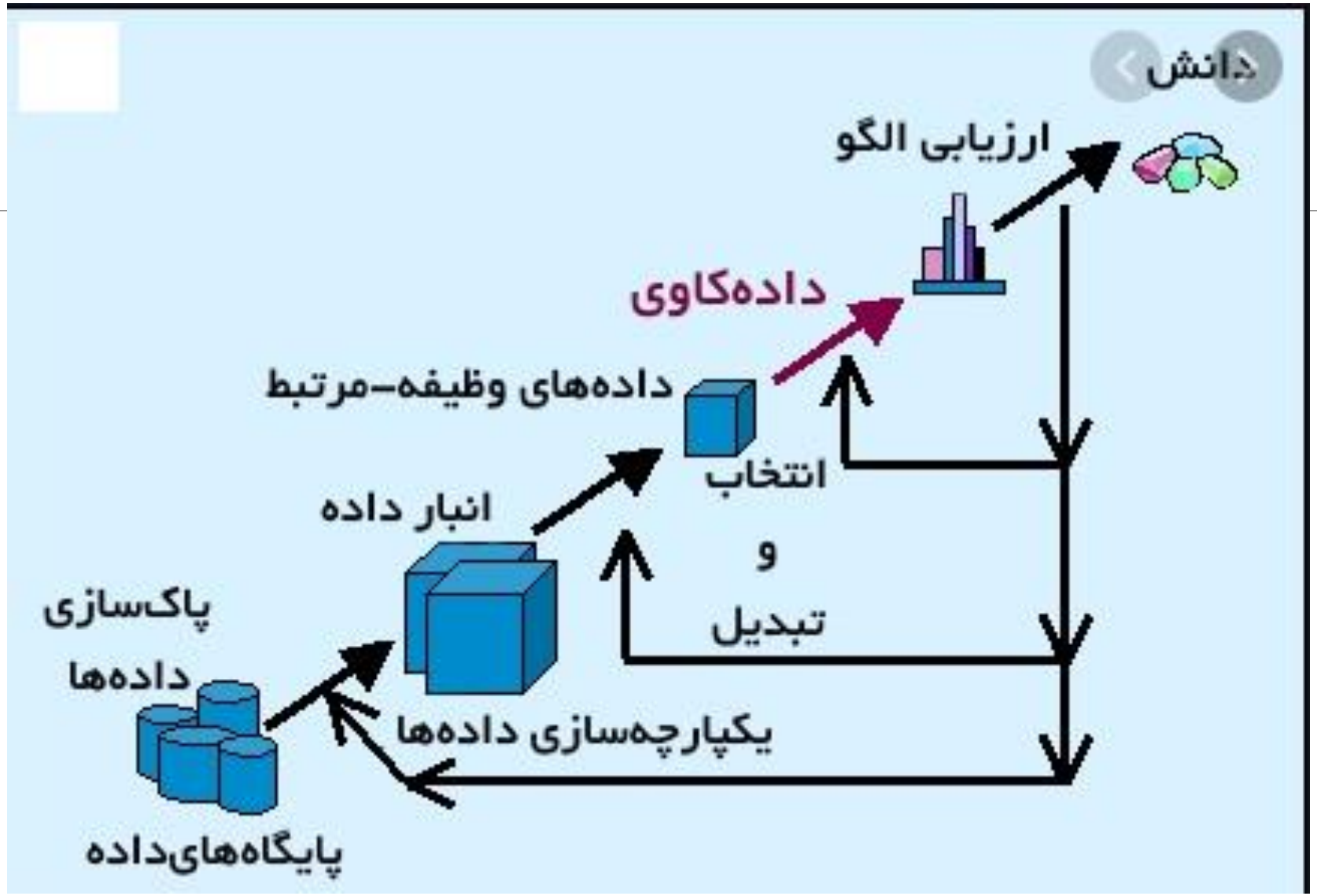
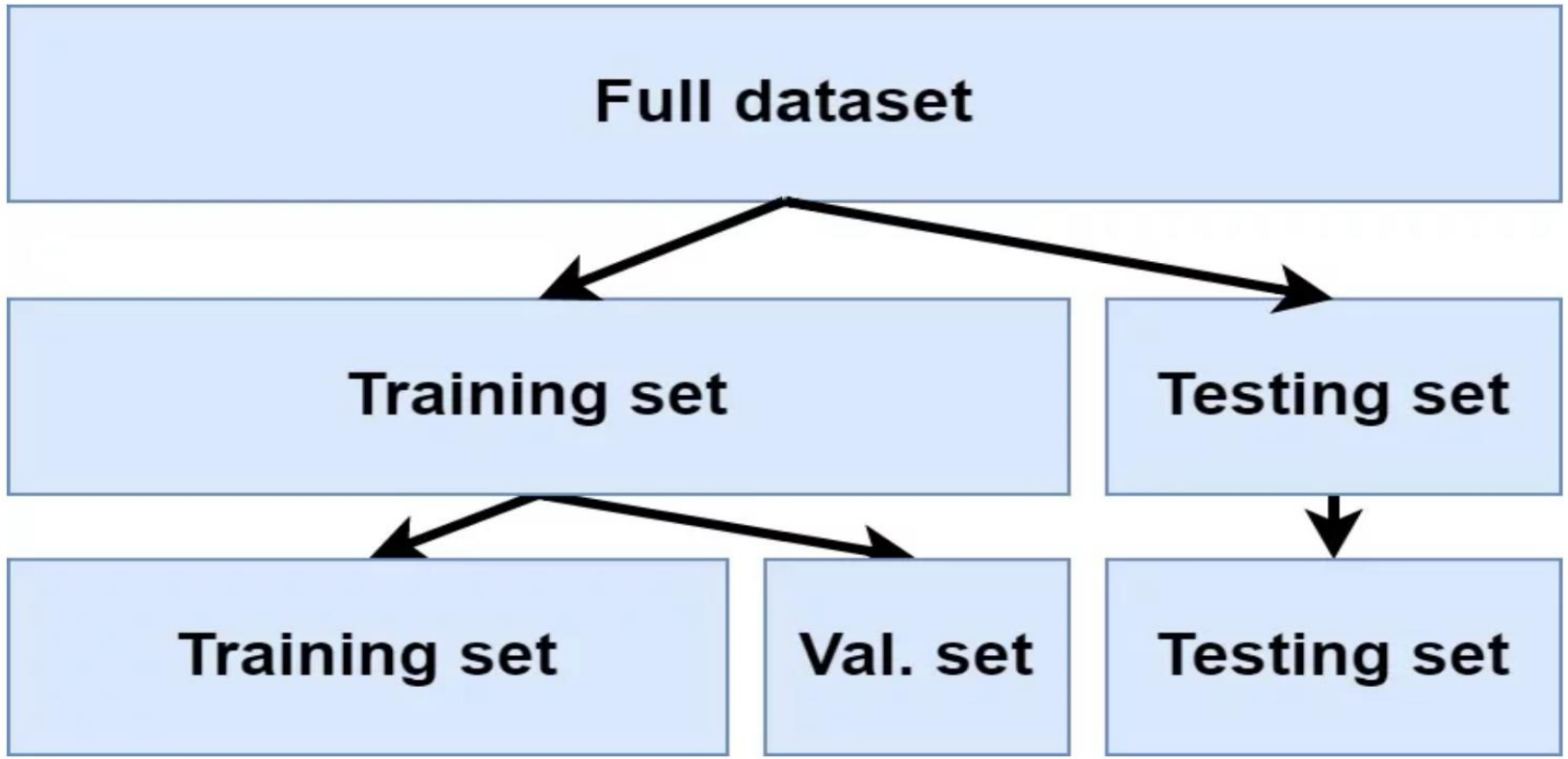
فصل سوم

معیارها و روش های

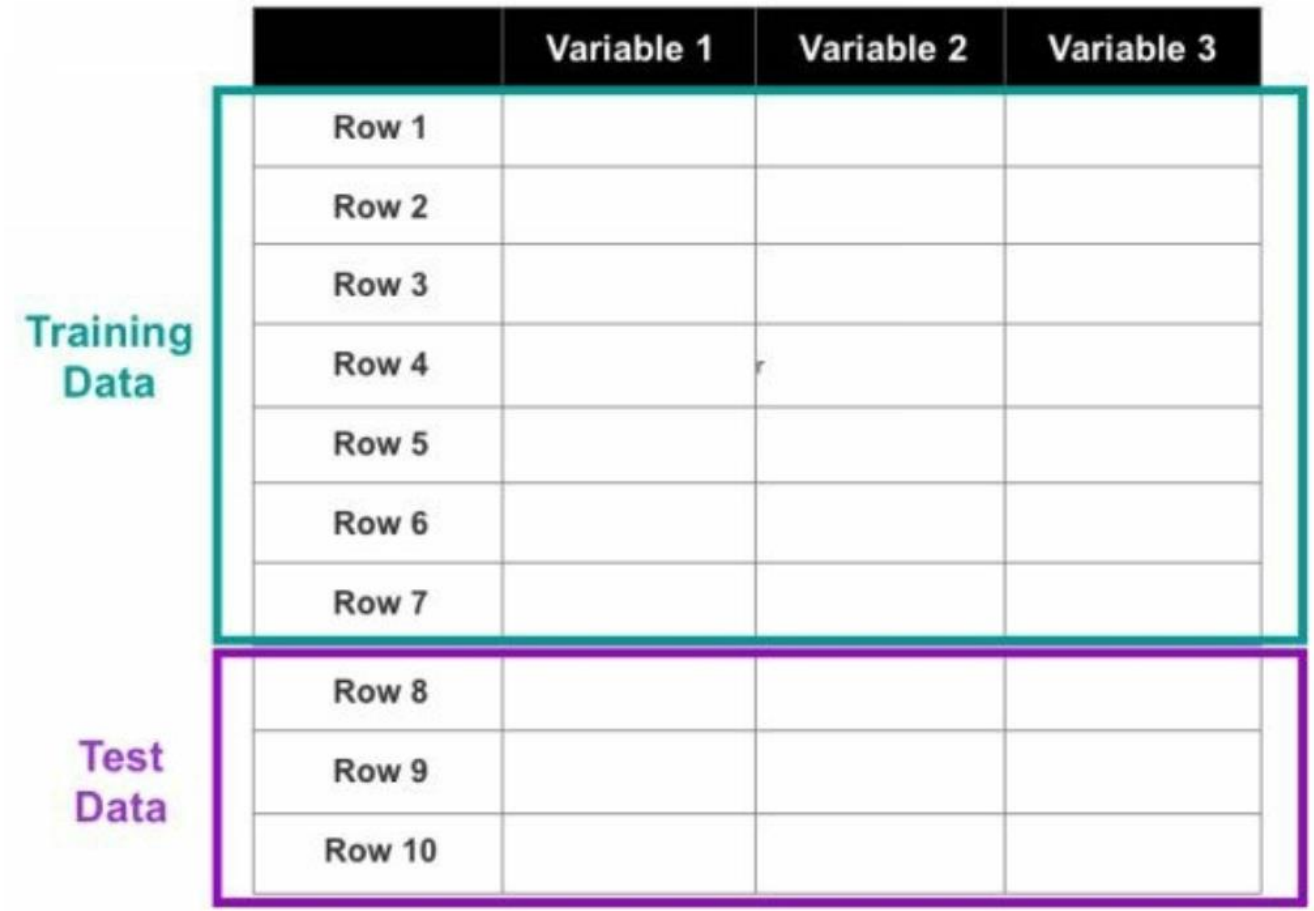
ارزیابی مدل، در داده کاوی


دانش
ارزیابی الگو
داده‌کاوی
داده‌های وظیفه–مرتبط
انتخاب
و
تبدیل
انبار داده
پاک‌سازی
داده‌ها
یکپارچه‌سازی داده‌ها
پایگاه‌های‌داده

داده های آموزش(Train)، اعتبارسنجی(Validation) و آزمایش(Test) در یادگیری ماشین به زبان ساده



برای یادگیری مدل در یادگیری ماشین، معمولا داده ها رو به سه دسته آموزش، اعتبارسنجی و آزمایش تقسیم می کنن. دلیل این مسئله اینه که ما با استفاده از داده های محدودی که داریم، مدلی رو آموزش بدیم که برای ورود به دنیای واقعی هم به کار بیاد. وگرنه آموزش مدلی با دقت صد در صد (بدون در نظر گرفتن داده اعتبارسنجی و آزمایشی و فقط با داده آموزش) کاری نداره و میشه راحت بهش رسید!

