AGREGAÇÃO DE BOOTSTRAP (BAGGING / ACONDIONAMENTO)

- O Bagging visa melhorar a precisão e o desempenho dos algoritmos de aprendizado de máquina.
- Ele faz isso pegando subconjuntos aleatórios de um conjunto de dados original e ajusta um classificador (para classificação) ou regressor (para regressão) a cada subconjunto.
- As previsões para cada subconjunto são então agregadas por meio de votação majoritária para classificação ou média para regressão, aumentando a precisão do modelo. Também reduz a variação e ajuda a evitar o overfitting.
- Embora seja geralmente aplicado a métodos de árvore de decisão, pode ser usado com qualquer tipo de método. O Bagging é um caso especial da abordagem de média do modelo.

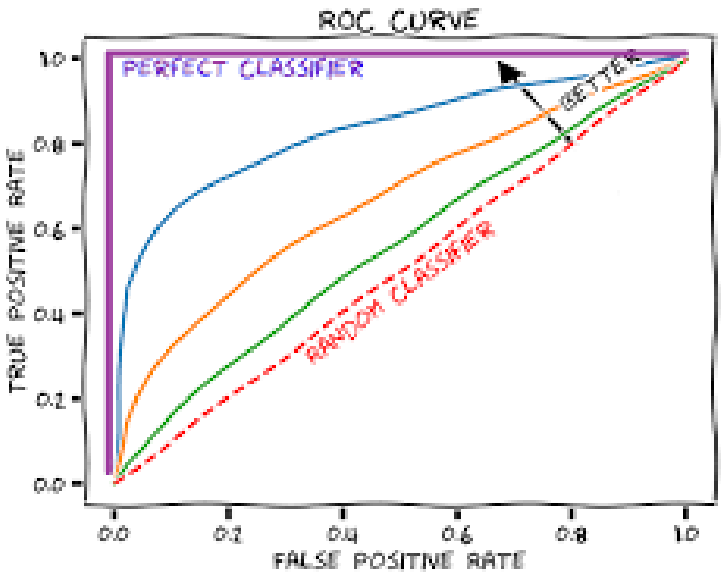
AGREGAÇÃO DE BOOTSTRAP (BAGGING / ACONDIONAMENTO)

- As etapas do algoritmo de Bagging consistem em 3 conjuntos de dados para serem processados pelo modelo de aprendizado de máquina:

- **O conjunto de dados de partida original:**
- Ex: (A, B, C, D, E, F, X, W, U, Q)
- **Bootstrap:** Gerado aleatoriamente com os dados e com a mesma quantidade de elementos do conjunto original (podendo ter elementos repetidos ou duplicados).
- Ex: (A, B, B, C, E, F, F, X, E, X)
- **Out-of-bag:** É formado por elementos resultantes da subtração entre o conjunto de dados original e o Bootstrap.
- Ex: (D, W, U, Q)

ROC CURVE

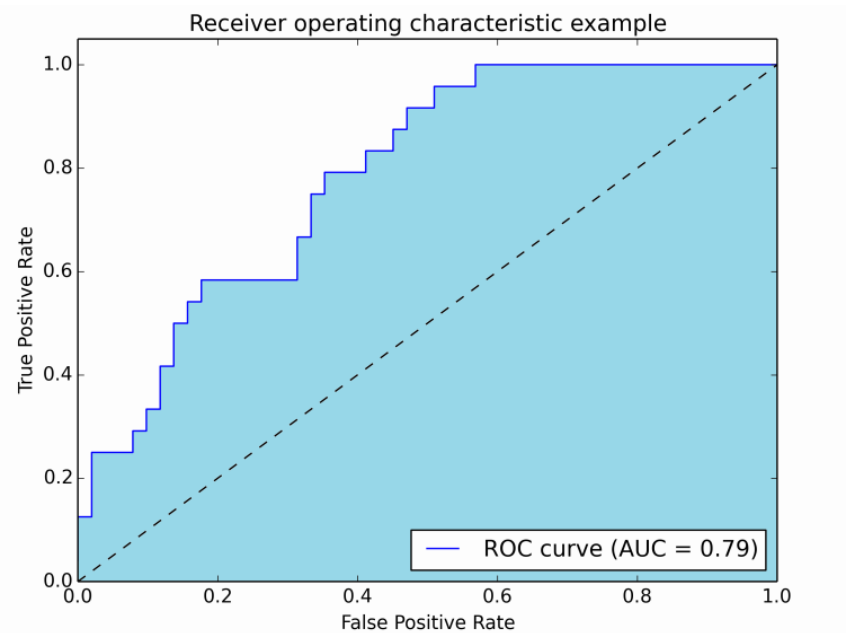
- A curva **ROC (Receiver Operator Characteristic)** mostra quão bom o determinado modelo pode distinguir entre duas coisas (já que é utilizado para classificação binária).
- Uma curva **ROC** traça "*True Positive Rate* vs. *False Positive Rate*" em diferentes limiares de classificação



