Applied Data Science

Session 29: Evaluation Metrics – Classification, Regression & Clustering

Dr. Soharab Hossain Shaikh

1

Regression Metrics

2

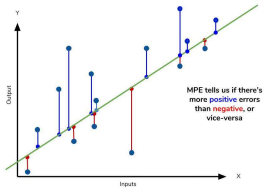
Regression Metrics

Mean squared error	$\text{MSE} = \frac{1}{n}\sum_{t=1}^{n} e_t^2$
Root mean squared error	$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_t^2}$
Mean absolute error	$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n} \lvert e_t \rvert$
Mean absolute percentage error	$\text{MAPE} = \frac{100\%}{n}\sum_{t=1}^{n} \lvert \frac{e_t}{y_t} \rvert$

$(y_i - \hat{y}_i)$

3

Regression Metric – MAPE



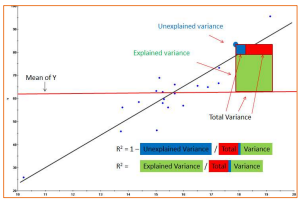
$$MAPE = \frac{100\%}{n}\Sigma\left|\frac{y-\hat{y}}{y}\right|$$

$\hat{y}$ is smaller than the actual value
n = 1 $\hat{y}$ = 10 y = 20
MAPE = 50%

$\hat{y}$ is greater than the actual value
n = 1 $\hat{y}$ = 20 y = 10
MAPE = 100%

4

R-Squared (R^2) - The Coefficient of Determination



Sum of Squares Regression → $SSR = \sum(\hat{y} - \bar{y})^2$

(measure of explained variation)

Sum of Squares Error → $SSE = \sum(y - \hat{y})^2$

(measure of unexplained variation)

Sum of Squares Total → $SST = \sum(y - \bar{y})^2$

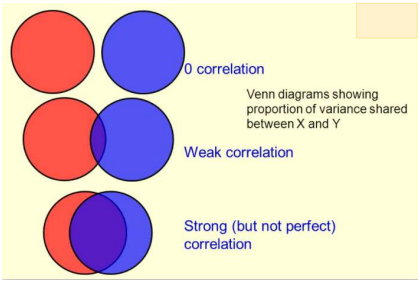
(measure of total variation in y)

The Total Sum of Squares SST = SSR + SSE.

Coefficient of Determination → $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$

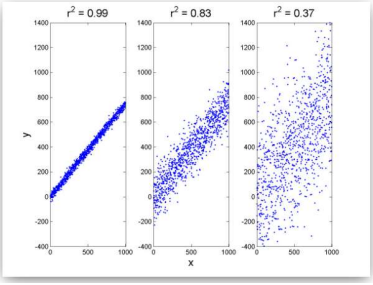
5

R^2 – The Coefficient of Determination



6

R^2 – The Coefficient of Determination



7

R-Squared (R^2) – Goodness of Fit

R^2 stastisic: The proportion of variance explained

TSS: Total Sum of Squares

How much variation is removed by the regression

$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ This is total variation in y.

$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$

RSS: Total Sum of Residuals

$RSS = \sum_{i=1}^{n}\hat{e}_i^2$ This is our estimation of variation in ϵ.

8

Adjusted R^2

$$R_a^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times (1 - R^2)\right]$$

where:

n = number of observations

k = number of independent variables

R_a^2 = adjusted R^2

Concerning R^2, there is an adjusted version, called **Adjusted R-squared**, which adjusts the R^2 for having too many variables in the model.

9

R-Squared (R^2) vs. Adjusted R^2

Vars	R-Sq	R-Sq(adj)
1	72.1	71.0
2	85.9	84.8
3	87.4	85.9
4	89.1	82.3
5	89.9	80.7

Here, the example shows how the adjusted R-squared increases up to a point and then decreases. On the other hand, R-squared blithely increases with each and every additional independent variable.

In this example, we might want to include only three independent variables in their regression model. An under-specified model (too few terms) can produce biased estimates. However, an over-specified model (too many terms) can reduce the model's precision.

In other words, both the coefficient estimates and predicted values can have larger margins of error around them.
That is why we don't want to include too many terms in the regression model!