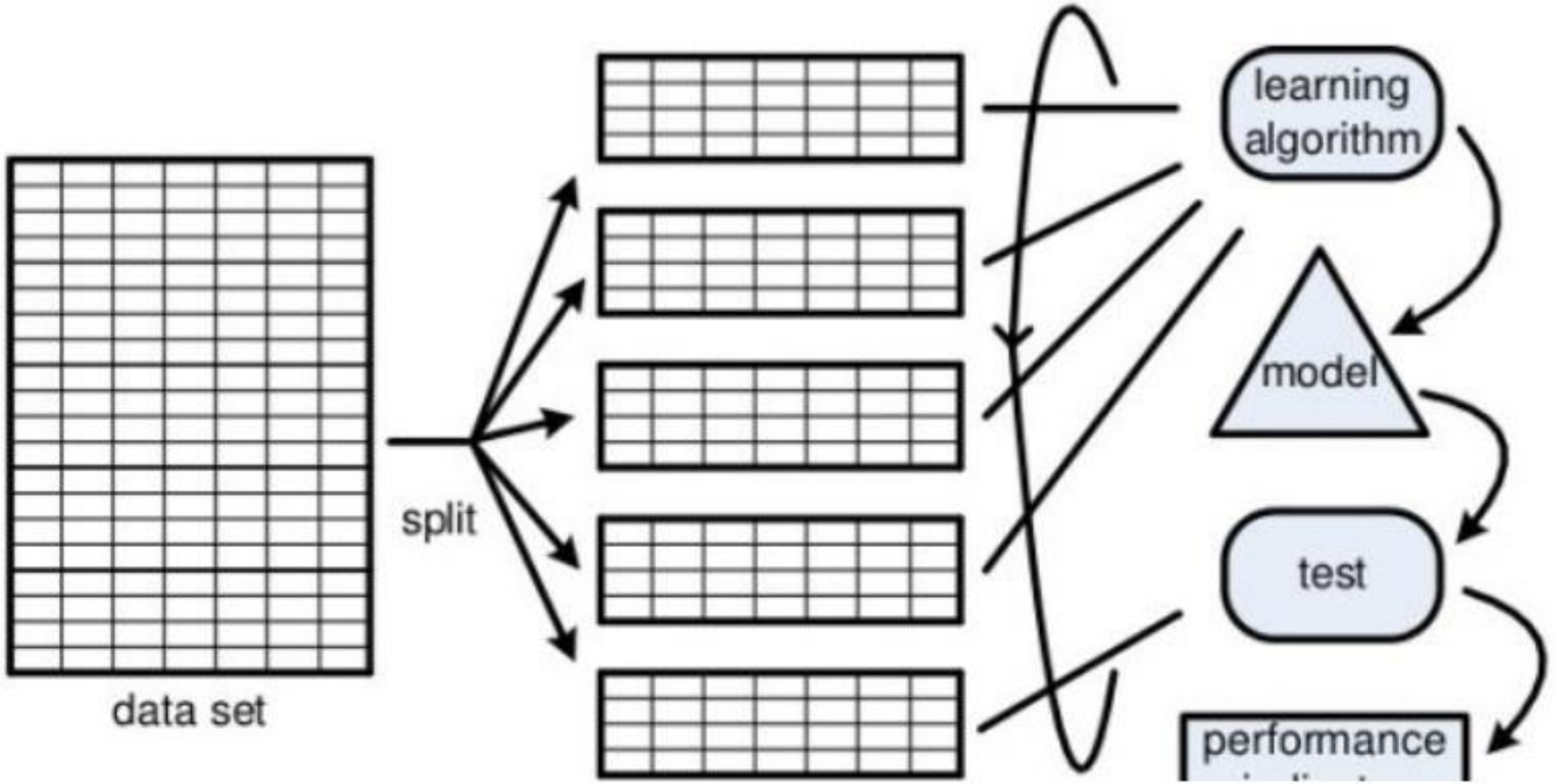

Variable 1
Variable 2
Variable 3
Row 1
Row 2
Row 3
Row 4
Row 5
Row 6
Row 7
Row 8
Row 9
Row 10
Training Data
Test Data

تقسیم داده به آموزش و تست در یادگیری ماشین

یکی از کارهای پایه در یادگیری ماشین تقسیم داده به دو قسمت آموزش و تست می باشد.

با داده های آموزش ، مدل را آموزش می دهیم و با داده های تست ، مدل آموزش یافته را تست می کنیم.

دانشی که در مرحله یادگیری مدل تولید می‌شود، می‌بایست در مرحله ارزیابی مورد تحلیل قرار گیرد تا بتوان ارزش آن را تعیین نمود و در پی آن کارائی الگوریتم یادگیرنده مدل را نیز مشخص کرد. این معیارها را می‌توان هم برای مجموعه داده‌های آموزشی در مرحله یادگیری و هم برای مجموعه رکوردهای آزمایشی در مرحله ارزیابی محاسبه نمود. همچنین لازمه موفقیت در بهره مندی از علم داده کاوی تفسیر دانش تولید و ارزیابی شده است.


learning algorithm
model
test
performance
split
data set

در هنگامِ آموزش، داده‌ها را به دو دسته‌ی آموزشی و ازریابی تقسیم‌بندی می‌کنیم. حال الگوریتم از روی داده‌های آموزشی یادگیری را انجام می‌دهد و از روی داده‌های ارزیابی یا همان داده‌های تست می‌توانید بفهمید که الگوریتم و مدلِ ساخته شده توسطِ آن، چقدر دقت داشته است. وقتی الگوریتمْ عملیاتِ یادگیری را انجام داد و در واقع یک مدل را از روی این داده‌ها ساخت، حالا می‌توان از روی این مدل، عملیاتِ داده‌کاوی را بر روی داده‌های جدید انجام داد.

یک م مجموعه داده آموزشی، مجموعه ای از نمونه‌ها است که در طول فرایند یادگیری استفاده می‌شود و برای ساخت مدل‌های پیشگو استفاده می‌شود. به عنوان مثال برای یک الگوریتم طبقه‌بندی کننده استفاده می‌شود

بیشتر رویکردهایی که از طریق داده‌های آموزشی برای روابط علمی جستجو می‌شوند، منجر به بیش برازش می‌شوند، به این معنی که آنها می‌توانند روابط آشکار را در داده‌های آموزشی که به‌طور کلی نگهداری نمی‌شوند، شناسایی و بهره‌برداری کنند.


Testing and Validation in
Data Mining
Training
Validation
Testing

داده های آموزشی(Train):

این داده ها برای آموزش مدل استفاده می شن. معمولا 70 درصد کل داده های در دسترس، به این دسته تعلق داره. وظیفه اصلی این داده ها تنظیم دقیق وزن هاست.

به عنوان مثال تو یه شبکه عصبی عمیق که از تعدادی لایه تشکیل شده، داده ها از لایه اول به لایه آخر میرن و با توجه به خروجی، طی back-propagation وزن ها آپدیت میشه. اما داده های اعتبارسنجی و آزمایشی به صورت مستقیم و به این شکل روی وزن ها تاثیری ندارن. به همین دلیله که معمولا حجم زیادی از داده ها به این دسته تعلق میگیره تا آپدیت شدن وزن ها با احتمال بیشتری متناسب با توزیع داده های واقعی باشه.

داده های اعتبارسنجی(Validation):

معمولا داده ها رو طی چندین بار به خورد مدل میدن و متناسب با اون وزن ها رو آپدیت میکنن. به هر بار دادن داده های آموزش به مدل و آپدیت شدن وزن ها توسط اون ها، میگن یک Epoch. این موضوع باعث میشه تا موضوع کم بودن داده ها رو با نشون دادن چندباره داده ها جبران کنیم. اما تا یه جایی این تکرار شدنه خوبه. از یه جایی به بعد به جای اینکه مدل از داده ها یاد بگیره، اون ها رو حفظ میکنه! اینجاست که داده های اعتبارسنجی به کمکمون میان.

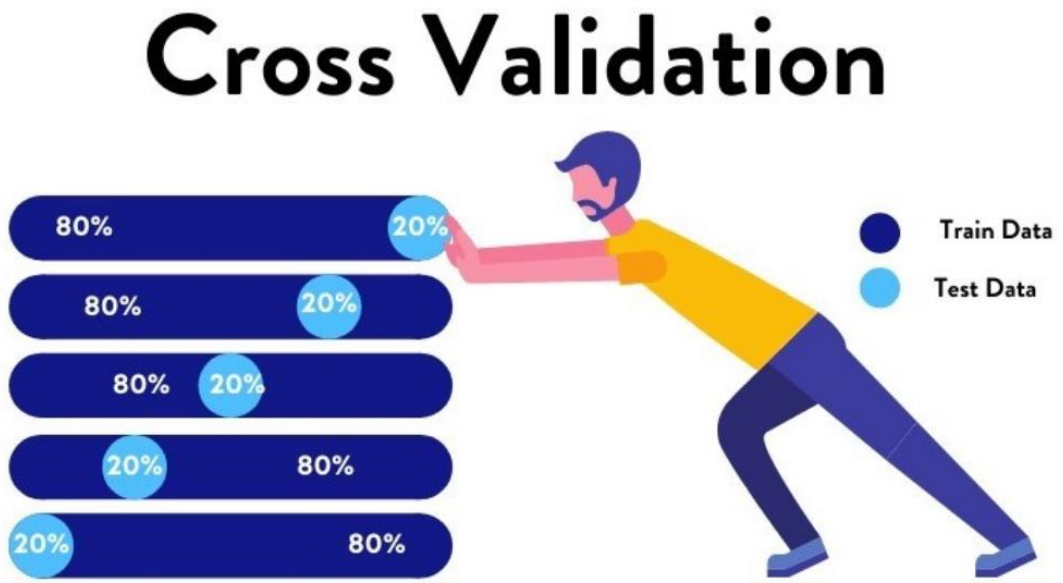
تو هر ایپوک(Epoch)، ما مدل رو روی داده های اعتبارسنجی امتحان میکنیم و دقت مدل رو درمیاریم. تا یه جایی با کم شدن خطای مدل روی داده های آموزش، خطا روی داده های اعتبارسنجی هم پایین میاد. اما از یه جایی به بعد که مدل داده های آموزش رو حفظ میکنه، که اصطلاحا بهش میگن Overfit شده، خطای مدل روی داده های اعتبارسنجی به جای کم شدن، بیشتر میشه. تو این نقطه است که آموزش باید متوقف بشه چرا که آموزش بیشتر نه تنها سودی نداره، بلکه باعث میشه عملکرد نهایی مدل پایین تر بیاد.

