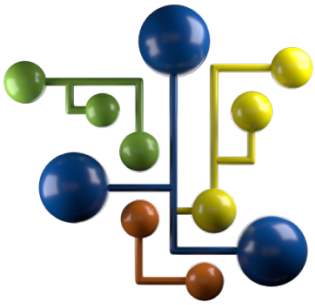

Data Science Academy

www.datascienceacademy.com.br

Machine Learning

Estado da Arte em Machine Learning

Métodos Ensemble têm uma técnica eficaz e poderosa para alcançar alta precisão em soluções supervisionadas e não supervisionadas. Diferentes modelos são eficientes e funcionam muito bem quando aplicados de forma isolada.

Os métodos ensemble permitem combinar modelos concorrentes para formar uma espécie de comitê, e tem havido muita pesquisa nesta área com um bom grau de sucesso. Os sistemas de recomendação e os aplicativos de mineração de texto baseados em fluxo, usam amplamente métodos ensemble.

E um princípio amplamente discutido em Machine Learning, é que acurácia e simplicidade do modelo levam ao estado da arte em análise preditiva. Mas conseguir os 2 normalmente é muito difícil, sendo um trade-off: um modelo mais complexo é mais flexível, mas consequentemente mais suscetível ao overfitting e provavelmente não vai generalizar bem em novos conjuntos de dados. Modelos mais simples podem não conseguir realizar o aprendizado de forma efetiva, embora reduzam o tempo total de treinamento.

Técnicas de regularização tem sido utilizadas como forma de se atingir o estado da arte em Machine Learning, aplicando uma função de erro que penaliza a complexidade do modelo. A regularização é apontada como uma das principais razões da performance muito superior de modelos de métodos ensemble.

Tem havido muitos estudos independentes realizados em grupos de aprendizagem supervisionados e não supervisionados. O tema comum observado é que muitos modelos diferentes, quando reunidos, fortaleceram os modelos fracos e geram um melhor desempenho global. Aprendizagem baseada em métodos ensemble é apenas uma das muitas categorias de modelos de aprendizagem.

Ensemble, em geral, significa um grupo de coisas que são geralmente vistos como um todo. Os conjuntos seguem uma abordagem de divisão e conquista usada para melhorar o desempenho.

E temos três categorias principais:

Bagging é usado para construção de múltiplos modelos (normalmente do mesmo tipo) a partir de diferentes subsets no dataset de treino.

Boosting é usado para construção de múltiplos modelos (normalmente do mesmo tipo), onde cada modelo aprende a corrigir os erros gerados pelo modelo anterior, dentro da sequência de modelos criados.

Voting é usado para construção de múltiplos modelos (normalmente de tipos diferentes). Estatísticas simples (como a média) são usadas para combinar as previsões.

Os métodos ensemble são, comprovadamente, poderosos métodos para melhorar a precisão e robustez de soluções supervisionadas, semi-supervisionadas e não supervisionadas. O conceito é simples. A partir do dataset original, geramos diferentes subsets que irão alimentar diferentes modelos individuais. Cada modelo individual, gera uma estimativa que serão combinadas para gerar um resultado final. Mais a frente vamos discutir as técnicas envolvidas em cada uma destas etapas.

Mas é importante ressaltar, que esta regra se aplica aos dados de treino. Uma vez criado o modelo final a partir do método ensemble, normalmente apenas o modelo final é aplicado ao dataset de teste, gerando assim a avaliação final do modelo.