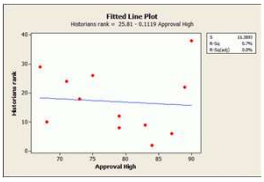
10

Predicted R-Squared

- **Predicted R-squared is used to determine how well a regression model makes predictions**.
- This statistic helps you identify cases where the model provides a good fit for the existing data but isn't as good at making predictions.
- However, even if one is not using the model to make predictions, predicted R-squared still offers valuable insights about your model.
- **Statistical software calculates predicted R-squared using the following procedure:**
  - **> It removes a data point from the dataset.**
  - **> Calculates the regression equation.**
  - **> Evaluates how well the model predicts the missing observation.**
  - **> And repeats this for all data points in the dataset.**
- Predicted R-squared helps you determine whether you are overfitting a regression model.
- Again, an overfit model includes an excessive number of terms, and it begins to fit the random noise in your sample.
- By its very definition, it is not possible to predict random noise. Consequently, if your model fits a lot of random noise, the predicted R-squared value must fall.
- A predicted R-squared that is distinctly smaller than R-squared is a warning sign that you are overfitting the model. Try reducing the number of terms.

11

Example of an Overfit Model and Predicted R-squared



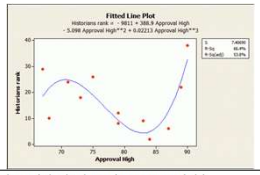
- These data come from an analysis performed that assessed the relationship between the *highest approval rating that a U.S. President achieved and their rank by historians*.
- There seems no correlation between these variables, as shown in the fitted line plot.
- It is nearly a perfect example of no relationship because it is a flat line with an R-squared of 0.7%!

Data Source: https://statisticsbyjim.com/wp-content/uploads/2017/04/PresidentRanking.csv

12

Example of an Overfit Model and Predicted R-squared

- Now, imagine that we are chasing a high R-squared and we fit the model using a cubic term that provides an S-shape.
- R-squared and adjusted R-squared look great! The coefficients are statistically significant because their p-values are all less than 0.05.
- We are just twisting the regression line to force it to connect the dots rather than finding an actual relationship.
- We overfit the model, and the **predicted R-squared of 0%** gives this away.
- If the predicted R-squared is small compared to R-squared, you might be over-fitting the model even if the independent variables are statistically significant.



Regression Analysis: Historians rank versus Approval High

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	3	867.10	289.034	5.27	0.027
Approval High	1	438.35	438.347	7.99	0.022
Approval High*Approval High	1	460.23	460.225	8.39	0.020
Approval High*Approval High*Approval High	1	481.55	481.552	8.78	0.018
Error	8	438.90	54.862		
Lack-of-Fit	7	430.90	61.557	7.69	0.271
Pure Error	1	8.00	8.000		
Total	11	1306.00			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.40690	66.39%	53.75%	0.00%

13

A Caution about the Problems of Chasing a High R-squared

- All study areas involve a certain amount of variability that cannot be explained.
- If we chase a high R-squared by including an excessive number of variables, we force the model to explain the unexplainable.
- This is not good. While this approach *can* obtain higher R-squared values, it comes at the cost of misleading regression coefficients, p-values, R-squared, and imprecise predictions.
- Adjusted R-squared and predicted R-square help you resist the urge to add too many independent variables to your model.
- Adjusted R-square compares models with different numbers of variables.
- Predicted R-square can guard against models that are too complicated.
- The great power that comes with multiple regression analysis requires user's restraint to use it wisely.

14

Probabilistic Model Selection

- The best model is chosen with the help of **probability framework of log-likelihood under Maximum Likelihood Estimation**.
- The quality of statistical methods can be measured by **Information Criteria** (IC) with some score. So, it refers to model selection methods **based on likelihood functions**. Lowest the score, best the model. This has come from the Information Theory of Statistics.
- Takes into account of **model performance and complexity** while another model selection technique of resampling checks only model performance.

> Model is chosen by a scoring method where scores are based on:
- **Performance** on train data is evaluated using **log-likelihood** which comes from the concept of MLE so as to optimize model parameters. It says about how well your model is fitted with your data.