داده های آزمایشی(Test):

با توجه به توضیحات گفته شده، در نهایت مدل آموزش میبینه و میشه ازش استفاده کرد(که التبه فعلا به کلیات داره پرداخته میشه و اینکه دقت چقد شده، مدلی که استفاده کردیم چی بوده و ... کاری نداریم).

اما نکته ای که اینجا پیش میاد اینه که من از کجا بفهمم مدلی که شما روی این داده ها آموزش دادی بهتره یا مدل من!؟ اینجا همونجاییه که داده های آزمایشی پا به میدان قضاوت(!) میذارن. از داده های آزمایش به عنوان معیاری برای مقایسه عملکرد مدل های مختلف که روی یه دیتاست واحد آموزش دیدن، استفاده میشه.

اعتبارسنجی متقابل (Cross Validation) چیست؟



با اینکه تقسیم داده های آموزش و تست در توسعه مدل از داده‌های موجود می‌تواند مؤثر باشد، این سؤال باقی می‌ماند که آیا مدل بر روی داده‌های جدید خوب کار می‌کند. اگر مجموعه‌داده موجود برای ساخت یک مدل خیلی کوچک باشد، یا اگر تقسیم داده‌ها به دو بخش داده های آموزش و تست از نقطه‌ی درست انجام نشده باشد، در نتیجه عملکرد مدل در دنیای واقعی ضعیف می‌شود.

خوشبختانه، یک راه‌حل مؤثر برای این مشکل وجود دارد. به‌جای تقسیم داده‌ها به دو قسمت (یکی برای آموزش و یکی برای آزمایش)، از روش **اعتبارسنجی متقابل (Cross Validation)** استفاده کنیم. این روش دسترسی داده‌های آموزشی را با تقسیم داده‌ها به چندین ترکیب و آزمایش هر ترکیب خاص، بیشینه می‌کند.

اعتبارسنجی متقابل به دو روش اصلی قابل انجام است. روش اول، اعتبارسنجی متقابل کامل (Exhaustive Cross Validation) است که شامل یافتن و آزمایش تمام ترکیب‌های ممکن برای تقسیم نمونه اصلی به مجموعه آموزشی و مجموعه آزمایشی است. روش دیگر و معمول‌تر اعتبارسنجی متقابل ناکامل (Non-exhaustive Cross Validation) است که به‌عنوان اعتبارسنجی k-fold شناخته می‌شود. این تکنیک شامل تقسیم داده به k سطل معین و نگه‌داری یکی از سطل‌ها، برای آزمایش مدل، در هر مرحله است.

تنظیم پارامترها به کمک اعتبارسنجی متقابل

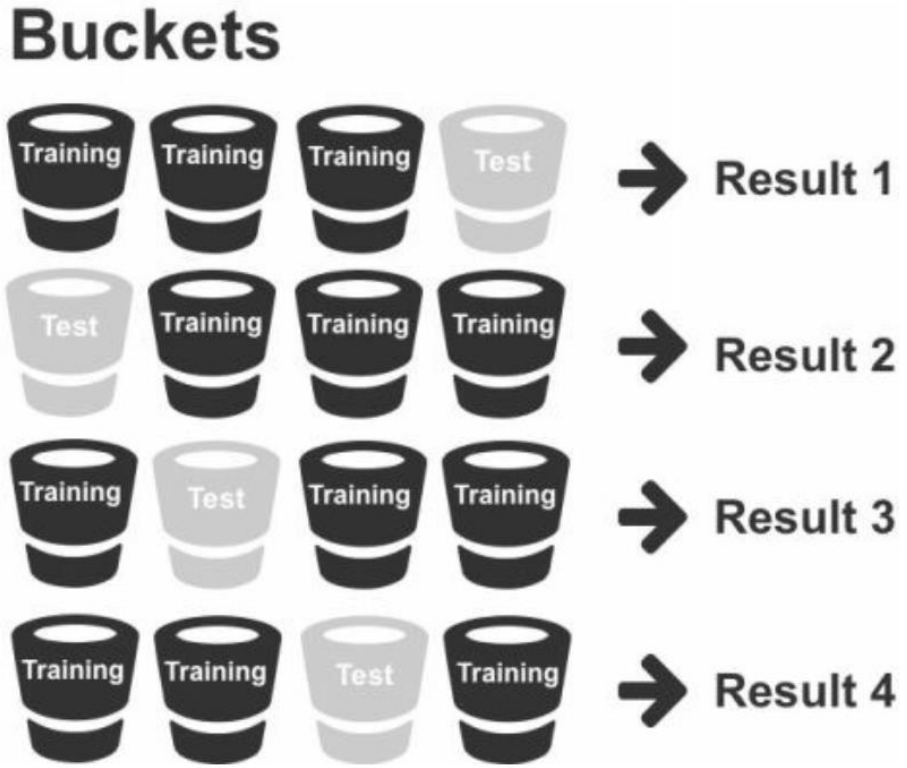
فرض کنید مشاهداتی از جامعه به صورت یک نمونه تصادفی در دسترس است که قرار است از آن‌ها در مدل‌سازی استفاده شود. هدف در اعتبارسنجی متقابل، دستیابی به مدلی است که تعداد پارامترهای آن بهینه باشد. یعنی پیدا کردن مدلی است که دچار بیش‌برازش نباشد. برای برای دستیابی به این هدف در «آموزش ماشین» (Machine Learning) معمولا داده‌ها را به دو قسمت تفکیک می‌کنند.

- **قسمت داده‌های آموزشی (Training set)**: از این بخش از داده‌ها به منظور ایجاد مدل و برآورد پارامترهای آن استفاده می‌شود.
- **قسمت داده‌های آزمایشی (Test set)**: این قسمت از داده‌ها برای بررسی کارایی مدل استفاده می‌شود. اهمیت این بخش از داده‌ها در این نکته است که این مشاهدات شامل مقدارهای متغیرهای مستقل (xها) و پاسخی (y) هستند که در مدل به کار نرفته ولی امکان مقایسه مقدار پیش‌بینی شده ($\hat{y}$) را با مقدار واقعی به ما می‌دهند. البته توجه داریم که این داده‌ها مدل را تحت تاثیر قرار نداده‌اند، پس در تعیین پارامترهای مدل نقشی نداشته و فقط برای ارزیابی مدل به کار می‌روند.

در هر مرحله از فرایند CV، مدل بدست آمده توسط داده‌های آزمایشی برای پیش‌بینی داده‌های CV به کار گرفته و «خطا» (Error) یا «دقت» (Accuracy) حاصل از برازش مدل روی داده‌های CV محاسبه می‌شود. معمولا میانگین این خطاها (دقت‌ها) به عنوان خطای (دقت) کلی مدل در نظر گرفته می‌شود. البته بهتر است انحراف معیار خطاها (دقت‌ها) نیز گزارش شود. به این ترتیب با توجه به تعداد پارامترهای مختلف (پیچیدگی مدل)، می‌توان مدل‌های متفاوتی تولید و خطای برآورد آن‌ها را به کمک روش CV اندازه‌گیری کرد. در انتها مدلی را به عنوان مدل مناسب انتخاب خواهیم کرد که دارای کمترین برآورد خطا باشد.

روش k-fold در یادگیری ماشین

در تقسیم داده های آموزش و تست، برای انجام اعتبارسنجی k-fold، داده‌ها در ابتدا به طور تصادفی به k سطل با سایز یکسان تقسیم می‌شوند. پس از آن‌یکی از سطل‌ها برای آزمایش رزرو می‌شود و برای اندازه‌گیری و ارزیابی عملکرد k-۱ سطل باقی‌مانده استفاده می‌شود. شکل زیر اعتبارسنجی k-fold را نشان می‌دهد.