Fonte: https://glassboxmedicine.files.wordpress.com/2019/02/roc-curve-v2.png?w=576&resize=398%2C299

AREA UNDER THE ROC CURVE (AUC)

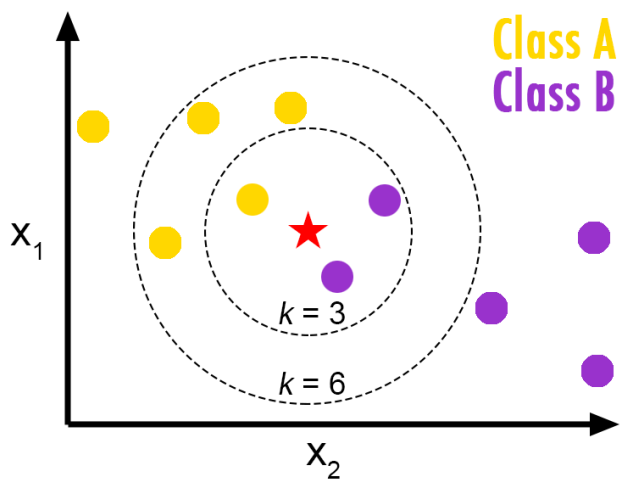
- A **AUC** nada mais é que uma maneira de resumir a curva **ROC** em um único valor, agregando todos os limiares da **ROC**, calculando a *"área sob a curva"*.
- O valor do **AUC** varia de 0,0 até 1,0. Quanto maior o valor, melhor.



Fonte: https://miro.medium.com/max/1050/1*RqK5DjVxcj4qZsCdN4FOSQ.png

K-NEAREST NEIGHBOUR (KNN)

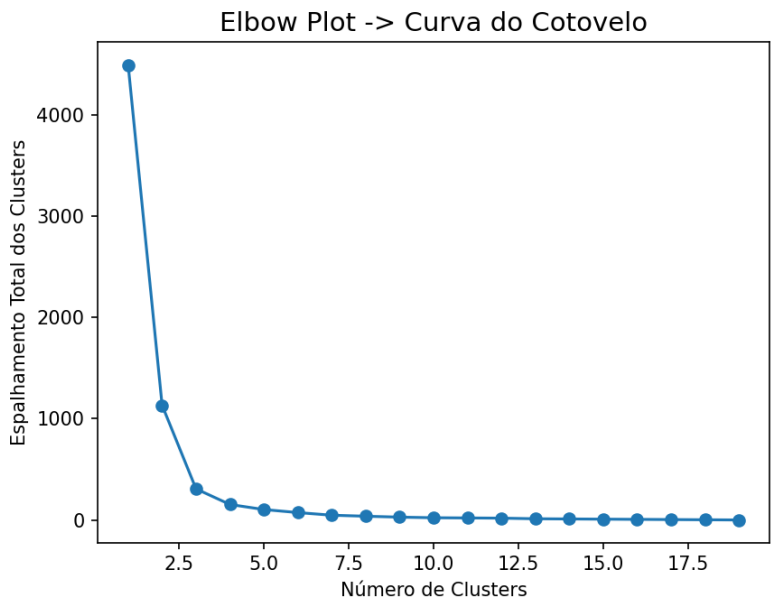
- É o tipo de modelo de classificação mais simples e intuitivo. Ao usar os dados de treinamento como base, é possível rotular um dado novo obtendo a maioria dos rótulos nos K-vizinhos mais próximos.
- Devido a natureza do algoritmo, é possível obter diferentes rótulos e resultados dependendo do valor de K escolhido. Deve-se verificar a partir das métricas e das taxas de erro qual o melhor valor para K.



Fonte:www.researchgate.net/profile/Reza_Arghandeh/publication/330400464/figure/fig9/AS:715398974562316@1547575815590/8-An-Example-of-K-nearest-Neighbors-10.png

K-MEANS

- É um método de aprendizado de máquina não supervisionado. O algoritmo divide os dados de maneira iterativa em grupos K. A cada iteração o algoritmo tenta reduzir a variância em cada grupo.
- K representa a quantidade de agrupamentos ou clusters que foi escolhido ou obtido pelo algoritmo.
- Uma maneira de determinar o melhor valor de K é através da Curva do Cotovelo (Elbow-Curve)



Exemplos

- BAGGING
- CURVAS ROC E AUC
- K-NEAREST NEIGHBORS (KNN)
- K-MEANS

RECOMENDAÇÕES

- Cada algoritmo tem uma série de parâmetros únicos, e cada um destes parâmetros podem ser modificados e irão influenciar os resultados. Além de dominar os algoritmos, deve-se dominar as características únicas de cada tipo.

- Comece sempre pelos modelos mais simples para validar a resolução do problema e só então prossiga para os métodos mais complexos.

- As métricas devem ser analisadas sempre em conjunto, principalmente para conclusões sobre comparação de modelos.