مدل چیست؟ تفاوت اون با الگوریتم چیه؟

|  |  |  |  |
| --- | --- | --- | --- |
| ساختار یا عبارت | ریاضیاتی | احتملاتی | پیروی داده ها از مدل |
|  |  |  |  |
| مدل شامل مجموعه ای از فرضیات است | الگوریتم شامل مجموعه ای از step ها است | الگوریتم ممکن است برای فیت کردن مدل به داده ها بکار برده شود | گاهی بیش از یک الگوریتم نیاز است برای برازش یک مدل |
|  |  |  |  |

هر نوع ساختار یا عبارتی میشه مدل؟ شناسایی ناپذیری در مدل چیه؟ میتونی یه مدل شناسایی ناپذیر مثال بزنی؟

|  |  |  |  |
| --- | --- | --- | --- |
| تعریف شناسایی پذیری | تعریف ریاضیاتی | مشکلات مدل در صورت عدم شناسایی پذیری | پیروی داده ها از مدل |
|  |  |  |  |

پیشامد چیست؟

|  |  |  |
| --- | --- | --- |
| فضای نمونه | فضای احتمال | اندازه پذیری |
|  |  |  |

متغیر تصادفی چیست؟

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| تابع | دامنه | برد | فضای احتمال | اندازه پذیر بودن |
|  |  |  |  |  |

فرض کنید دو کیسه داریم که به ترتیب ۳ و ۵ گوی باشد. کیسه اول حاوی گوی‌های به رنگ قرمز یا آبی است. کیسه دوم حاوی گوی‌های به رنگ سیاه و سفید است. فرض کنید یه کیسه به تصادف انتخاب میکنیم و یه گوی بر می‌داریم. الف) فرض کنید A پیشامد اینکه این گوی سفید باشد. مطلوب است محاسبه احتمال وقوع این پیشامد. ب)فرض کنید متغیر تصادفی X برابر صفر است اگر A رخ دهد و در غیر این صورت صفر است. در مورد X چه می‌توان گفت.

تکیه‌گاه متغیر تصادفی چیست؟ چطور یک نقطه خاص می‌تواند عضو تکیه‌گاه باشد؟ تعریفی کلی ارائه دهید که هم برای متغیرهای تصادفی گسسته و هم پیوسته صدق کند.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| اشاره به برد | تعریف گسسته | تعریف پیوسته | تعریف مبتنی بر تابع توزیع | تعریف مبتنی بر همسایگی |
|  |  |  |  |  |

روش بیزی و روش حداکثر درستنمایی را توضیح دهید. تفاوت این دو در نحوه برآورد پارامتر Ɵ را توضیح دهید.

|  |  |  |  |
| --- | --- | --- | --- |
| اشاره به تابع لایکلیهود | اشاره به پارمترهای مدل | اشاره با ماکسیمم سازس | فلسفه ماکسیمم سازی |
|  |  |  |  |
| متغیر تصادفی بودن پارامتر | فلسفه تصادفی بودن پارامتر | اشاره به توزیع پیشین | اشاره به توزیع پیسن |
|  |  |  |  |

اگر از روش‌های بیزی استفاده کردین میتونین بگین روش‌های مونت کارلو چیه؟ و کاربرد اون توی بیز چیه؟

|  |  |  |  |
| --- | --- | --- | --- |
| اشاره محاسبه انتگرال به صورت سامیشن | اشاره به تولید نمونه | اشاره به تولید نمونه از تووزیع پسین | اشاره به مساله حدی |
|  |  |  |  |
| اشاره به محاسبه آماره مورد نظر بر اساس مشاهدات |  |  |  |
|  |  |  |  |

پیش بینی چیست؟ پیش گویی چیست؟ forecast and prediction

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| تعریف اولیه | ترجمه درست به فارسی | شرح تفاوت | مثال تفاوت | اشاره به اینکه یکی حالت خاص اونیکی است |
|  |  |  |  |  |

تفاوت برآورد و پیش گویی چیست؟ estimation and prediction

|  |  |  |  |
| --- | --- | --- | --- |
| ترجمه درست به فارسی | اشاره به پارامتر | اشاره به متغیر تصادفی | شرح تفاوت |
|  |  |  |  |