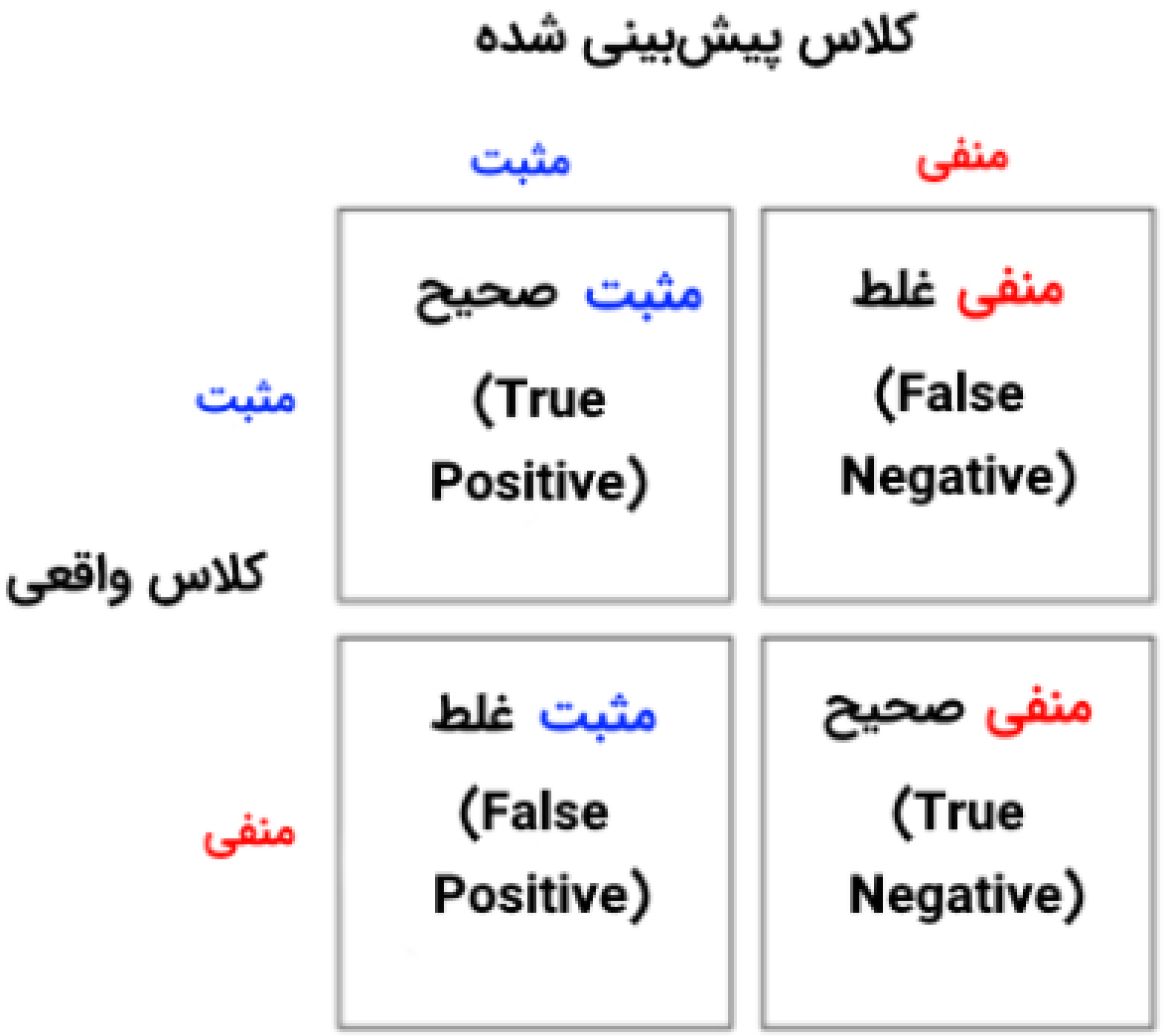


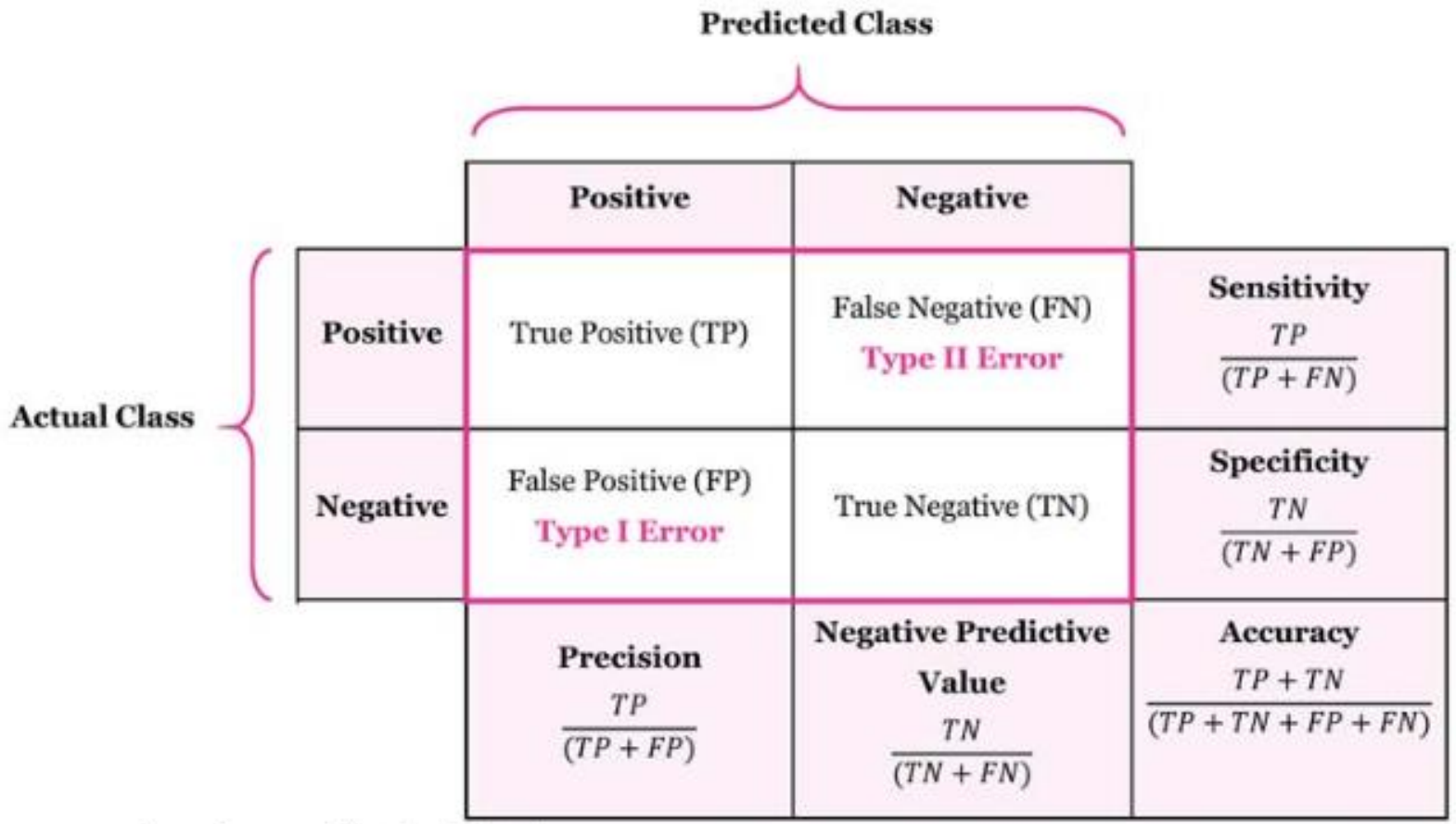
فرایند اعتبارسنجی متقابل به‌اندازه k بار (fold) تکرار خواهد شد. در هر fold، یک سطل دست‌نخورده باقی می‌ماند تا مدلی که به کمک دیگر سطل‌ها آموزش‌دیده را، آزمایش کند. این فرایند تا جایی تکرار خواهد شد که تمام سطل‌ها به‌عنوان سطل آموزش و آزمایش مورداستفاده قرار بگیرند. پس از آن نتایج ترکیب می‌شوند تا یک مدل واحد را فرموله‌سازی کنند.

با استفاده از تمام داده‌ها در هر دو بخش آموزش و تست، تکنیک اعتبارسنجی k-fold خطاهای بالقوه (مثل برازش بیش از حد، Overfitting) که با اتکا بر روی یک تقسیم ثابت از داده آموزشی و آزمایشی اتفاق می‌افتد را به‌شدت کم می‌کند.

ارزیابی در الگوریتم های دسته بندی

برای سادگی معیارهای ارزیابی الگوریتم‌های دسته بندی، آنها را برای یک مسئله با دو دسته ارائه خواهیم نمود. در ابتدا با مفهوم ماتریس درهم ریختگی Classification Matrix آشنا می‌شویم. این ماتریس چگونگی عملکرد الگوریتم دسته بندی را با توجه به مجموعه داده ورودی به تفکیک انواع دسته‌های مساله دسته بندی، نمایش می‌دهد.


کلاس پیش‌بینی شده
مثبت
منفی
مثبت صحیح
(True Positive)
منفی غلط
(False Negative)
مثبت
کلاس واقعی
مثبت غلط
(False Positive)
منفی صحیح
(True Negative)
منفی



معیارهای ارزیابی الگوریتم های هوش مصنوعی

عناصر ماتریس درهم ریختگی

هر یک از عناصر ماتریس به شرح ذیل می‌باشد:

TN:بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی منفی تشخیص داده است.
TP:بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی نیز دسته آنها را بدرستی مثبت تشخیص داده است.
FP:بیانگر تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته بندی دسته آنها را به اشتباه مثبت تشخیص داده است.
FN:بیانگر تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته بندی دسته آنها را به اشتباه منفی تشخیص داده است.

ماتریس درهم ریختگی Confusion Matrix

این ماتریس برای متغیر برچسب دار که دارای دوسطح پاسخ میباشد به صورت زیر ساخته میشود:

رکورد در مدل تخمین زده رکورد واقعی	-	+
+ موفقیت	FN	TP
- شکست	TN	FP

TN: تشخیص درست شکست (-)

TP: تشخیص درست موفقیت (+)

FN: تشخیص نادرست شکست (-)

FP: تشخیص نادرست موفقیت (+)

با کمک نتایج ماتریس درهم ریختگی میتوان ضرایب یا معیارهای زیر را محاسبه نمود که برای ارزیابی دقت مدل مفید میباشید.

معایب دقت دسته بندی Classification Accuracy Rate

$$CA = \frac{TN+TP}{TN+TP+FN+FP}$$

نسبت فوق، نسبت رکوردهایی است که درست دسته بندی شده اند، هرچه این نسبت به یک نزدیک تر باشد دقت مدل بیشتر است.

به طور مثال فرض کنید داده ها توسط دو مدل درخت تصمیم و رگرسیون لوژستیک دسته بندی شده باشند، برای مقایسه دو مدل میتوان از معیار دقت استفاده کرد، هرروشی که معیار دقت مدل ان بیشتر باشد از دقت بالاتری برای پیش بینی برخوردار است.

مهمترین معیار برای تعین کارایی یک الگوریتم دسته بندی دقت یا نرخ دسته بندی Classification Accuracy - Rateاست که این معیار دقت کل یک دسته بندی را محاسبه می‌کند. در واقع این معیار مشهورترین و عمومی‌ترین معیار محاسبه کارایی الگوریتم‌های دسته بندی است که نشان می‌دهد، دسته بند طراحی شده چند درصد از کل مجموعه رکوردهای آزمایشی را به درستی دسته بندی کرده است.

دقت دسته بندی با استفاده از رابطه CA بدست می‌آید که بیان می‌کند دو مقدار TPو TNمهمترین مقادیری هستند که در یک مسئله دودسته ای باید بیشینه شوند. (در مسائل چند دسته ای مقادیر قرار گرفته روی قطر اصلی این ماتریس - که در صورت کسر محاسبه CA قرار می‌گیرند - باید بیشینه باشند.

معیار خطای دسته بندی Error Rate دقیقاً برعکس معیار دقت دسته بندی است که با استفاده از رابطه ER بدست می‌آید. کمترین مقدار آن برابر صفر است زمانی که بهترین کارایی را داریم و بطور مشابه بیشترین مقدار آن برابر یک است زمانی که کمترین کارائی را داریم.

معیار نرخ خطا:Error Rate

$$ER = \frac{(FN+FP)}{TN+TP+FN+FP}$$

به طوریکه $ER = 1 - CA$ هرچه این معیار به صفر نزدیکتر باشد مدل از دقت بالاتری برخوردار است.

معیار یادآوری (حساسیت): Recall Rate

$$Recall^{-} = \frac{TN}{FP+TN}$$

نسبت رکوردهای درست پیش بینی شده منفی شده به کل رکوردهای منفی

$$Recall^{+} = \frac{TP}{FN+TP}$$

نسبت رکوردهایی که درست پیش بینی + شده به کل رکوردهای +

هر چه $Recall^{-}$ و $Recall^{+}$ به یک نزدیکتر باشند دقت مدل در پیش بینی بیشتر است.

معیار دقت: Precision Rate

$$Precision^{-} = \frac{TN}{TN+FN}$$

نسبت داده هایی که پیش بینی درست (-) شده است به کل داده هایی که پیش بینی (-) برای ان ها صورت گرفته است.

$$Precision^{+} = \frac{TP}{TP+FP}$$

نسبت داده هایی که پیش بینی درست (+) شده است به کل داده هایی که پیش بینی (+) برای ان ها صورت گرفته است.

نکته

معیار دقت نشان میدهد چند درصد از رکوردهایی که منفی پیش بینی شده اند (یا +) درست هستند؛ همچنین، ریکال منفی ($Recall^-$) نشان میدهد چه نسبتی از رکوردهایی که منفی درست پیش بینی شده اند به کل رکوردهای منفی میباشد.

