

Treinamento e teste

Transcrição

Quando nós clicamos em "*Start*" no Weka, o que estamos fazendo é treinar um algoritmo para classificação.

Esse algoritmo está sendo treinado a partir da base de dados que passamos para ele, a do banco. A ideia é que após esse treinamento, esse classificador se torne capaz de classificar novos clientes, que ele não viu durante o treinamento. Assim, o ideal é que passemos uma base de dados que não foi vista durante o treinamento para testar se esse classificador está realmente funcionando.

Se clicarmos em "*Test options > Use training set > Start*", será utilizada a **base de dados de treinamento** para verificar se o algoritmo está funcionando. Ou seja, ele testará o classificador treinado com essa base de dados de treinamento, com dados que ele já viu durante o treinamento. Por isso, nesse caso a taxa de acerto aumentará.

Mas se fornecermos uma base de dados externa que não foi vista durante o treinamento, essa taxa de acerto deve diminuir, voltando a um valor de acerto menor ao anterior à base de dados de treinamento. Para inserir a **base de dados de teste**, clicaremos em "*Test options > Supplied test set > Set*". Se não tivermos essa base externa, podemos clicar em "*Percentage split*", e ele dividirá a base de dados que temos em uma parte para treinamento e outra para o teste.

Nossa opção padrão em "*Percentage split*" está com 66% para o treinamento, e os demais 34% serão para o teste. Normalmente, a maior porcentagem ficará para o treinamento. Clicaremos em "*Start*" e veremos que a taxa de acerto diminuirá, porém ele estará acertando em 84% das vezes numa base de dado que ele não viu durante o treinamento, o que é muito importante.

Por fim, a opção final é a de marcar o campo referente ao "*Cross-validation*" ou "*Stratified Cross-validation*" (Validação cruzada estratificada), que terá o número 10. Isso significa que a base de dados será dividida em 10 partes e uma dessas partes será usada para testar a base. O detalhe é que isso não será feito uma única vez, mas durante as 10 vezes uma das partes será usada para teste e as demais para treinamento. Depois, novamente a base será dividida em 10 partes, sendo uma para treinamento e uma para testes, e assim ocorrerá por 10 vezes.

A taxa de acerto mostrada será uma média das taxas de acerto obtidas em cada uma dessas vezes em que a base de dados foi dividida em um teste.