08

Conclusão

Transcrição

[00:00] Nós terminamos esse nosso curso de Machine Learning, mais especificamente voltado pra aprendizado supervisionado. E o que nós vimos foi a infinidade de métodos ou aplicações que existem pra nós resolvermos problemas de aprendizado supervisionado.

[00:17] Então relembrando, o que é aprendizado supervisionado? É aquele tipo de aprendizado que você tem um conjunto de características e uma variável resposta, e você treina em cima desses caras e depois você joga uma variável, um conjunto de características, lembrando do nosso caso de Zootopia, com uma resposta que nós não sabemos qual é, e o modelo tenta dizer isso pra nós.

[00:39] No curso passado aqui da Alura nós vimos isso com o caso do Naive Bayes, e agora, nesse curso, nós vimos a diferença entre o que é um classificador, o que é um método de classificação e o que é um método de regressão. E a existência de infinidades de métodos.

[00:54] Se nós voltarmos aqui, nós começamos lá com regressão linear, eu vou até entrar aqui no meu curso, pra nós vermos de novo. Estou aqui, André Machine Learning, olha só, eu tenho aqui o curso, nós começamos como exploração, demos a entender a ideia do R quadrado, como os dados da correlação linear.

[01:23] Nós vimos a regressão linear pra dados simples, depois regressão linear múltipla e depois vimos como nós prevemos o nosso caso de bilheteria do nosso chefe.

[01:34] Depois nós fomos pra essa ideia de classificação, então tudo isso foi regressão, que a regressão é o número que nós queremos prever e a classificação é mais uma variável categórica, gostou ou não gostou, e nós trabalhamos com isso, com a ideia da regressão logística, comparamos com o Naive Bayes.

[01:53] E depois disso nós fomos pra uma área totalmente diferente, mais não linear, que são as regras de tomada de decisão, que são as nossas árvores. Se nós formos ver aqui, nós tínhamos a árvore de decisão e como que essas árvores de decisão são criadas, como elas são feitas e como elas são comparadas com os modelos que nós já tínhamos aprendido, no caso da regressão linear simples.

[02:17] E nós vimos que elas podem ser usadas, essas regras, que nós vamos crescendo essa árvore, elas podem ser usadas tanto pra regressão como também pra classificação. E nós vimos e falamos, “mas uma árvore isolada ali, sozinha, ela talvez não funcione muito bem”, talvez funcione, mas em geral teria não funcionado, principalmente quando comparada com outros métodos.

[02:37] E nós vimos como nós podemos combinar, foi a última aula, o último conjunto de aulas que nós vimos, que é o que nós chamamos de ensemble. Que nós vamos ter o bagging, nós vamos ter como esse bagging é feito, o random forest. Nós vimos, então nós podemos combinar e mesmo a combinação, na hora que nós formos criar essas árvores, essas combinações são diferentes.

[03:07] E as estratégias que eu uso nos dão uma infinidade de algoritmos diferentes que nós podemos treinar em cima. Nós vimos o bagging, nós vimos o random forest, nós vimos o boosting, adaptive boosting, nós vimos o gradient boosting também, sem entrar muito no detalhe, mas vendo em geral qual é a ideia, qual é o princípio por trás deles. E nós vimos que eles podem ser tão poderosos quanto a regressão linear, a regressão logística, pra um determinado conjunto de dados que nós estivermos trabalhando.

[03:40] A grande lição que nós tiramos aqui é que, novamente, vou repetir o que eu falei em uma das aulas atrás, o Machine Learning, aprendizado de máquina, não é um método canônico pra resolver um único tipo de problema.

[03:54] Cabe a você, um engenheiro de machine learning ou cientista de dados, analisar a fundo quais são os dados que você tem, a distribuição de dados que você tem, o tipo de problema que você quer resolver, e você ter uma infinidade de métodos à sua disposição e cada um deles tende a funcionar pra determinado tipo de problema.

[04:14] E numa situação real, o ideal é você pegar nossos dados de treino e treinar diferentes modelo nesses dados de treino, e separar esses dados de treino em treino e validação. E esses dados serem treinados em dados de treino e validação, e nós validarmos o nosso modelo ideal em cima disso e depois jogarmos em cima dos dados de teste.

[04:33] Nós demos uma roubada nesse curso, porque nós vimos só treino e teste, porque nós não estávamos treinando mais em um caso real, em dados que nós nunca vimos.

[04:43] E a regra de ouro, isso é uma coisa que nós temos que aprender aqui, nós nunca treinamos em dados de teste, sempre treina em treino, parte do treino vira treino e validação, compara os nossos modelos com base nesses dados de validação, vem pro caso real, nós usamos os dados de teste. Dados de teste podem ser, inclusive, nossos dados de treino lá e nós comparamos em cima disso.

[05:03] É isso. Diferentes métodos, diferentes características, todos eles resolvem o mesmo tipo de problema, cada um na sua forma, um pode ser melhor do que o outro, depende da forma como você estrutura, dos parâmetros que você escolhe, da distribuição dos seus dados. O problema, no fundo, é como estão seus dados, é isso que você tem que entender quando você estiver trabalhando com um caso real. Espero que vocês tenham gostado desse curso.