معیار F: (میانگین هارمونیک)

$$F^{+} = \frac{2Recall^{+}Precision^{+}}{Recall^{+}+Precision^{+}}$$

$$F^{-} = \frac{2Recall^{-}Precision^{-}}{Recall^{-}+Precision^{-}}$$

ترکیت دومعیار دقت و حساسیت به تفکیک + و – معیار جدیدی که میانگین هارمونیک یا معیار F است را ایجاد میکند از این معیار در زمانی استفاده میشود که تفاوتی در پیش بینی معیار دقت و حساسیت برای محقق وجود نداشته باشد.

ذکر این نکته ضروری است که در مسائل واقعی، معیار دقت دسته بندی به هیچ عنوان معیار مناسبی برای ارزیابی کارایی الگوریتم‌های دسته بندی نمی‌باشد، به این دلیل که در رابطه دقت دسته بندی، ارزش رکوردهای دسته‌های مختلف یکسان در نظر گرفته می‌شوند. بنابراین در مسائلی که با دسته‌های نامتعادل سروکار داریم، به بیان دیگر در مسائلی که ارزش دسته ای در مقایسه با دسته دیگر متفاوت است، از معیارهای دیگری استفاده می‌شود.

همچنین در مسائل واقعی معیارهای دیگری نظیر DRو FARکه به ترتیب از **روابط III و IV**بدست می‌آیند، اهمیت ویژه ای دارند. این معیارها که توجه بیشتری به دسته بند مثبت نشان می‌دهند، توانایی دسته بند را در تشخیص دسته مثبت و بطور مشابه تاوان این توانایی تشخیص را تبیین می‌کنند. معیار DRنشان می‌دهد که دقت تشخیص دسته مثبت چه مقدار است و معیار FARنرخ هشدار غلط را با توجه به دسته منفی بیان می‌کند.

تمرین

ماتریس درهم ریختگی زیر برای تشخیص نوعی سرطان در بین 1500 بیمار تهیه شده است. مطلوبست محاسبه و تفسیر انواع معیارهای ارزیابی با کمک این ماتریس.

پیش بینی واقعی	سرطان ندارد (-)	سرطان دارد (+)
سرطان ندارد(-)	100 TN	50 FP
سرطان دارد(+)	10 FN	500 TP

ماتریس هزینه: Cost Matrix

در مواردی که ارزش دسته ها با یکدیگر برابر نباشند (تشخیص نادرست مثبت و تشخیص نادرست منفی) استفاده از معیارهایی که به هزینه این تشخیص نادرست توجه نمیکنند روش درستی نیست اگر بتوان میزان اهمیت هر دسته را تعیین کنیم و در نتیجه هزینه های مربوط به خطاها در فرآیند دسته بندی مشخص نمائیم، بهترین معیار ارزیابی برای داده های دسته بندی شده معیار هزینه یا Cost میباشد. دراین صورت ماتریس درهم ریختگی با توجه به هزینه به صورت زیر میباشد.

	-	+
+	C +|-	C +|+
-	C -|-	C -|+

$$Cost = TNC_{-|-} + FPC_{-|+} + FNC_{+|-} + TPC_{+|+}$$

$$Cost = FPC_{-|+} + FNC_{+|-}$$

نکته

در مقایسه روشهای مدل سازی در صورت حضور عامل هزینه در دسته بندی داده ها برای مقایسه مدل ها جهت پیش بینی حتما باید به ماتریس هزینه و محاسبه Cost در این ماتریس توجه کرد.

Area Under Curve (AUC)

معیار مهم دیگری که برای تعیین میزان کارایی یک دسته بند استفاده می‌شود معیار **AUC** (Area Under Curve) است.

AUC نشان دهنده سطح زیر نمودار **ROC** (Receiver Operating Characteristic) می‌باشد که هر چه مقدار این عدد مربوط به یک روش دسته بندی بزرگتر باشد کارایی نهایی دسته بند مطلوب‌تر ارزیابی می‌شود. نمودار ROC روشی برای بررسی کارایی دسته بندها می‌باشد. در واقع منحنی‌های ROC منحنی‌های دو بعدی هستند که در آنها DR یا همان نرخ تشخیص صحیح دسته مثبت (True Positive Rate) - TPR روی محور Y و بطور مشابه FAR یا همان نرخ تشخیص غلط دسته منفی (False Positive Rate) - FPR روی محور X رسم می‌شوند. به بیان دیگر یک منحنی ROC مصالحه نسبی میان سودها و هزینه‌ها را نشان می‌دهد.

بسیاری از دسته بندها همانند روش‌های مبتنی بر درخت تصمیم ، به گونه ای طراحی شده اند که تنها یک خروجی دودویی (مبنی بر تعلق ورودی به یکی از دو دسته ممکن) تولید می‌کنند. به این نوع دسته بندها که تنها یک خروجی مشخص برای هر ورودی تولید می‌کنند، دسته بندهای گسسته گفته می‌شود که این دسته بندها تنها یک نقطه در فضای ROC تولید می‌کنند.

مقدار AUCبرای یک دسته بند که بطور تصادفی، دسته نمونه مورد بررسی را تعیین می‌کند برابر 0.5 است. همچنین بیشترین مقدار این معیار برابر یک بوده و برای وضعیتی رخ می‌دهد که دسته بند ایده آل بوده و بتواند کلیه نمونه‌های مثبت را بدون هرگونه هشدار غلطی تشخیص دهد. معیار AUCبرخلاف دیگر معیارهای تعیین کارایی دسته بندها مستقل از آستانه تصمیم گیری دسته بند می‌باشد. بنابراین این معیار نشان دهنده میزان قابل اعتماد بودن خروجی یک دسته بند مشخص به ازای مجموعه داده‌های متفاوت است که این مفهوم توسط سایر معیارهای ارزیابی کارایی دسته بندها قابل محاسبه نمی‌باشد. در برخی از مواقع سطح زیر منحنی‌های ROCمربوط به دو دسته بند با یکدیگر برابر است ولی ارزش آنها برای کاربردهای مختلف یکسان نیست که باید در نظر داشت در این گونه مسائل که ارزش دسته‌ها با یکدیگر برابر نیست، استفاده از معیار AUCمطلوب نمی‌باشد. به همین دلیل در این گونه مسائل استفاده از معیار دیگری به جزء هزینه Cost Matrixمنطقی به نظر نمی‌رسد. در انتها باید توجه نمود در کنار معیارهای بررسی شده که همگی به نوعی دقت دسته بند را محاسبه می‌کردند، در دسته بندهای قابل تفسیر نظیر دسته بندهای مبتنی بر قانون و یا درخت تصمیم، پیچیدگی نهایی و قابل تفسیر بودن مدل یاد گرفته شده نیز از اهمیت بالایی برخوردار است.

معیار AUC: Area Under Curve

این معیار نشان دهنده مساحت زیر منحنی ROC میباشد،

ROC = Receiver Operating Characteristic

منحنی ROC برای یک مدل نشان دهنده موازنه بین نرخ (+) واقعی یا TPR (True Positive Rate) و نرخ (+) اشتباه یا FPR (False Positive Rate) میباشد.

برای محاسبه معیار AUC نیاز به ترسیم منحنی ROC داریم که برای این منظور با کمک اطلاعات ماتریس درهم ریختگی ابتدا TPR و FPR با کمک فرمول های زیر محاسبه میشوند.

