترکیب اطلاعات در سطح تصمیم، در حالتی که خروجی طبقه‌بندها عدد واضح^۲ باشد از روش‌های Majority Voting و Weighted Majority Voting می‌توان استفاده کرد. حال تحقیق کنید که اگر خروجی طبقه‌بندها به صورت عدد واضح نبوده و مقادیر احتمالاتی باشد از چه روشی با مفهومی مشابه روش‌های نامبرده شده می‌توان استفاده کرد؟ این روش را به صورت فرمول‌بندی شده توضیح دهید. (حتما مراجع استفاده شده برای پاسخ به این سوال را به طور دقیق در انتهای سوال ذکر کنید)

تمرین ۴

دیتاست *Sis Fall* را از لینک مورد نظر (کلیک کنید) دانلود نمایید. این دیتاست شامل ۲۰۹۷۱۵۰ سطر و ۱۰ ستون است. این مجموعه داده برای تشخیص افتادن افراد، جمع‌آوری شده است. در جریان جمع‌آوری این داده ۳۸ داوطلب شرکت کرده‌اند که این داوطلبین ۱۵ نوع سناریوی افتادن و ۱۹ سناریو از دسته فعالیت‌های روزانه شبیه به افتادن را انجام داده‌اند. ستون‌های این دیتاست شامل داده‌های ۳ محور $x-y-z$ از ۳ سنسور (۹ خروجی) و برچسب‌های مربوط به هر داده (*Fall or Not Fall*) است.

الف) از بین طبقه‌بندهای *Support Vector Machine*، *K-Nearest Neighbors*، *Logistic regression*، *Naive Bayes*، *Random Forest* و *XGBoost* سه طبقه‌بندی که از نظر شما برای این کاربرد و دیتاست مناسب است، انتخاب نموده و با ارائه دلیل انتخاب خود را توجیه کنید.

ب) حال داده‌های هر ۳ سنسور را با استفاده از فیلتر کالمن با هم در سطح داده ترکیب نمایید.

ج) سپس داده‌های ترکیب شده در سطح داده را در هر ۳ محور با پنجره زمانی به طول ۱ ثانیه با همپوشانی ۵۰٪ بین هر دو پنجره متوالی، پنجره بندی نمایید. سپس ویژگی‌های آورده شده در جدول زیر را برای هر پنجره از داده محاسبه کرده و ذخیره کنید. در مرحله بعدی داده‌های به دست آمده در فضای ویژگی را به دو بخش داده‌های *train/test* با نسبت ۸۰ به ۲۰ تقسیم کنید. دقت نمایید که فرکانس نمونه برداری سنسورها در این مجموعه داده 200 Hz است. (فایل کد به همراه توضیحات کامل، در گزارش آپلود شود).

د) نهایتا صحت نتایج هر طبقه‌بند را در قالب جدولی گزارش کرده و نتایج این ۳ طبقه‌بند را با روش‌های *Bagging* و *Majority Voting* و *Bagging+Weighted Majority Voting* در سطح طبقه‌بندها (تصمیم) ترکیب نمایید. نتایج این ترکیب را در قالب جدول گزارش کرده و نتایج این دو روش را با هم و با حالت ترکیب نشده مقایسه کنید.

ه) (امتیازی) تحقیق کنید که در روش *Stacking* در چه حالتی ریسک بروز ابربرازش^۳ وجود دارد؟ (حتما مراجع استفاده

^۲crisp

^۳over fitting

جدول ۱: ویژگی‌های استخراج شده

Type	Code	Feature	Equation
Amplitude	F1	Sum vector magnitude	$F_1[k] = \text{RMS}(\tilde{a}[k]) = \sqrt{a_x^2[k] + a_y^2[k] + a_z^2[k]}$
Orientation	F2	Angle between z-axis and vertical	$F_2[k] = \text{atan2} \left(\sqrt{(\tilde{a}_x[k])^2 + (\tilde{a}_z[k])^2}, -\tilde{a}_y[k] \right)$
Statistics	F3	Standard deviation magnitude	$F_3[k] = \sqrt{\sigma_x^2[k] + \sigma_y^2[k] + \sigma_z^2[k]}$

شده در این سوال را به طور دقیق در انتهای سوال ذکر کنید)

ی) (امتیازی) یک طبقه‌بند از نوع *Convolutional Neural Network* به سه طبقه بندی انتخابی خود در بخش ب بیافزاید. نتایج این طبقه‌بند را با سایر طبقه‌بندها مقایسه و روش ترکیب اطلاعات Bagging را برای ۴ طبقه‌بند حاصل پیاده‌سازی نمایید.

نکات مهم در حل تمرین‌ها

۱. کلیدی فایل‌های مربوط به هر سری تمرین را در یک پوشه قرار داده و آن را با نام HWNumber-StudentNumber به صورت زیپ شده در سایت درس بارگزاری کنید. توجه داشته باشید که فایل‌های شبیه‌سازی باید به همراه فایل گزارش ارسال شده باشند. همچنین در نظر داشته باشید که شبیه‌سازی بدون گزارش مناسب فاقد ارزش است.
۲. تمرینات فقط به صورت تایپ شده در فرمت لیتک داده شده تحویل گرفته می‌شوند.
۳. در هر سری از تمرین‌ها ممکن است تعدادی سوال امتیازی مطرح شود که بسته به صلاحدید تا ۱۰ درصد نمره‌ی آن تمرین، امتیاز مازاد برای آن در نظر گرفته می‌شود.
۴. برای تمرین‌های این درس مجموعاً یک هفته بودجه‌ی تاخیر در نظر گرفته شده است. در نتیجه امکان تمدید مهلت تمرین‌ها وجود ندارد.
۵. مشورت در مورد روال حل تمرین مجاز است ولی نتیجه‌گیری و حل نهایی هر تمرین باید به صورت انفرادی انجام پذیرد. فرض کنید پس از تصحیح تمرین‌ها مشخص می‌شود دانشجوی a و b تمرین‌ها را کاملاً به کمک هم حل کرده‌اند نمره‌ی این عزیزان به این صورت محاسبه خواهد شد:

$$StudentAscore : S_a$$

$$StudentBscore : S_b$$

$$CHEAT! \rightarrow S_a = S_b = 0.6 * \frac{(\min(S_a, S_b))}{2}$$

۶. بسته به نوع تمرین‌ها برخی از آن‌ها ممکن است ارائه‌ی شفاهی داشته باشند که زمانبندی آن از قبل اعلام خواهد شد.
۷. لطفاً هر گونه ابهام یا سوالی در مورد حل این تمرین را با تدریس‌یار مربوطه در میان بگذارید:

Email address: elahe.radmanesh@ut.ac.ir

Email caption: DF-0203

لطفاً ایمیل‌ها را با کپشن مشخص شده بفرستید تا احتمال گم شدن و بی‌پاسخ ماندن آنها به حداقل برسد.

موفق باشید.