متغیر پنهان چیست؟ در طول دوره کاریتون چه تجربه ای داشتین؟

|  |  |  |  |
| --- | --- | --- | --- |
| متغیر بودن | غیر قابل مشاهده بودن | اثر گذاری روی مدل | یک مثال |
|  |  |  |  |
| چگونگی کشف وجود متغیر پنهان | مشاوره با متخصص جهت دلیل این اثر |  |  |
|  |  |  |  |

سه مثال بزنید که نتوان از روش‌های شبکه‌های عصبی استفاده نمود؟

|  |  |  |  |
| --- | --- | --- | --- |
| اشاره به حجم داده ها | محاسبه اثر | تفسیرپذیری | اشاره به محدودیت های NN |
|  |  |  |  |

چرا نمی‌توان از رگرسیون برای کلاس بندی استفاده کرد؟

مساله کلاس بندی به این صورت است که بیماری در بیمارستان است که دچار سکته شده باشد، آوردوز کرده یا دچار صرح شده است. فرض کنید بر اساس یک سری متغیر کمکی بخواهیم این مساله کلاسه بندی رو با رگرسیون خطی حل کنیم. متغیر تصادفی Y سه مقدار ۱، ۲ و ۳ را قبول میکند.....

|  |  |  |
| --- | --- | --- |
| اشاره به یکسان نبودن تکیه گاه در دو سمت معادله | اشاره مساله شباهت ۱ با ۲ و ۳ | اشاره به مساله order کردن با ۱، ۲ و ۳(بزگتر و کوچکتر) |
|  |  |  |
| یکسان بودن تفاوت۲ و۳ با تفاوت ۲و۱ | مساله جابجایی سکته بجای ۱ بشود ۳ | مساله نسبت دادن اعداد دیگر ۰، ۲و۹ که موجب تغییر نتایج و تفسیر آن می‌شود |
|  |  |  |

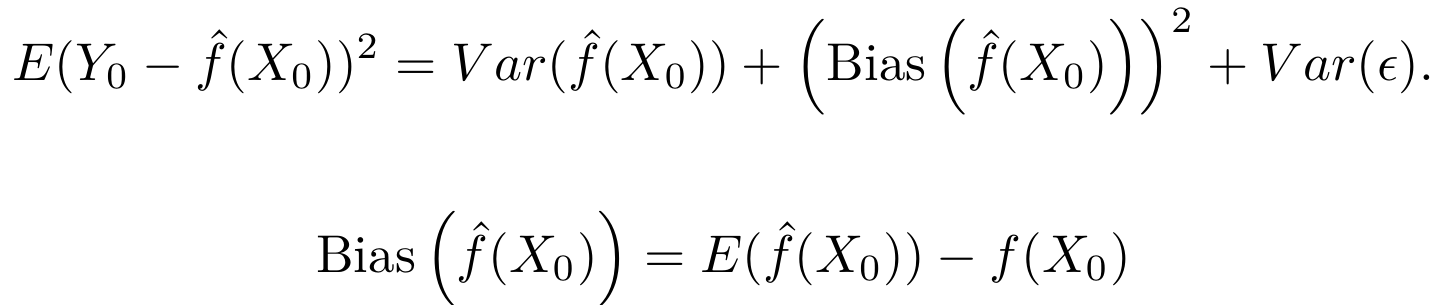
فرض کنید متغیر Y نسبت افرادی که دچار قند خون به کل افراد است یا مثلا نسبت کل خریداران به نسبت کل بازدید کنند‌گان است. فقط Y در دسترس است(تعداد بازدیدکنندگان و خریدران در دسترس نیست) یه سری متغیر کمکی هم داریم. نظرتون در مورد رگرسیون خطی چیه در اینجا؟ مدلی خاص مد نظرتون هست که به این داده‌ها برازش بدیم؟

|  |  |  |
| --- | --- | --- |
| اشاره به یکسان نبودن تکیه گاه در دو سمت معادله | اشاره به نحوه مدل سازی درست | اشاره به GLM |
|  |  |  |

مشکلات داده های با بعد بالا چیه؟ یه مثال بزن در مورد این داده‌ها

|  |  |  |
| --- | --- | --- |
| اشاره به اینکه محاسبات معمول قابل انجام نیست | اشاره به overfitting بخاطر اینکه بیشتر داده ها در حاشیه هستن و یک نمونه تا نمونه دیگه کاملا متفاوت است | ممکمن است با ناپذیر بودن ماتریس X^TX مواجه شویم |
|  |  |  |
| حتی در صورت معکوس پذیری overfitting. مثلا رگرسیون با دو متغیر و دو مشاهده | واریانس برآوردش صفر می‌شود | مساله هم‌خطی چندگانه زیرا به دلیل بعد زیاد می توان هر متغیر کمکی را با بقیه تببین نمود و بنابراین تفسیر پذیری با احتیاط صورت می گیرد |
|  |  |  |