> It provides an indication of total error.
- **Model Complexity** is evaluated using number of parameters (or degrees of freedom) in model.

15

IC-based Regression Metrics

1. AIC stands for (*Akaike's Information Criteria*), a metric developped by the Japanese Statistician, Hirotugu Akaike, 1970. The basic idea of AIC is to penalize the inclusion of additional variables to a model. It adds a penalty that increases the error when including additional terms. The lower the AIC, the better the model.
2. BIC (or *Bayesian information criteria*) is a variant of AIC with a stronger penalty for including additional variables to the model.
3. Mallow's Cp: A variant of AIC developed by Colin Mallows.

16

IC-based Regression Metrics

Score = kp - 2 log(L)

Model Complexity — Model Performance

where

k = For AIC: 2
For BIC: log (sample-size)
L = Likelihood function (mse, log_loss)
p = No of parameters

Applicability to both linear and non-linear models: Since AIC/BIC are based on log-likelihood function for a model which you can have for both linear and non-linear models.

17

AIC (*Akaike's Information Criteria*) and AIC Corrected

AIC = 2k - 2log(L)

k= number of independent variables to build model
L= maximum likelihood estimate of model

- For maximizing likelihood log(L), more variables supposed to be added in model, leads to overfitting.
- Hence, "2k" penalty term introduced which doesn't remove overfitting completely. Because of weak penalty included.
- This formula doesn't take observations into account instead only the model parameters.

$$AICc = AIC + \frac{2k(k+1)}{n-k-1}$$

- In case of **small samples**, most chances of having over-fitting as AIC will end up selecting many parameters.
- Thus, **AIC corrected** was introduced to address this issue.
- **Smaller the value, lesser information be lost and best model fit.**

18

BIC (*Bayesian Information Criteria*)

BIC = log(n) k - 2log(L)

k= number of independent variables to build model
L= maximum likelihood estimate of model
n = sample size (#observations)
log-base = e(natural log)

For maximizing likelihood log(L), more variables supposed to be added in model, lead to over-fitting.

BIC tackles this issue by including a strong penalty of "log(n)k" which instead may fall you in under-fitting i.e. too simple models.
Simple models aren't capable of catching variations in data.

19

IC-based Regression Metrics

Model Selection Criteria : *Cp*, AIC, BIC, or adjusted *R2*, CV
1. Adjust Training Error: *Cp*, AIC, BIC, or adjusted-*R2*
2. Estimating Test Error : K fold Cross Validation

Mallow's *Cp*

$$C_p = \frac{1}{n}(\text{RSS} + 2d\hat{\sigma}^2)$$

number of predictors — estimate of error variance

AIC(Akaike Information Criterion)

$$\text{AIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + 2d\hat{\sigma}^2\right)$$

BIC(Bayesian Information Criterion)

$$\text{BIC} = \frac{1}{n\hat{\sigma}^2}\left(\text{RSS} + \log(n)d\hat{\sigma}^2\right)$$

Adjusted R-Square

$$R^2_{adj} = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}$$

20

Other Regression Metrics

Notice the formula for Cp, AIC , and BIC are in the form of "**MSE(RSS) + penalty** ".

Also notice that the penalty term consists of number variables and the estimation of the error variance.

We can see it as the cost we're imposing on models for having extra parameters.

Every new parameter has got to pay that cost by reducing the MSE by at least a certain amount; if it doesn't, the extra parameter isn't worth it.

BIC being a bit more conservative than the other two, and Adjusted R-square is only half as good as others, therefore not a very good choice for model selection.

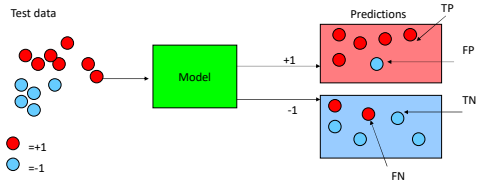
We will see its usefulness as a scoring metric due to its closeness in interpretability to R-Square and the ability to penalize complex models.

