پروژه اول درس یادگیری ماشین

پریا پاسه‌ورز

شماره دانشجویی: 810101393

سوال اول

الف) توضیح درباره cross validation و توضیح انواع آن:

Cross validation معیاری برای ارزیابی کارکرد مدل‌مان روی دیتا مشاهده نشده است. برای این کار، اطلاعات مسئله را به چندین زیرمجموعه یا fold می‌شکنیم و و از یکی از این foldها برای validation و از بقیه برای train کردن مدل‌مان استفاده می‌کنیم. باید این پروسه چندین بار تکرار شود و در هر مرحله یک fold جدید را برای validation در نظر بگیریم. در نهایت، میانگین نتیجه‌ای که از هر بار تکرار این مرحله گرفته‌ایم، performance نهایی مدل را مشخص می‌کند.

مزیت اصلی این روش این است که از overfitting تا حد زیادی جلوگیری می‌کند، چون overfitting وقتی اتفاق می‌افتد که مدل روی دیتای train عملکرد خوبی دارد، ولی در مواجهه با دیتای test ضعیف عمل می‌کند. چون در این روش دیتای validation set مقادیر مختلفی را اختیار می‌کند، پس به ما دید بهتری نسبت به عملکرد مدل در برابر دیتای test می‌دهد.

عیب این روش این است که چون باید cycleهای متوالی روی اطلاعات process انجام دهیم، پس زمان زیادی می‌برد. همچنین هزینه محاسباتی بالایی نیز دارد، پس به resource زیادی احتیاج دارد.

انواع cross validation:

1. Holdout Cross validation

در این روش، بخشی از اطلاعات موجود به بخش train(معمولا 70 یا 80 درصد) و بقیه داده‌ها به test اختصاص یافته می‌شود. مشکل اصلی این است که ممکن است اطلاعاتی که برای بخش test اختصاص یافته است شامل اطلاعات مهمی باشد که در train شدن مدل آورده نشود و عملکرد مدل را دچار اشکال کند، به عبارت دیگر، ممکن است دیتای train به خوبی کل دیتا ما را represent نکند.

2. K fold cross validation

در این روش، دیتا را به k fold تقسیم می‌کنیم و هر بار از یکی برای test و از بقیه برای train کردن مدل استفاده می‌کنیم. این مراحل k بار تکرار می‌شود. در نهایت، از نتیجه به دست آمده میانگین گرفته می‌شود.

معمولا مقدار k عددی زیر 10 در نظر گرفته‌می‌شود و در حقیقت یک hyperparameter است.

این روش در مدلهایی که دیتای محدود دارند کاربردی است، در واقع با در نظر گرفتن subsetهای مختلف برای دیتاهای train و test، حداکثر استفاده از دیتای موجود انجام می‌شود و اطمینان حاصل می‌کنیم که هر data point هم در بخش test و هم در بخش train حضور دارد.

عیب این روش این است که زمان‌بر خواهد بود، چون نیاز است که الگوریتم یکبار از اول برای هر k اجرا شود، اما نتیجه مطلوبی خواهد داشت.

3. Stratified Cross validation

هدف اصلی این روش این است که هر fold، distribution class مشابهی با بقیه کلاسها داشته باشد، یعنی حضور نمونه‌هایی از کلاسهای مختلف به طور همگون در هر fold ظاهر شود.

پس برای استفاده از این روش، دیتا را مطابق توضیحات داده شده به k fold تقسیم می‌کنیم و هر بار از یکی برای test و از بقیه برای train کردن مدل استفاده می‌کنیم. این مراحل k بار تکرار می‌شود.

این روش زمانی کاربردی است که Balance بودن foldها در عملکرد مدل در برابر دیتای مشاهده نشده موثر است.

4. Leave one out cross validation

در این روش اگر فرض کنیم که n تا sample داشته باشیم، هر بار روی n-1 تا از sampleها، training را انجام می‌دهیم و یک sample را برای تست باقی می‌گذاریم. این روند برای تمام sampleها تکرار می‌شود.

مزیت این روش این است که از تمام sampleها استفاده می‌شود پس bias کمی دارد.

عیب این روش این است که چون هربار کل n-1 تا sample را در برابر یک sample تست می‌کنیم، ممکن است آن یک sample در واقع یک Outlier باشد و سبب ناپایداری مدل شود. علاوه بر این، پردازش سنگینی نیز احتیاج دارد چون تمام دیتاها یکبار به عنوان test sample استفاده می‌شوند. پس این روش بیشتر زمانی کاربرد دارد که دیتاست کوچک باشد یا بخواهیم تاثیر هر دیتای را روی مدلمان مشاهده کنیم.

در واقع این روش همان k fold cross validation است که در اینجا k = 1.

5. Leave P-out cross validation

در این روش، همیشه به تعداد p sample، که p یک عدد ثابت است، از sampleهای موجود کنار گذاشته می‌شود که برای تست کردن استفاده می‌شوند و بقیه sampleها در Training استفاده می‌شوند. این مرحله برای هر ترکیب p تایی تکرار می‌شود. مزیت این روش این است که یک تخمین unbias از مدل می‌دهد، چون زیرمجموعه‌های مختلفی از دیتا را در نظر می‌گیرد.