در مورد Bias-variance trade-off توضیح دهید؟ Regularization چیه؟ رگرسیون لاسو و ریج رو توضیح بدید؟



|  |  |
| --- | --- |
| The equation tells us that in order to minimize the expected test error, we need to select a statistical learning method that simultaneously achieves low variance and low bias | مفهوم واریانس و overfitting  Variance refers to the amount by which ˆf would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different ˆf. |
|  |  |
| اشاره به ماهیت نمیتونیم هر دو خطارو همزمان کم کنیم  In general, more flexible statistical methods exhibit higher variance and lower bias. | چرا با افزایش انعطاف پذیری واریانس ممکن است افزایش چشمگیری داشته باشد Conversely, flexible models often become closely fitted to the training data, leading to greater variability between training data and new data. |
|  |  |

معایب روش تبدیل چیست؟

|  |  |  |
| --- | --- | --- |
| تفسیر پذیری پایین | بایاس در موقع تبدیل معکوس | از دست دادن اطلاعات(زیرا تابعی روی آن اعمال میشود-افزار نهایی زیرمجموعه ای از افراز اولی است) |
|  |  |  |
| تبدیل‌ها گاهی موجب می‌شوند پیش بینی دقت کافی نداشته باشد. | گاهی برخی نقاط داده های نامتعارف حاوی اطلاعات مهمی است اما تبدیلی لگاریتمی باعث می شود داده تبدیل یافته نرمال به نظر برسد | ممکن تبدیل های متفاوت منجر به ارتباطات متفاوت بین متغیرها شود و ممکن اس حساسیت به تبدیل داشته باشیم. |
|  |  |  |

در یه رگرسیون خطی ساده، فرض کنید شما یک یا چندتا داده نامتعارف شناسایی کردید؟ روال شما چی هست؟ چیکار می کنید؟

|  |  |  |
| --- | --- | --- |
| بیان یک رویکرد مشخص | حذف در صورت خطای انسانی | تبدیل روی داده های |
|  |  |  |
| تغییر در مدل(دم کلفت تی) یا آمیخته نمودن آن | جایگزینی | بررسی حاوی اطلاع بودن داده |
|  |  |  |

آنومالی چیست؟ چه تفاوتی با نویز و داده های پرت داره؟

|  |  |  |
| --- | --- | --- |
| آنومالی: مورد انتظار نبودن آن | عدم اهمیت نویز | مورد انتظار بودن داده پرت |
|  |  |  |
|  | نویز را میخواهیم حذف کنیم | ممکن است نتیجه خطای انسانی باشند |
|  |  |  |
| آنومالی:اهمیت (حاوی اطلاعات مهمی ) | نویز به داده‌ها اضافه شده و می‌تواند منجر به اختلال در تحلیل و تفسیر داده‌ها شود. | ممکن است نتیجه خطای اندازه گیری غیر انسانی باشند(دستگاه با خطایی خاص اندازه گیری میکند) |
|  |  |  |
| نویز می‌تواند کیفیت داده‌ها را کاهش دهد و اثرات واقعی را مخفی کن |  |  |
|  |  |  |

انواع انومالی رو بگو برای هر کدوم یه مثال بزن

|  |  |  |
| --- | --- | --- |
| Pattern-based anomalies | Context-based anomalies | Data point–based anomalies |
|  |  |  |
| مثال | مثال | مثال |
|  |  |  |

مدل‌های خطی چیه؟ مدل‌های Generalized Linear Models چیه؟

|  |  |  |
| --- | --- | --- |
| مدل خطی: خطی بودن نسبت به پارامتر | یک مثال مدل خطی | یک مثال مدل غیر خطی |
|  |  |  |
| تعریف مدل GLM و اینگه تعمیمی ازLM هستند | فلسفه G و چطور تعمیمی از قبلی است | مثال GLM |
|  |  |  |

از معیار‌های مناسبت مدل ۳ نمونه را بیان کنید.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Log-likelihood | AIC | BIC | DIC | C\_P |
|  |  |  |  |  |
| R^2 | CrossValidation |  |  |  |
|  |  |  |  |  |