21

Classification Metrics

22

True/False Positives/Negatives



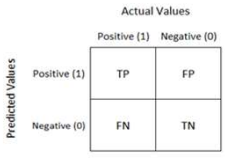
23

23

Confusion Matrix

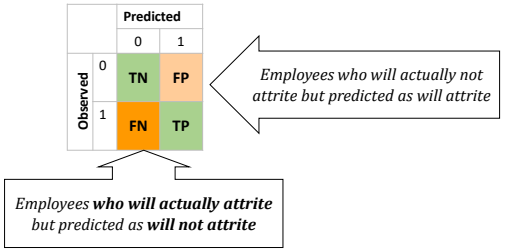
Well, it is a performance measurement for machine learning classification problem where output can be two or more classes.

It is a table with 4 different combinations of predicted and actual values.



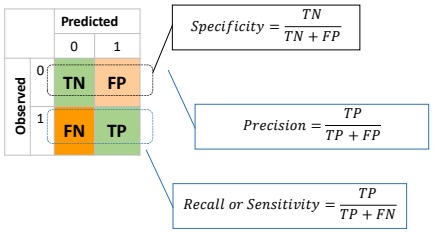
24

Confusion Matrix



25

Confusion Matrix



26

Classification Metrics

Accuracy = (Number of correctly classified samples)/ (total number of samples)

Not a good measure for problems with **heavy class imbalance**.

Better metrics required:

> Precision
> Recall (Sensitivity/TPR)
> F-score
> Specificity/TNR

$$\text{Accuracy} = \frac{tp + tn}{tp + tn + fp + fn}$$

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

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

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

27

27

Classification Metrics

- Sensitivity/Recall/True Positive rate: $\frac{tp}{tp+fn}$
- Specificity/True Negative rate: $\frac{tn}{tn+fp}$
- Precision: $\frac{tp}{tp+f}$
- F1 Score: $\frac{2\times precision\times recall}{precision+recall}$
- Accuracy: $\frac{tp+tn}{tp+tn+fp+fn}$

1 – Specificity = 1 - $\frac{tn}{tn+fp}$

= $\frac{fp}{fp+tn}$

= False Positive Rate

28

ROC and AUC

- ROC (Receiver Operating Characteristics) is a probability curve with True positive rate in the vertical axis and False positive rate on the horizontal axis for different threshold values
- AUC is the area under the ROC curve

29

Introduction to ROC Curves

- *ROC = Receiver Operating Characteristic*
- Started in electronic signal detection theory (1940s - 1950s)

"The ROC curve was first developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle fields and was soon introduced to psychology to account for perceptual detection of stimuli."

"ROC analysis since then has been used in medicine, radiology, biometrics, and other areas for many decades and is increasingly used in machine learning and data mining research."

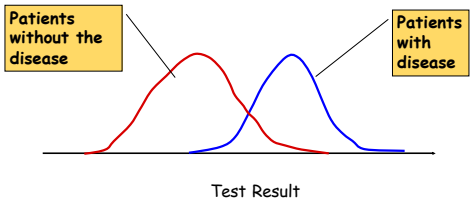
30

ROC Curves: Example

- Consider Diagnostic test for a Disease
- Test has 2 possible outcomes:
  - 'positive' = suggesting presence of disease
  - 'negative' = suggesting absence of disease
- An individual can test either positive or negative for the disease.

