

Erro médio e erro médio quadrado

Vimos que o K-Means utiliza o erro para ver a qualidade do modelo. Este erro é a distância do centróide para cada dado pertencente ao grupo.

Mas como essa distância é calculada de fato?

O K-Means utiliza a seguinte função de erro. Vamos dizer que a distância de um ponto de um grupo para o centróide vale **2**. Quando somamos a distância de todos os pontos, chamamos de erro médio absoluto, ou *mean absolute error* (MAE) (https://en.wikipedia.org/wiki/Mean_absolute_error) em inglês.

Porém, queremos penalizar os pontos mais distantes do centróide - pois, um ponto muito distante do centróide pode não pertencer àquele grupo.

Uma maneira efetiva de penalizar esses dados é elevando suas distâncias ao quadrado. Isto é, aquela distancia de valor **2**, quando elevada ao quadrado, vale **4**. Portanto, o erro é maior do que antes.

Esse método é chamado de erro médio quadrado, ou *mean squarred error* (MSE) (https://pt.wikipedia.org/wiki/Erro_quadr%C3%A1tico_m%C3%A9dio).

Existe uma discussão no [Stack Exchange sobre esses dois tipos de erro](https://stats.stackexchange.com/questions/48267/mean-absolute-error-or-root-mean-squared-error) (<https://stats.stackexchange.com/questions/48267/mean-absolute-error-or-root-mean-squared-error>). É interessante dar uma olhada para ver como outras pessoas explicam a respeito do assunto. (Link em inglês)