در مورد Prediction Accuracy – Interpretability trade off توضیح دهید؟ اهمیت تفسیر پذیری چیست؟ یک مثال می تونی بزنی؟ به کارشناس تردد توضیح میدی که چه ارتباطی بین تصادف، تردد و اجرائیات است؟

|  |  |  |
| --- | --- | --- |
| توضیح مساله | اشاره به اهمیت تفسیر پذیری | مثال های مناسب |
|  |  |  |

روش‌های کلاس بندی Logistic Regression ، Linear Discriminant Analysis و Quadratic Discriminant Analysis، Naive Bayes

|  |  |  |  |
| --- | --- | --- | --- |
| مدل | اشاره به تابع زیان | اشاره به کلاسیفیار | مدل ریاضی |
| Logistic Regression | I(Y-^Y) | P(Y=1|X=x\_0) مستقیم |  |
| K-Nearest Neighbors | I(Y-^Y) | P(Y=1|X=x\_0) برآورد |  |
| LDA |  |  |  |
| QDA |  |  |
| Naive Bayes |  |  |

hyperplane، Support Vector Machines، Support Vector Classifiers، Maximal Margin Classifier

|  |  |  |
| --- | --- | --- |
| تعریف hyperplane | اشاره به فرمول | تقسیم فضای p بعدی به دو بخش  So we can think of the hyperplane as dividing p-dimensional space into two halves. |
|  |  |  |
| توضیح فرمول | اشاره به |  |
|  |  |  |
| Maximal Margin Classifier فرمول | Support Vector Classifiers فرمول | فرمول Support Vector Machines |
|  |  |  |
|  |  |  |

Kernel function چیه؟

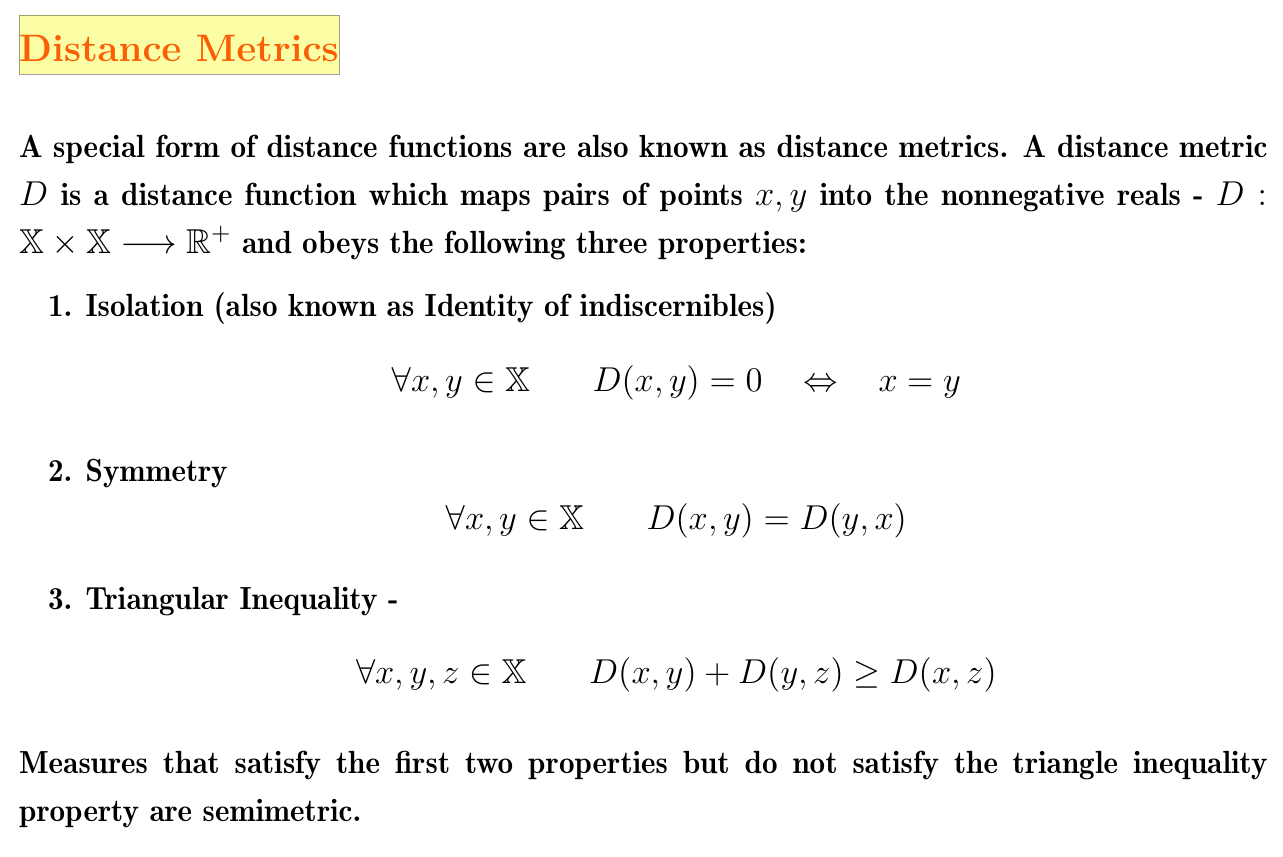
|  |  |  |
| --- | --- | --- |
| اشاره به عملگر ضرب داخلی | اشاره | اشاره به برد k که اعداد حقیقی است(یک بعدی) |
|  |  |  |
| تعریف | یک مثال از k | کاربرد |
|  |  |  |