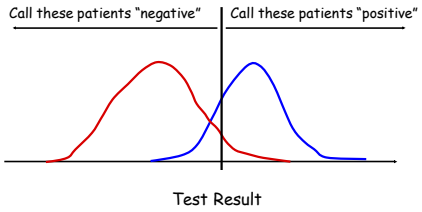
31

Specific Example



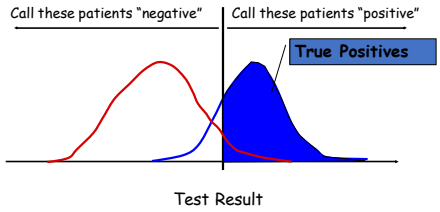
32

Threshold



33

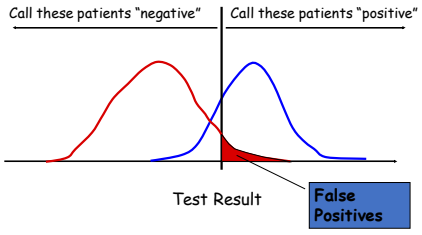
Some Definitions



without the disease
with the disease

34

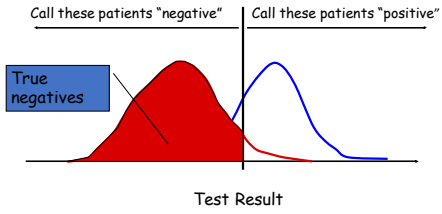
Some Definitions



without the disease
with the disease

35

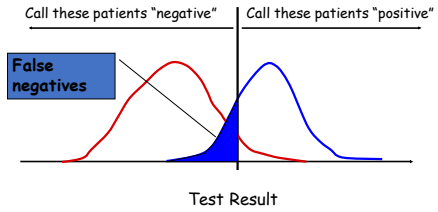
Some Definitions



without the disease
with the disease

36

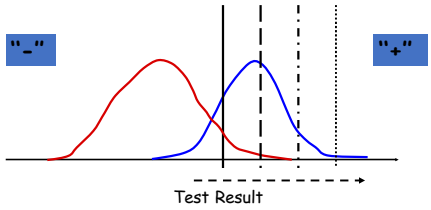
Some Definitions



without the disease
with the disease

37

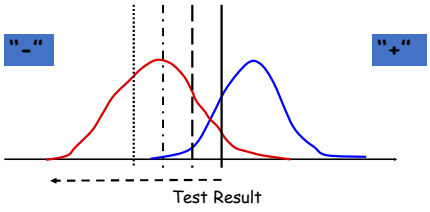
Moving the Threshold: right



without the disease
with the disease

38

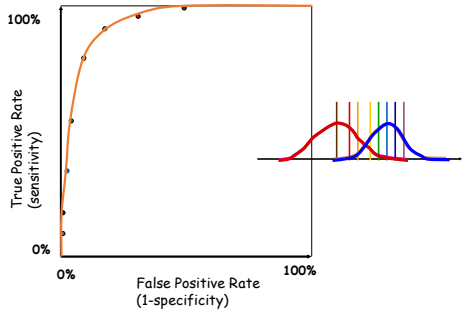
Moving the Threshold: left



without the disease
with the disease

39

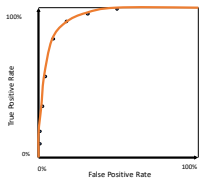
ROC Curve



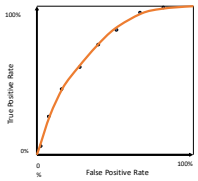
40

ROC Curve Comparison

A good test:

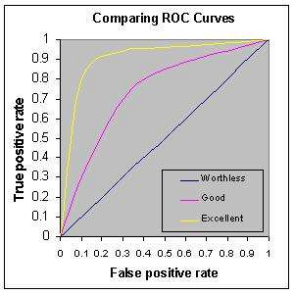


A poor test:



41

Receiver Operating Characteristic Methodology



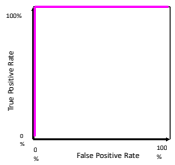
Summarizing the performance of a Binary classifier for many different threshold values.

Better than reporting a single confusion matrix.

42

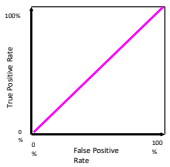
ROC Curve Extremes

Best Test:



The distributions don't overlap at all

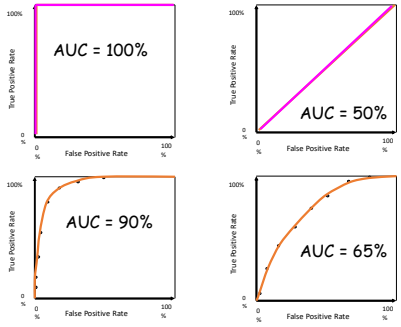
Worst test:



The distributions overlap completely

43

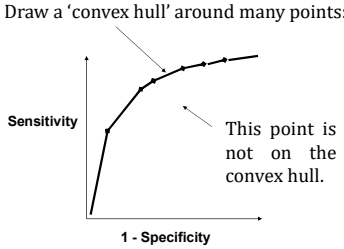
AUC for ROC Curves



44

ROC Curve

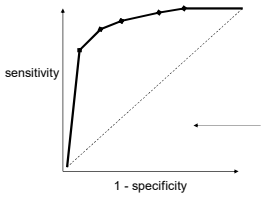
Draw a 'convex hull' around many points:



This point is not on the convex hull.