$$\text{TPR} = \frac{TP}{FN + TP}$$

$$\text{FPR} = \frac{FP}{TN + FP}$$

$$TPR = \frac{TP}{FN+TP}$$

$$FPR = \frac{FP}{TN+FP}$$

	+	-	
+	TP	FN	TP+FN
-	FP	TN	FP+TN

مثبت واقعی

منفی واقعی

معیار AUC

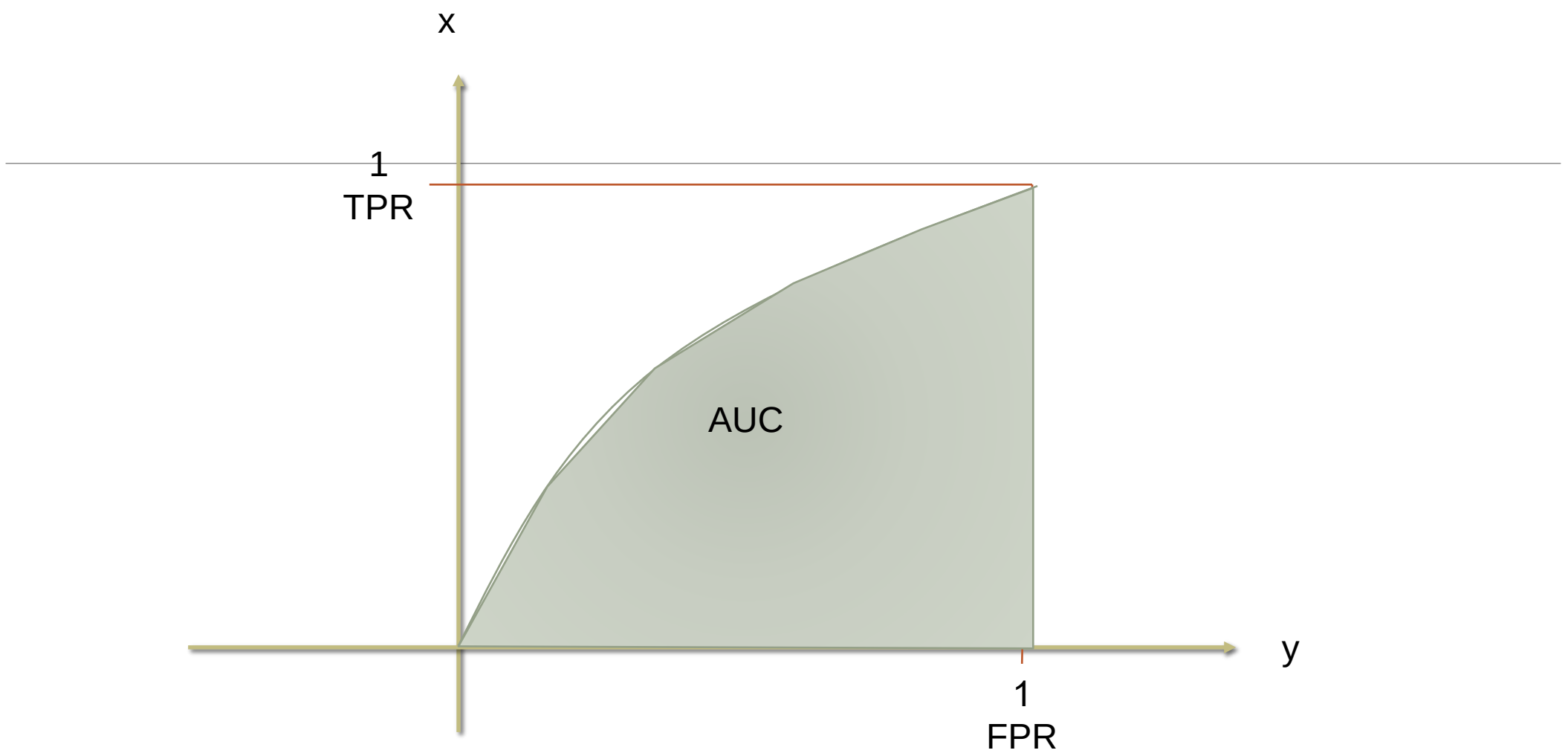
منحنی ROC – > منحنی مشخصه عملکرد سیستم

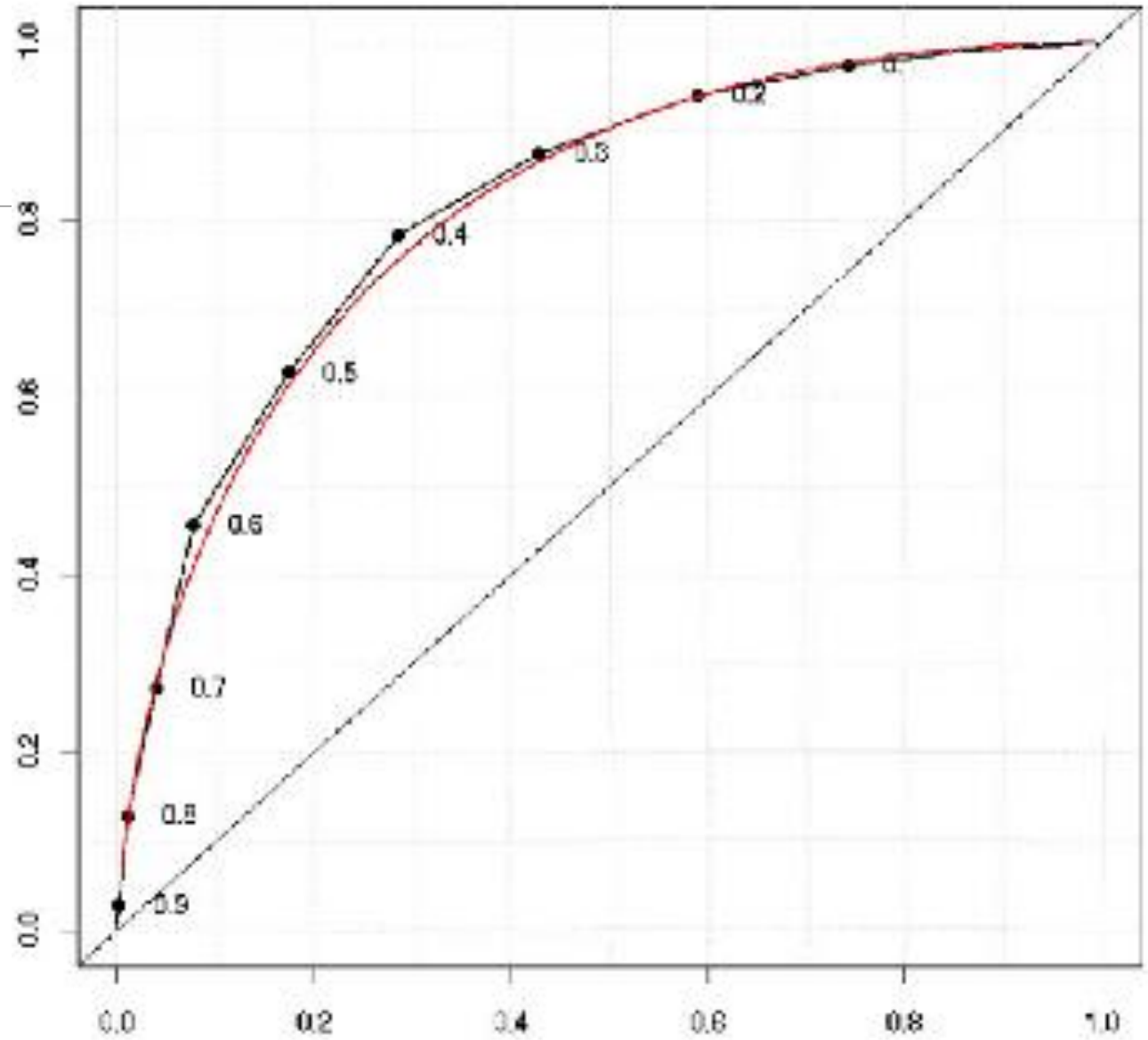
تاریخچه این منحنی به جنک جهانی دوم برمیگردد که در آن توانایی سیستم رادار در تشخیص درست شی مورد نظر استفاده شده است که این شی کشتی دشمن یا یک شی بی ضرر میباشد. امروزه منحنی ROC کاربردهای فراوانی دارد از جمله توسط پزشکان برای ارزیابی آزمایش های پزشکی به کار میرود. این منحنی برای یک مدل برازش داده شده نشان دهنده یک موازنه بین نرخ مثبت واقعی TPR و نرخ مثبت اشتباه FPR میباشد.

منحنی ROC

نقاط در صفحه منحنی ROC داخل مربع یک در یک فوق قرار میگیرند. نقطه (0,1) نقطه ایده آل این منحنی میباشد. نقاط ترسیم شده توسط مدل پیشنهادی (سیستم مورد استفاده) در این صفحه تشکلی یک منحنی محدب رامیدهد که مساحت زیر این منحنی معیار AUC میباشد که هرچه این مساحت بیشتر باشد دقت مدل در پیش بینی بیشتر است.

توجه کنید که یک نقطه در فضای ROC بهتر از نقطه دیگر است اگروفقط اگر در شمال غربی تر از آن نقطه قرار گرفته باشد.


x
1
TPR
AUC
y
1
FPR


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

ترسیم منحنی ROC

ابتدا احتمال تعلق رکوردها به کلاس مثبت واقعی محاسبه میشود سپس، بر اساس این احتمالات مرتب میشوند در ادامه آستانه تصمیم گیری برای دسته بندی را از کمترین مقدار احتمال تا بیشترین مقدار آن تغییر داده و دو شاخص FPR و TPR را برای هروضعیت محاسبه میکنیم. با مشخص نمودن این دو شاخص نقاط در صفحه منحنی ROC مشخص و با احتمال این نقاط میتوان مساحت زیر منحنی ROC را محاسبه و ارزیابی نمود.

مثال

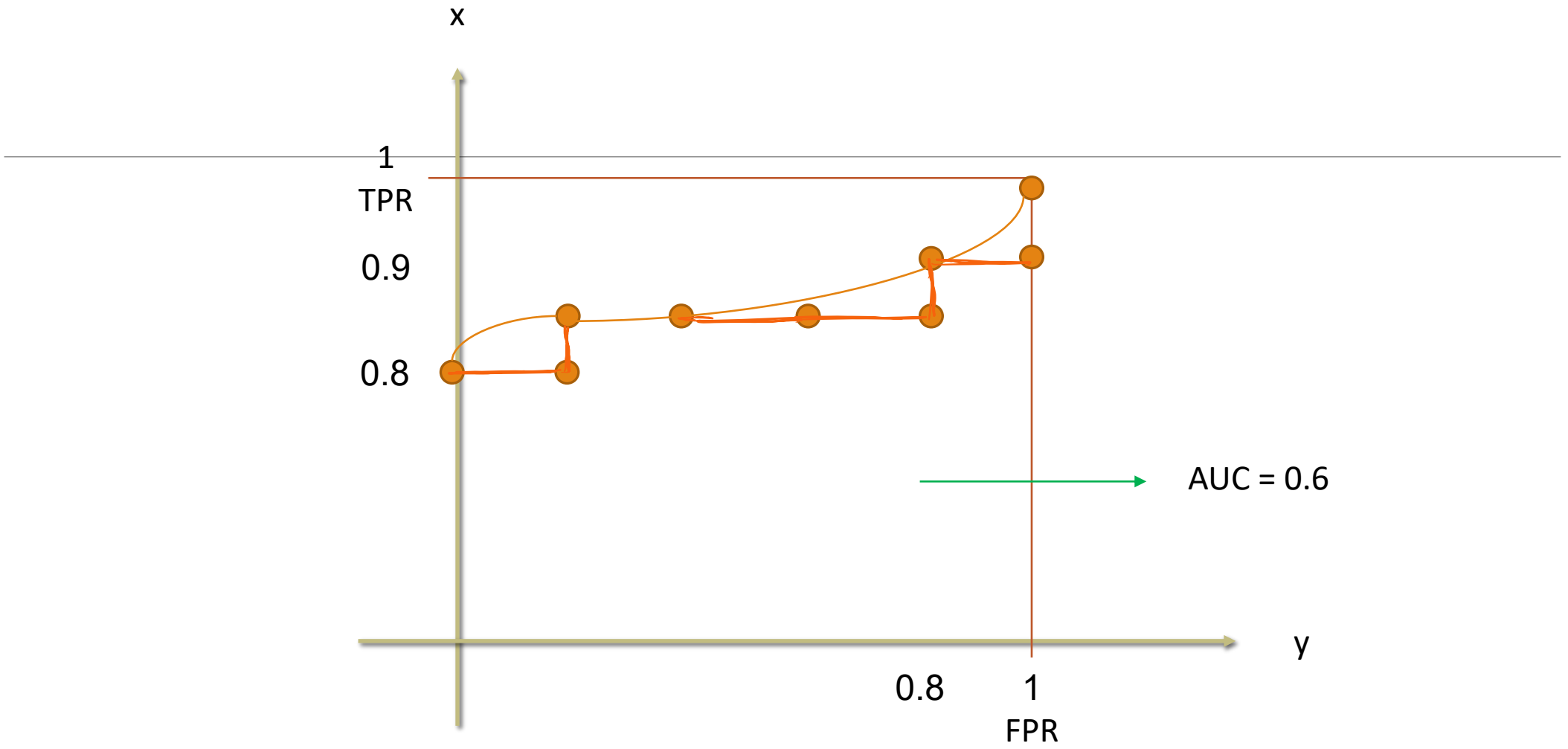
اطلاعات 10 رکورد شامل کلاس واقعی و احتمال تعلق به کلاس مثبت به طور مرتب شده در جدول زیر آمده است. مطلوبست رسم منحنی ROC یا مساحت AUC.

