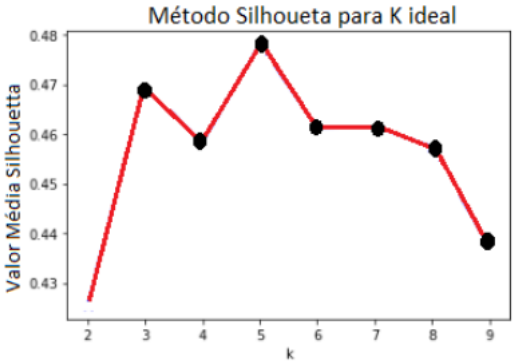
07

Para saber mais: Silhouette

O método silhueta (*silhouette*, em inglês) é uma técnica estatística para auxiliar na escolha da quantidade de cluster mais perto do ideal, esse cálculo tem como objetivo de identificar o quanto os cluster são distinto entre eles, ou seja, o quanto os cluster possuem registros bem semelhantes dentro do próprio cluster e que seja diferente entre os outros cluster.

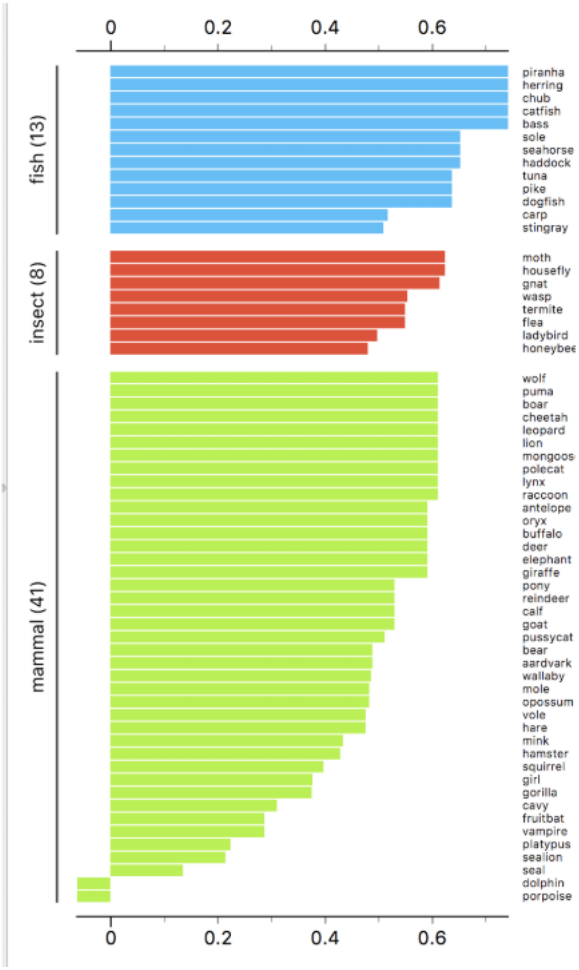
A silhueta indicará valores entre `-1` e `1`, sendo que, quanto mais próximo de `1` indica que aquela quantidade de cluster é a mais ideal, ou seja, que os cluster estão bem definidos e heterogêneos entre eles. Por exemplo, vamos visualizar o gráfico abaixo, no eixo X tem a quantidade de cluster e no eixo Y a média da silhueta.



Podemos observar no gráfico acima que o valor mais próximo de `1` é o `K=5`, ou seja, por este gráfico podemos definir a quantidade de cluster como `5`.

Muitos autores preferem este método ao *Elbow*, porque demonstra o valor mais significativo de forma clara. Ao estudarmos sobre *Elbow*, você pode observar que tivemos um pouco de dificuldade para identificar o ponto mais ideal do valor de K, e isso é muito normal nesse método, por isso que muitos preferem o método silhueta ao *Elbow*, mas se preferir pode utilizar as duas para fazer uma comparação de resultados.

No método *silhoutte*, utilizado durante a aula e no exemplo acima, foi utilizada a média entre todos os valores, porém, também é comum encontrar outra forma de visualização, como demonstrado na figura abaixo.



(fonte: https://en.wikipedia.org/wiki/Silhouette_(clustering) (https://en.wikipedia.org/wiki/Silhouette_%28clustering%29))

A ideia do gráfico acima é a seguinte: mostrar o valor da silhueta para cada elemento dentro do cluster. Por exemplo, no cluster em verde, o primeiro elemento *wolf* (em português, lobo) tem um valor aproximado `0,60`, ou seja, próximo a `1`, indica que esse elemento está no cluster correto. Porém, lá no final os dois últimos elementos, *dolphin* (golfinho) e *porpoise* (toninha) possuem valores negativos, ou seja, esses elementos são os valores discrepantes (*outliers*) dentro desse cluster.

Referências

- *Selecting optimal number of clusters in KMeans Algorithm (Silhouette Score)* (https://medium.com/@jyotiyadav99111/selecting-optimal-number-of-clusters-in-kmeans-algorithm-silhouette-score-c0d9ebb11308), artigo do Medium (https://medium.com/);
- Técnicas de agrupamento, de Estatística Avançada (http://www.estatisticaclassica.com/ea/tecnicas-de-agrupamento.html);
- Aprendizado Não Supervisionado com K-means (https://dev.to/giselyalves13/aprendizado-nao-supervisionado-com-k-means-106f);
- *Silhouette (clustering)* (https://en.wikipedia.org/wiki/Silhouette_%28clustering%29), na Wikipedia.