Suppose we have a mapping that brings our vectors in to some feature space . Then

the dot product of x and y in this space is . A kernel is a function k that corresponds to this dot product, i.e. k(x, y) = ϕ(x)T ϕ(y), , Gaussian kernel

Proximity Measures, Distance Functions , Distance Metrics and Similarity Functions

|  |  |  |  |
| --- | --- | --- | --- |
| تعریف تابع فاصله | اشاره به میزان عدم شباهت، دوری و نزدیکی | تعریف مشابهت | ارتباط مشابهت و فاصله |
|  |  |  |  |
| تعریف متریک فاصله | اشاره به isolation | اشاره به symmetry | اشاره به نامساوی مثلثی |
|  |  |  |  |



Multidimensional Scaling

|  |  |
| --- | --- |
| فرمول | توضیحات کلی |
|  |  |
|  |  |
| اشاره به similarity | مثال |
|  |  |
|  |  |

آنتروپی چی هست؟ میتونی فرمولشو بر حسب امید ریاضی بنویسی؟

|  |  |
| --- | --- |
| اشاره به عدم قطعیت | فرمول |
|  |  |
|  |  |
| یه مثال در مورد پرتاب سکه بزن و قطعیت رو دستکاری کن | اشاره به اینکه چرا عدم قطعیت رو نشون میده |
|  |  |
| تابع چگالی حاوی کلیه اطلاعات | چرا این معیار جذابه؟ چه تفاوتی با واریانس داره؟ مقایسه کن؟ |
|  |  |

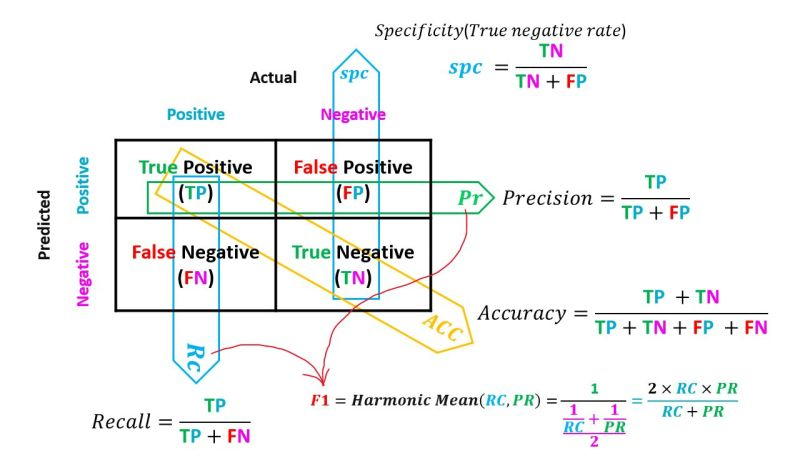
Bagging, Random Forests, Boosting, and Bayesian Additive Regression Trees

