Random Forest - Teoria

Ciclo	Ciclo 06: Algoritmos baseado em árvores
# Aula	47
Created	@March 9, 2023 10:11 AM
Done	☑
Ready	☑

Objetivo da Aula:

- [ ] O conceito de Week Learning
- [ ] Como a Random Forest funciona?
- [ ] Principais características
- [ ] Limitações
- [ ] Resumo
- [ ] Próxima aula

Conteúdo:

▼ 1. O conceito de Week Learning

O conceito de Weak Learning é uma ideia fundamental por trás do algoritmo de Random Forest. Em vez de criar uma única árvore de decisão muito complexa que se ajusta perfeitamente aos dados de treinamento, o algoritmo de Random Forest cria muitas árvores mais simples, com desempenho ligeiramente melhor do que aleatório.

O termo "Weak" se refere ao fato de que cada árvore individual pode ter uma precisão limitada, ou seja, são árvores que não conseguem separar perfeitamente os dados de treino. No entanto, a combinação das previsões de todas as árvores individuais resulta em um modelo robusto e geralmente mais preciso overall.

Isso acontece porque as árvores individuais são construídas de maneira diferente, usando diferentes subconjuntos dos dados de treinamento e variáveis de entrada. Além disso, as árvores são treinadas de forma independente, o que significa que cada árvore aprende com os dados de treinamento de maneira diferente.

Ao combinar as previsões de todas as árvores individuais, a Random Forest é capaz de reduzir o impacto de erros individuais e produzir um modelo mais geral e preciso. Essa abordagem é especialmente útil quando lidamos com conjuntos de dados grandes e complexos, onde uma única árvore de decisão pode não ser suficiente para capturar todas as nuances e padrões dos dados.

▼ 2. Como a Random Forest funciona?

https://docs.google.com/presentation/d/1-z-vTwSDb9Gpb_iRPG204VPg7q36qol2jVAeQoG5CNk/edit?usp=sharing

▼ 3. A importância dos atributos

Com a Random Forest, é possível obter uma medida da importância das variáveis de entrada.

Isso é feito calculando a redução média de impureza, no caso de uma Random Forest Classifier ou uma redução no erro quadrático médio, no caso de uma Random Forest Regressor, de cada variável em todas as árvores da floresta.

Quanto maior for a redução média de impureza ou do erro quadrático médio, mais importante é a variável.

Essas medidas de importância podem ser usadas para selecionar as variáveis mais importantes para um modelo ou para entender melhor quais variáveis contribuem mais para a previsão do resultado.

Além disso, as medidas de importância podem ser usadas para detectar variáveis irrelevantes ou redundantes, o que pode ajudar a simplificar o modelo e melhorar sua precisão.

▼ 4. Principais características

1. Apresenta robustez contra outliers
2. Não é necessário normalizar e nem padronizar as variáveis.
3. É capaz de fornecer uma medida de importância das variáveis de entrada.

▼ 5. Limitações

1. Não faz extrapolação do conjunto de dados. É capaz de prever valores menores do que o máximo valor do conjunto de dados.
2. É menos interpretável do que uma única árvore de decisão, pois envolve a combinação de várias árvores.

▼ 6. Resumo

- O algoritmo de Random Forest utiliza o conceito de Weak Learning para criar múltiplas árvores de decisão mais simples, com desempenho ligeiramente melhor do que aleatório, que combinadas produzem um modelo mais geral e preciso.
- É possível medir a importância das variáveis de entrada na Random Forest calculando a redução média de impureza ou do erro quadrático médio de cada variável em todas as árvores da floresta.
- A Random Forest apresenta robustez contra outliers, não requer normalização ou padronização das variáveis e é capaz de fornecer uma medida de importância das variáveis de entrada, mas não faz extrapolação do conjunto de dados e é menos interpretável do que uma única árvore de decisão.

▼ 7. Próxima aula

Random Forest na prática