04

Overfitting (Sobreajuste)

Transcrição

Agora abriremos outra base de dados, a "banco70.arff", que tem em torno de 70% dos dados que estávamos utilizando no curso anteriormente.

Novamente vamos a "*Classifier > trees > RandomForest*". Primeiro, vamos rodar esse algoritmo selecionando "*Use training set*", ou seja, comparando a taxa de acerto com os dados vistos durante o treinamento. Assim, teremos 99% de taxa de acertos, pelos dados já serem conhecidos.

Agora, escolheremos o campo "*Supplied test set*", clicaremos em "*Open file*" e abriremos o arquivo do "banco30.arff", que conforme poderíamos prever, terá 30% dos dados que estávamos usando anteriormente. Fazendo o teste novamente, a taxa de acerto será bem menor em comparação aos 99% anteriores, pois esses dados ainda não foram vistos durante o treinamento.

Normalmente os algoritmos pegam os dados de forma aleatório. Pegamos somente a parte final desses dados, então a taxa de acertos com a base de dados que ele não tinha visto durante o treinamento é mais baixa, já que os dados podem estar tendenciosos nesse caso, que não foi aleatório.

Os acertos serão muito maiores se tratando da base de dados já vista. Quando o ajuste é grande demais com relação aos dados já vistos e se torna impossível resolver exemplos novos, dizemos que estamos tendo um *Overfitting*, ou sobre-ajuste no nosso modelo de *Machine Learning*.

Sempre existirá uma discrepância sobre o quanto o classificador acertará quanto aos dados que já viu e os que não viu, sendo os acertos sobre os dados conhecidos maiores. Mas há um problema quando a distância entre esses valores é muito grande.

Como dissemos, a situação apresentada pode estar enviesada por não termos feito a amostragem aleatória dos dados que não foram vistos. Então, podemos utilizar, por exemplo, o "*Cross-validation*" para uma classificação. Se os acertos estiverem em torno dos 60%, enquanto usando o "*Use training set*" ele acertar muito mais, teremos problemas.

Como estamos usando só 70% dos dados, os acertos com "*Cross-validation*" serão mais altos também, havendo uma diferença com relação aos nossos testes anteriores.

Um ponto importante é que podemos visualizar o número de iterações ocorridas com esse algoritmo. Em "*numIterations*" veremos "100". Vamos diminui-la para "50" e realizaremos os testes novamente. Com a base de dados de treino, novamente teremos 99% de acertos e com a base de testes, haverá uma diferença com relação ao resultado anterior, os acertos serão um pouco menores. Mas em alguns casos, quando diminuímos o número de iterações, o *Overfitting* pode diminuir, porque ele não se ajustará tanto aos dados, permitindo certa liberdade, a depender do algoritmo, para a resolução de problemas que ele ainda não viu.