تفاوت Bagging و Random Forests چیه؟

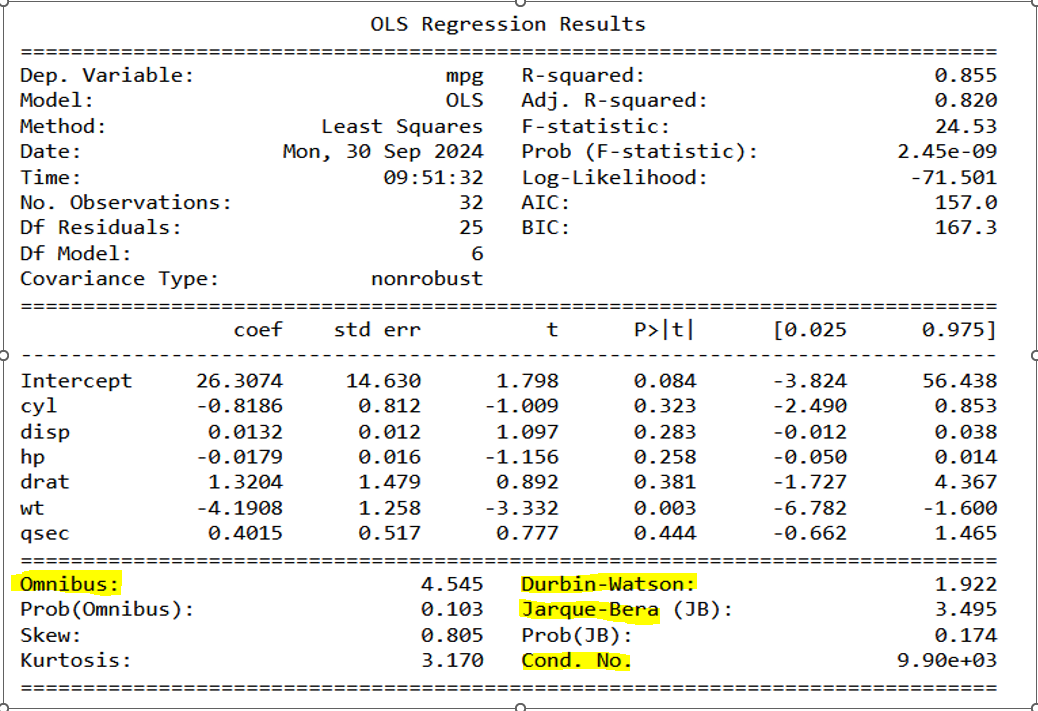
|  |  |
| --- | --- |
| تفاوت اصلی | چرا انتخاب زیر مجموعه‌ای از متغیرهای بجای کل اونها؟ |
| As in bagging, we build a number of decision trees on bootstrapped training samples. But when building these decision trees, each time a split in a tree is considered, a random sample of m=sqrt(p) predictors is chosen as split candidates from the full set of p predictors. |  |
|  |  |
| اشاره به دلیل همبستگی درخت ها | اشاره به همبسگی احتمالی درخت ها |
| Suppose that there is one very strong predictor in the data set. most or all of the trees will use this strong predictor in the top split. all of the bagged trees will look quite similar to each other. | Hence the predictions from the bagged trees will be highly correlated. |
|  |  |
| چرا همبستگی درخت ها مناسب نیست؟ (پیش بینی ها نزدیک هم هستن، همبستگی مثبت دارن و تنوع کاهش پیدا میکنه) | چرا همبستگی درخت ها مناسب نیست؟ (فرمول) |
|  |  |
|  |  |

در مورد تصویر زیر نظر شما چیست؟

|  |  |
| --- | --- |
| Accuracy: This metric shows us to what extent the predicted data by our model matches the actual data. | Precision: This metric shows us how much of the data that our model classified as positive are actually positive. |
|  |  |
| Recall: This metric indicates the proportion of the actual positive data that our model classified as positive. In other words, this metric shows how sensitive our model is to positive data. | Specificity: This metric indicates the proportion of the actual negative data that our model classified as negative. In other words, this metric shows how sensitive our model is to negative data. |
|  |  |
| F1-measure: This metric shows us how accurate and sensitive our model is in detecting positive data. In other words, this metric is a combination of Precision and Recall. | اگر من بخوام معیاری معرفی کنم که دقت و حساسیت رو برای داده‌هایی که مقدار واقعی شون negative است چی پیشنهاد میدی؟ |
|  |  |



میتونی بخش های هایلایت شده رو تفسیر کنی؟



میتونی از اندازه های فاصله بین دو کلاستر چندتا رو بگی؟

|  |  |  |
| --- | --- | --- |
| Single linkage: It measures the distance between the closest members of the clusters. | Complete linkage: It measures the distance between the  most distant members. | • Comparison of centroids: It measures the distance between the centers of the clusters. |
|  |  |  |