کاربرد آن زمانی است که دیتاست بزرگ باشد و یا منابع محاسباتی و زمان اندک باشد.

مزیت این روش این است که چون subsetهای مختلفی از اطلاعات را در نظر می‌گیریم، پس bias کمی خواهیم داشت و در نهایت میانگین نتایج به دست آمده به عنوان نتیجه نهایی اعلام می‌شود.

ب) نشان دهید اگر عناصر هر بعد یک متریک فاصله اقلیدسی d بعدی را در یک عدد حقیقی غیر صفر ضرب کنیم، این متریک همچنان یک فاصله استاندارد خواهد بود:

$$\boldsymbol{D(x,y) = \sqrt{\sum_{k=1}^{d} (x_k - y_k)^2}}$$

$$x`_k = a_k x_k \ \ for\ k = 1,2,\dots,d$$

یک فاصله متریک استاندارد ویژگیهای زیر را داراست:

1. non-negativity:

$$d(x,y) \geq 0 \ and\ d(x,y) = 0 \ if\ and\ only\ \ if\ x = y$$

2. symmetry:

$$d(x,y) = d(y,x)$$

3. triangle inequality:

$$d(x,z) \leq d(x,y) + d(y,z)\ for\ all\ points\ x,y,z$$

4. identity of indiscernible:

$$d(x,y) = 0 \ if\ and\ only\ \ if\ x = y$$

پس باید نشان دهیم اگر عناصر هر بعد این متریک را در یک عدد حقیقی غیر صفر ضرب کنیم، همچنان این چهار ویژگی برقرار خواهد بود.

ویژگی اول:

$$\boldsymbol{D(x',y) = \sqrt{\sum_{k=1}^{d} (x`_k - y_k)^2}}$$

$$x`_k = a_k x_k \ \ for\ k = 1,2,\dots,d$$

همچنان رابطه $D(x,y) \geq 0$ صادق است، چرا که این فاصله از طریق محاسبه یک عبارت رادیکالی محاسبه می‌شود که حاصل آن همیشه بزرگ‌تر یا مساوی صفر است.

پس این ویژگی برقرار است.

ویژگی دوم:

$$\boldsymbol{D(x', y)} = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{d}} (x`_k - \boldsymbol{y_k})^2}$$

$$\boldsymbol{D(y, x')} = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{d}} (y_k - \boldsymbol{x'_k})^2}$$

تفاوت $(x`_k - \boldsymbol{y_k})$ **با** $(y_k - \boldsymbol{x'_k})$ به ازای هر k در این است که یکی مقدار منفی دیگری را دارد. دقت کنید فاصله اقلیدسی جذر حاصل جمع توان دو هر یک از این عبارات است، پس تاثیر علامت در آن از بین می‌رود و این ویژگی برقرار است.

ویژگی سوم:

می‌دانیم تعریف فاصله اقلیدسی در واقع معادل محاسبه نرم اختلاف x , y است، یعنی:

$$\boldsymbol{D(x, y)} = \sqrt{\sum_{\boldsymbol{k=1}}^{\boldsymbol{d}} (\boldsymbol{x_k} - \boldsymbol{y_k})^2} = \boldsymbol{||x - y||}$$

یعنی کافی است نشان دهیم:

$$\boldsymbol{||x' - z||} \leq \boldsymbol{||x' - y||} + \boldsymbol{||y - z||}$$

حال رابطه نامساوی مثلث را برای بردارهای a , b در نظر بگیرید:

$$||a + b|| \leq ||a|| + ||b||$$

پس کافی است قرار دهیم:

$$a = x' - y$$

$$b = y - z$$

در این صورت صحت ویژگی سوم نیز ثابت می‌شود.

ویژگی چهارم:

طرف اول:

$$d(x', y) = 0 \rightarrow x' = y$$

$D(x, y)$ در واقع حاصل یک رادیکال است که زیر این رادیکال جمع تعدادی عبارت توان دو قرار دارد. برای اینکه حاصل جمع تعدادی عبارت توان دو صفر شود، باید تک‌تک این عبارات صفر باشند، یعنی داشته باشیم:

$$(x`_k - \boldsymbol{y_k})^2 = \boldsymbol{0\ for\ k = 1, 2, \dots, d}$$

پس:

$$\boldsymbol{x'_k = y_k \; for \; k \; = \; 1, 2, \dots, d}$$

و این یعنی x و y با هم برابرند.

طرف دوم:

$$x' = y \rightarrow d(x', y) = 0$$

اینکه $x' \; = \; y$، یعنی $\boldsymbol{x'_k = y_k \; for \; k \; = \; 1, 2, \dots, d}$، پس حاصل $(x`_k - \boldsymbol{y_k})^{\boldsymbol{2}}$ به ازای تمام k ها صفر خواهد شد. پس جمع آنها نیز صفر خواهد شد و عبارت $D(x', y) \; = \; 0$.

نشان دادیم که دو طرف رابطه برقرار است، پس این ویژگی نیز صادق است.