45

ROC Analysis



All the optimal detectors lie on the convex hull.

Any detector on this side can lead to a better detector by flipping its output.

Take-home point : You should always quote sensitivity and specificity for your algorithm, if possible, plotting an ROC graph.
Any statistic you quote should be an average over a suitable range of tests for your algorithm.

46

ROC Analysis

- Suitable for evaluating the performance of a Binary Classifier.
- ROC and AUC is insensitive to whether the output probability reported by the model/classifier is actually calibrated to the actual probability.
- ROC is useful even when the predicted probabilities are not properly calibrated.

47

ROC Analysis

- Can be extended to measure accuracy of a model for more than two output classes – we take an approach – one vs. all.
- Let's consider there are three output classes/categories A, B and C.
- Class A vs. all (classes B and C) [one curve]
- Class B vs. all (classes A and C) [another curve]
- Class C vs. all (classes A and B) [another curve]

48

How to choose the threshold?

- Business Decision – Minimize false Positive Rate or Maximize True Positive Rate
- Ex. Credit Card Transaction – Fraud Detection
- We may choose a low threshold – this might cause many False Positive but minimize False Negative.

49

How to choose the threshold?

Application Dependent:

Maximizing TP is a common objective.

 # System for Authentication for Access to Secure Services (e.g. finger print based authentication): **minimize FP**
We do not want any unauthorized person to access the secure services, therefore, we want to minimize false positives.
FN can be detected through further authentication mechanism (e.g. password, tag).

 # Medical Diagnosis System: **minimize FN**
We do not want to leave a patient having a decease but not detected as a positive by the system.
FP can be detected in the further steps of the medical procedures/tests etc. (however cost of diagnosis/tests etc. is a concern!!!).

50

50

Why accuracy is not a good model performance measure?

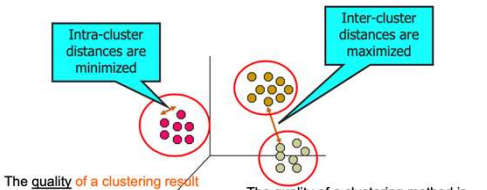
		Predicted	
		0	1
Observed	0	TN = 92	FP
	1	FN = 8	TP

51

Clustering Metrics

52

Intra-cluster and Inter-cluster Distances

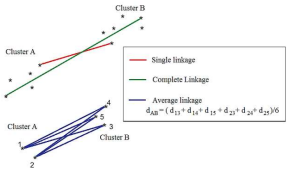


The quality of a clustering result depends on both the similarity measure used by the method and its implementation.

The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns

53

Linkage Methods



Single Linkage Method: The defining feature of the method is that the distance between groups is defined as that of the closest pair of samples, where only pairs consisting of one sample from each group are considered .

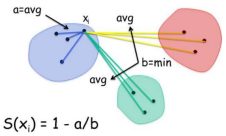
Complete Linkage Method: It is similar to the single linkage method except that the distance between two clusters is now defined as the largest distance between pairs of samples in each cluster, rather than the smallest.

Average Linkage Method: Also known as the unweighted pair-group method using the average approach (UPGMA) – the distance between two clusters is the average of the distance between all pairs of samples that are made up of one sample from each group.

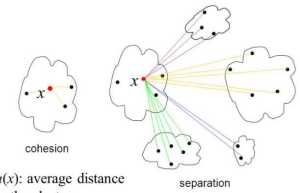
54

Silhouette Coefficient

- The idea...