رکوردها	کلاس واقعی	احتمال تعلق به کلاس +
1	+	0.95
2	+	0.93
3	-	0.87
4	-	0.85
5	-	0.85
6	+	0.85
7	-	0.76
8	+	0.53
9	-	0.43
10	+	0.25

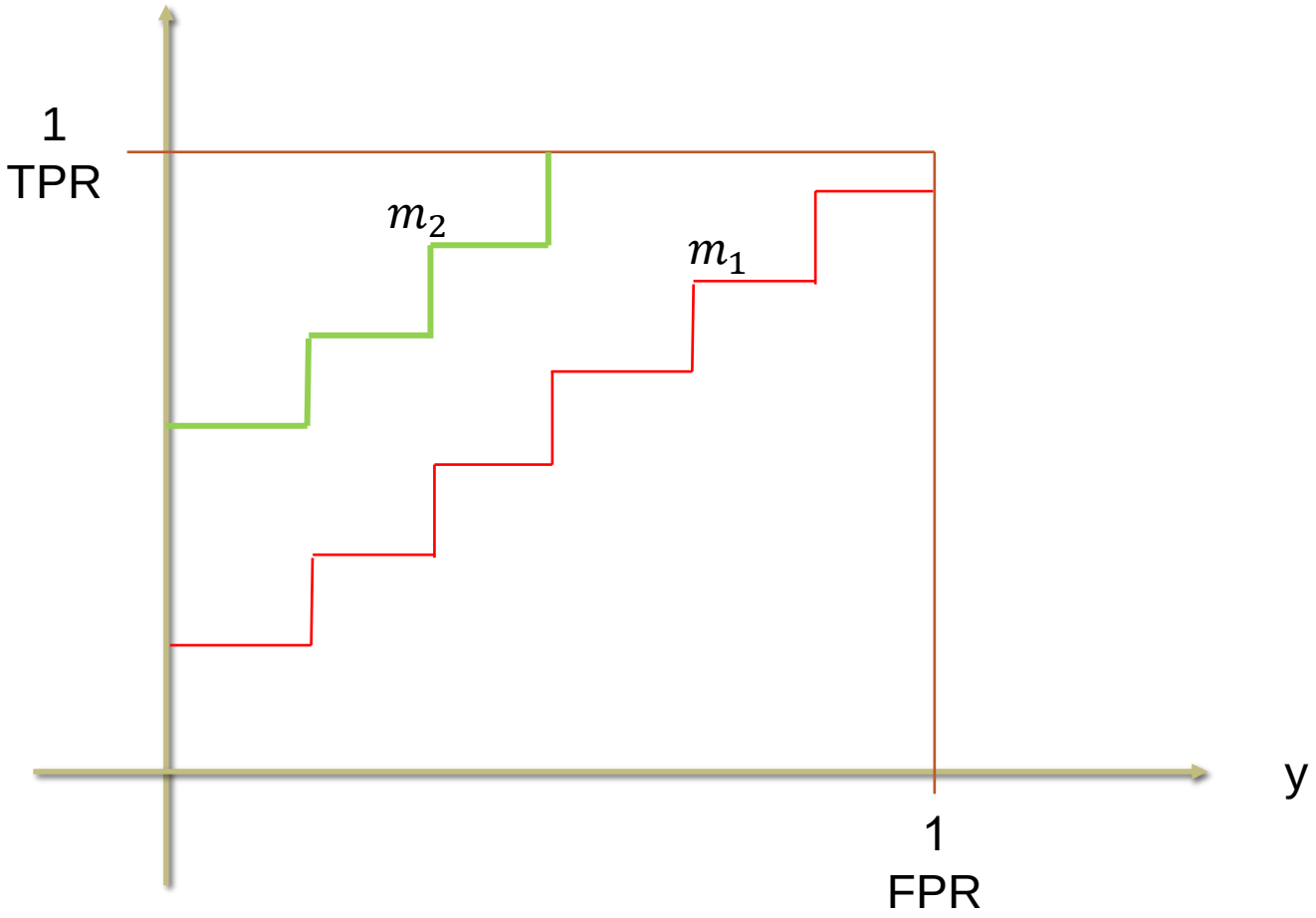
کلاس واقعی	+	-	+	-	+	-	-	-	+	+
حدآستانه تصمیم گیری	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95
TP	5	4	4	3	3	2	2	2	2	1
FP	5	5	4	4	3	3	2	1	0	0
TN	0	0	1	1	2	2	3	4	5	5
FN	0	1	1	2	2	3	3	3	3	4
TPR	1	0.8	0.8	0.6	0.6	0.4	0.4	0.4	0.4	0.2
FPR	1	1	0.8	0.8	0.6	0.6	0.4	0.2	0	0

محاسبه AUC

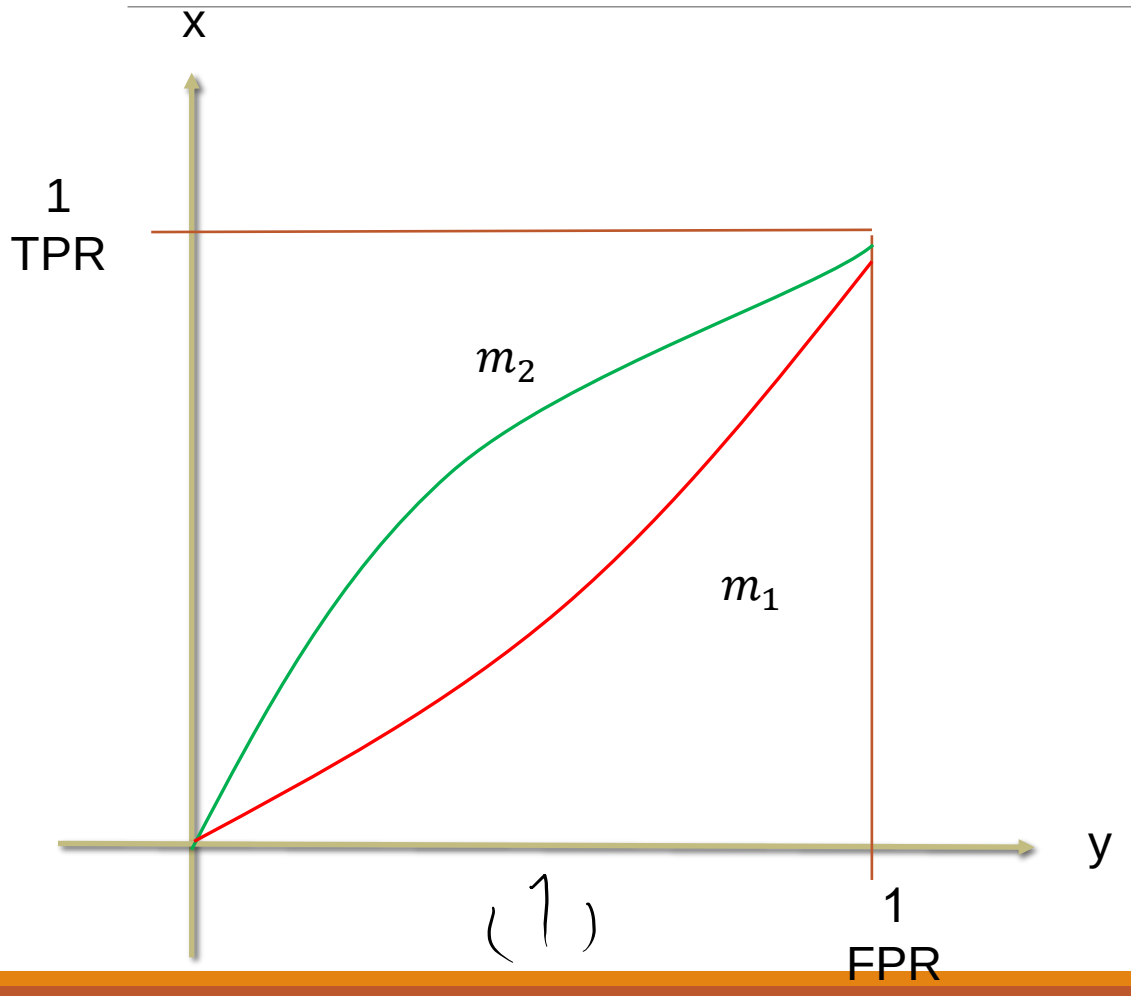
رکورد	احتمال تعلق به مثبت	کلاس واقعی	TP	FP	TN	FN	TPR	FPR
1	0.95	+	۱	۰	۵	۴	۰.۲	۰
2	0.93	+	۲	۰	۵	۳	۰.۴	۰
3	0.87	−	۲	۱	۴	۳	۰.۴	۰.۲
4	0.85	−	۲	۲	۳	۳	۰.۴	۰.۴
5	0.85	−	۲	۳	۲	۳	۰.۴	۰.۶
6	0.85	+	۳	۳	۲	۲	۰.۶	۰.۶
7	0.76	−	۳	۴	۱	۲	۰.۶	۰.۸
8	0.53	+	۴	۴	۱	۱	۰.۸	۰.۸
9	0.43	−	۴	۵	۰	۱	۰.۸	۱
10	0.25	+	۵	۵	۰	۰	۱	۱

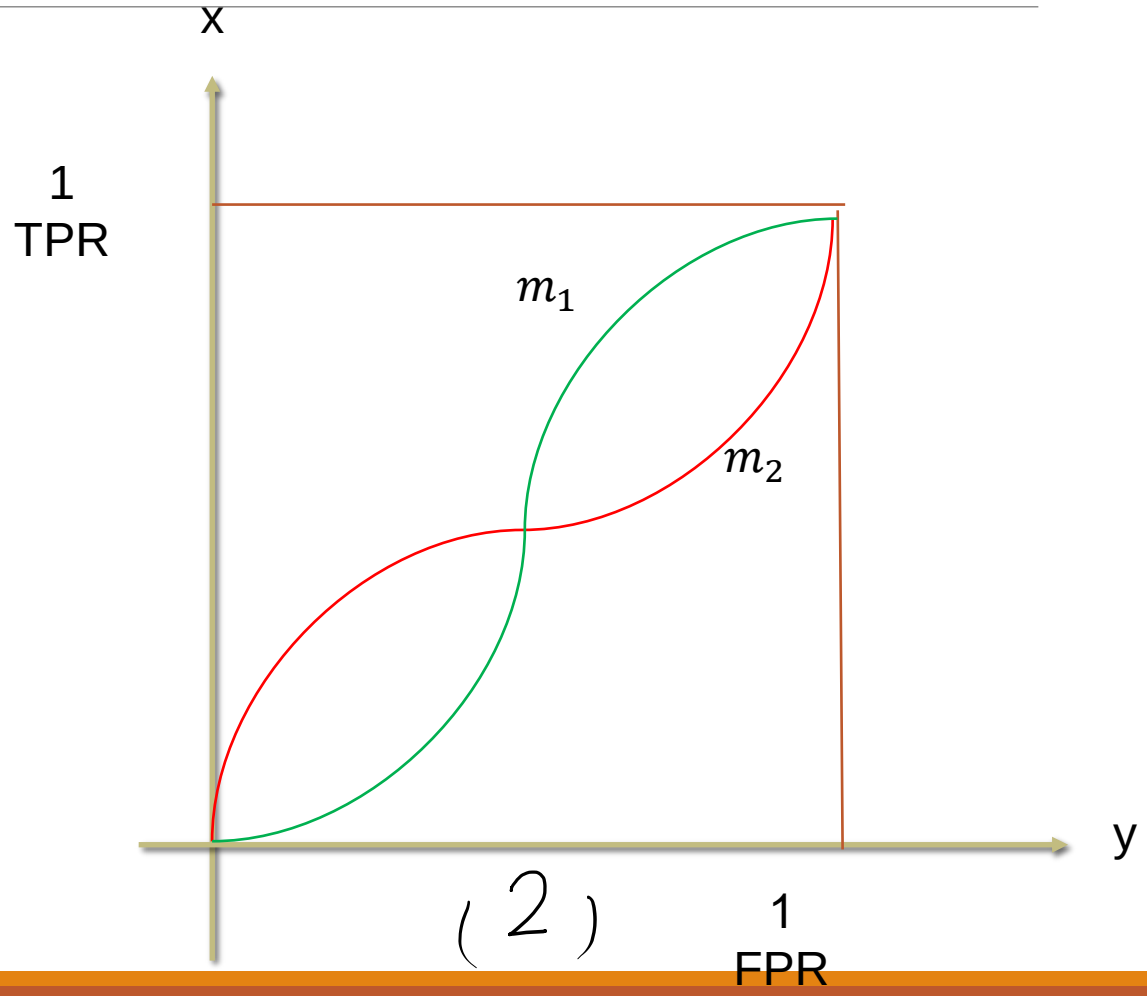

x
1
TPR
0.9
0.8
AUC = 0.6
y
0.8
1
FPR

مقایسه دو مدل m_1 و m_2 با کمک منحنی ROC



در نمودار 1 مدل m_2 نسبت به مدل m_1 همواره از دقت بالاتری برخودار است اما در نمودار 2 مشاهده میشود مساحت زیر منحنی ROC برای هر دو مدل برابر است.





در این نوع رخدادها بسته به مثال تصمیم گیری میشود مثلا مدل m_2 در مسائلی مطلوب تر است که FPR در مقایسه با TPR با اهمیت تر باشد.

مثال مربوط به این حالت تشخیص نفوذ به یک شبکه کامپیوتری میباشد، همچنینm_1 در مسائلی که معیار TPR نسبت به FPR مهم تر باشد ترجیح داده میشود مانند تشخصی مسافرین غیرمجاز در فرودگاه(بستگی به مسئله دارد)

توجه کنید ممکن است در ترسیم منحنی ROC یک از شکل های زیر اتفاق بیفتد.

ارزیابی در الگوریتم های خوشه بندی

ارزیابی در الگوهای خوشه بندی (بدون ناظر)

ارزابی الگوهای خوشه بندی را میتوان به 2 دسته زیر تقسیم کرد:

الف) معیارهای ارزیابی داخلی

ب) معیارهای ارزیابی خارجی

معیارهای ارزیابی داخلی به معیارهایی گفته میشود که تعیین کیفیت خوشه بندی با توجه به اطلاعات موجود در خود داده ها صورت گیرد اما در معیارهای ارزیابی خارجی عملکرد الگوریتم های خوشه بندی با توجه به اطلاعات اضافی و خارج از مجموعه داده های مدل بندی شده صورت میگیرد.

ارزیابی داخلی: کاری به ستون برچسب نداریم.

ارزیابی خارجی: با کمک ستون برچسب ارزیابی صورت میگیرد.

میدانیم مهمترین ویژگی های الگوریتم های خوشه بندی باید به گونه ای باشند که فاصله درون خوشه ها را مینیمم و فاصله بین خوشه ها را ماکزیمم نماید؛ زیرا، در اینصورت خوشه ها مجزاتر و تحلیل ها دقیق تر خواهد بود.

به عبارتی معیار مربوطه باید دارای دو ویژگی باشد : تراکم خوشه ای و جدایی خوشه ای

Sum of square error :SSE معیار

$$SSE = \sum_{i=1}^{k} \sum_{x \in c_i} (x - m_i)^2$$

K: تعداد خوشه ها

m_i: مرکز خوشه i ام

C_i: خوشه i ام

این معیار برای تراکم خوشه ای استفاده میشود که هرچه مقدار آن کمتر باشد الگوریتم دقیق تر است.

معیارBSSE: Between Sum of Square Error

$$BSSE = \sum_{i=1}^{k} n_i (m - m_i)^2$$

که در آن k تعداد خوشه ها

m_i مرکز خوشه i ام

m مرکز همه داده ها

n_i حجم خوشه i ام

این معیار جدایی خوشه ها را اندازه گیری میکند. الگوریتمی که BSSE بیشتری داشته باشد، دقیق تر است.

معیار Average Silhouette Coefficient :ASC

این معیار که ترکیبی از تراکم خوشه ای و جدایی خوشه ای میباشد به صورت زیر تعریف میشود. برای محاسبه این معیار باید دو معیار $a(x)$ و $b(x)$ محاسبه و در یک رابطه به نام $s(x)$ قرار گیرند.

$$a(x) = \frac{1}{n_i} \sum_{y \in c_i; y \neq x} dis(x,y)$$

که در آن c_i داده های خوشه i ام بوده و $x \in c_i$ و n_i حجم خوشه i ام است.

(متوسط فاصله بین نقطه x تا نقاط دیگر متعلق به همان خوشه)

$$b(x) = \min_{j;j\neq i}\{\frac{1}{n_j}\sum_{y\in c_j}^{k} dis(x,y) \qquad i \neq j \qquad \begin{cases} x \in c_i \\ y \in c_j \end{cases}$$

با کمک $a(x)$ و $b(x)$ معیار $s(x)$ به صورت زیر تعریف میشود.

$$s(x) = \frac{b(x)-a(x)}{max\{b(x),a(x)\}}$$

هرچه $s(x)$ بیشتر باشد الگوریتم خوشه بندی در خوشه i ام دقیق تر است.

$$\bar{s}(x) = \frac{1}{N}\sum_{x\in N} s(x)$$

N: تعداد رکوردها

$\bar{s}(x)$ متوسط $s(x)$ برای تمام x های متعلق به خوشه ها میباشد.

$\bar{s}(x)$ نشان میدهد به طور متوسط چقدر نقاط خوشه i ام به هم نزدیک و از سایز خوش ها دور میباشند. این معیار نماینده کیفیت خوشه بندی میباشد که هرچه این مقدار بیشتر باشد کیفیت خوشه بندی بالاتر است.

معیارهای ارزیابی خارجی

در این معیارهای ارزیابی الگوریتم های خوشه بندی با کمک داده هایی خارج از مجموعه داده های مورد بررسی انجام میشود و در پایان کیفیت الگوریتم ها مقایسه میشود. اصولا این داده ها همان ستون برچست و یا ستون اطلاعات با ناظر میباشند.

اگرچه در روش خوشه بندی از این ستون استفاده نمیکنیم ولی میتوان برای ارزیابی عملکرد الگوریتم های خوشه بندی از این ستون برچسب کمک بگیریم

سوال: درصورتی که در داده ها موجو ستون برچسب دراختیار ماست به چه دلیلی از الگوریتم های دسته بندی استفاده نمیکنیم؟

پاسخ:

استفاده از الگوریتم های خوشه بندی به معنای حذف روش های دسته بندی نیست بلکه با استفاده روشهای خوشه بندی میخواهیم به اطلاعات دیگری در درون داده ها دست یابیم و از ستون برچسب تنها برای ارزایب الگوریتم های خوشه بندی استفاده میکنیم.

شاخص Random Index :RI

در این شاخص با کمک ستون برچسب (در صورت وجود این ستون) میتوان اطلاعات جدول زیر را کامل کرد.

پیش بینی با کمک خوشه بندی اطلاعات موجود	خوشه یکسان	خوشه متفاوت
دسته یکسان	TP	FN
دسته متفاوت	FP	TN

TP: تعداد رکوردهایی که در یک دسته و در یک خوشه قرار دارند.
TN: تعداد رکوردهایی که در دسته های متفاوت بودند و در خوشه های متفاوت قرار گرفتند.
FN: تعداد رکوردهایی که در یک دسته هستند ولی در خوشه های متفاوت قرار دارند.
FP: تعداد رکوردهایی که در درسته های متفاوتی بودند و در خوشه های یکسان قرار گرفتند.

$$RI = \frac{(TP+TN)}{TP+FN+FP+TN}$$

هرچه مقدار شاخص RI بیشتر باشد (نسبت رکوردهایی که متتناسب با ستون برچسب خوشه بندی شدند)، الگوریتم خوشه بندی دقیق تر است.

مثال:

اطلاعات زیر برای یک الگوریتم خوشه بندی محاسبه شده است. مطلوبست شاخص RI.

	+	-
+	20	24
-	20	72

$$RI = \frac{20+72}{20+24+20+72} = 0.6765$$

معیار آنتروپی:

در این معیار P_{ij} را به صورت زیر تعریف میکنیم:

$$p_{ij} = \frac{n_{ij}}{n_i}$$

که در آن n_{ij} تعداد رکوردهای دسته i ام میباشد که در خوشه j ام قرار گرفته است.

n_i تعداد رکوردهای دسته i ام است.

در این صورت:

$$e_j = -\sum_{i=1}^{k} p_{ij} \log(p_{ij}) \qquad j = 1, \dots, k'$$

k تعداد دسته ها و k' تعداد خوشه ها

آنتروپی کل

$$Entropy = \sum_{j=1}^{k'} \frac{n_j}{n} e_j$$

n تعداد کل رکورد ها

میدانیم هرچه آنتروپی کمتر باشد الگوریتم مورد نظر بهتر است.

مثال

17 رکورد وجود دارد که به سه دسته (3 نوع رنگ سبز، نارنجی و قرمز) تقسیم شده است. اگر این 17 مشاهد در سه خوشه قرار گرفته باشند، با کمک اطلاعات جدول زیر مطلوبست محاسبه آنتروپی کل این الگوریتم خوشه بندی.

یک سری از مشاهدات ممکن است در هیچ کدام از خوشه ها قرار نگیرند یا یک مشاهده در دو خوشه قرار داشته باشد.

	خوشه اول	خوشه دوم	خوشه سوم
دسته سبز 8	$p_{11}=\frac{5}{8}$	$p_{12}=\frac{1}{8}$	$p_{13}=\frac{0}{8}$
دسته نارنجی 5	$p_{21}=\frac{1}{5}$	$p_{22}=\frac{4}{5}$	$p_{23}=\frac{1}{5}$
دسته قرمز 4	$p_{13}=\frac{2}{4}$	$p_{32}=\frac{0}{4}$	$p_{33}=\frac{3}{4}$

$$n=17$$
$$k'=k=3$$

$$e_1 = -\frac{5}{8}\log\left(\frac{5}{8}\right) - \frac{1}{5}\log\left(\frac{1}{5}\right) - \frac{2}{4}\log\left(\frac{2}{4}\right) = 0.9622$$

$$e_2 = -\frac{1}{8}\log\left(\frac{1}{8}\right) - \frac{4}{5}\log\left(\frac{4}{5}\right) = 0.4384$$

$$e_3 = -\frac{1}{5}\log\left(\frac{1}{5}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.5376$$

$$Entropy = \sum_{j=1}^{3} \frac{n_j}{n} e_j = \frac{8}{17} e_1 + \frac{5}{17} e_2 + \frac{4}{17} e_3 = 0.4528+0.1289+0.1265=0.7082$$

معیار خالصی: Purity

$$p_j = \max(p_{ij}) \qquad ; j = 1, \dots, k'$$

$$Purity = \sum_{j=1}^{k} \frac{n_j}{n} p_j$$

مثال: مطلوبست معیار خالصی در الگوریتم خوشه بندی برای مثال قبل.

n_j	8	5	4
p_j	$p_1 = \frac{5}{8}$	$p_2 = \frac{4}{5}$	$p_3 = \frac{3}{4}$

$$Purity = \frac{8}{17}\left(\frac{5}{8}\right) + \frac{5}{17}\left(\frac{4}{5}\right) + \frac{4}{17}\left(\frac{3}{4}\right) = 0.294+0.235+0.176=0.705$$