- Usually, $S(x_i) = 1 - a/b$

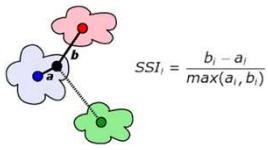


$a(x)$: average distance in the cluster

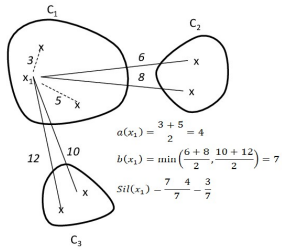
$b(x)$: average distances to others clusters, find minimal

55

Silhouette Coefficient



$$SSI_i = \frac{b_i - a_i}{max(a_i, b_i)}$$



$a(x_1) = \frac{3+5}{2} = 4$

$b(x_1) = min\left(\frac{6+8}{2}, \frac{10+12}{2}\right) = 7$

$Sil(x_1) - \frac{7}{7} - \frac{4}{7} - \frac{3}{7}$

56

Silhouette Coefficient

Assume the data have been clustered via any technique, such as k-means, into k clusters.

For data point $i \in C_i$ (data point i in the cluster C_i), let

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i, j)$$

be the mean distance between i and all other data points in the same cluster, where $d(i, j)$ is the distance between data points i and j in the cluster C_i (we divide by $|C_i| - 1$ because we do not include the distance $d(i, i)$ in the sum). We can interpret $a(i)$ as a measure of how well i is assigned to its cluster (the smaller the value, the better the assignment).

We then define the mean dissimilarity of point i to some cluster C_k as the mean of the distance from i to all points in C_k (where $C_k \neq C_i$).

For each data point $i \in C_i$, we now define

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i, j)$$

to be the *smallest* (hence the $\min$ operator in the formula) mean distance of i to all points in any other cluster, of which i is not a member. The cluster with this smallest mean dissimilarity is said to be the "neighboring cluster" of i because it is the next best fit cluster for point i.

57

Silhouette Coefficient

We now define a *silhouette* (value) of one data point i

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

and

$$s(i) = 0, \text{ if } |C_i| = 1$$

Which can be also written as:

$$s(i) = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

From the above definition it is clear that

$$-1 \leq s(i) \leq 1$$

Also, note that score is 0 for clusters with size = 1.

This constraint is added to prevent the number of clusters from increasing significantly.

58

Silhouette Coefficient

- For ***s(i)*** to be close to 1 we require ***a(i)<< b(i).***
- As ***a(i)*** is a measure of how dissimilar ***i*** is to its own cluster, a small value means it is well matched.
- Furthermore, a large ***b(i)*** implies that ***i*** is badly matched to its neighbouring cluster.
- Thus an ***s(i)*** close to one means that the data is appropriately clustered.
- If ***s(i)*** is close to negative one, then by the same logic we see that ***i*** would be more appropriate if it was clustered in its neighbouring cluster.
- An ***s(i)*** near zero means that the datum is on the border of two natural clusters.

- The mean ***s(i)*** over all points of a cluster is a measure of how tightly grouped all the points in the cluster are.
- Thus the mean ***s(i)*** over all data of the entire dataset is a measure of how appropriately the data have been clustered.
- If there are too many or too few clusters, as may occur when a poor choice of ***k*** is used in the clustering algorithm (e.g.: ***k-means***), some of the clusters will typically display much narrower silhouettes than the rest.
- Thus silhouette plots and means may be used to determine the natural number of clusters within a dataset.
- One can also increase the likelihood of the silhouette being maximized at the correct number of clusters by re-scaling the data using feature weights that are cluster specific.

59

Cophenetic Correlation Coefficient

- In statistics, **cophenetic correlation** (more precisely, the **cophenetic correlation coefficient**) is a measure of how faithfully a dendrogram preserves the pairwise distances between the original unmodeled data points.
- Suppose that the original data ***{X_i}*** have been modeled using a cluster method to produce a dendrogram ***{T_i};*** that is, a simplified model in which data that are "close" have been grouped into a hierarchical tree.
- Define the following distance measures.
- ***x(i, j) = | X_i – X_j |,*** the ordinary Euclidean distance between the *i-th* and *j-th* observations.
- ***t(i, j)*** = the dendrogrammatic distance between the model points $\boldsymbol{T_i}$ and $\boldsymbol{T_j}$.
- This distance is the height of the node at which these two points are first joined together.
- Then, letting $\overline{\mathbf{x}}$ be the average of the **x(i, j),** and letting $\overline{\mathbf{t}}$ be the average of the ***t(i, j),*** the cophenetic correlation coefficient ***c*** is given by

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}$$

60

61

To be continued in the next session…..

62