Aprendizado de Máquina

1. O que é uma base de dados desbalanceada?

R: Base de dados desbalanceada é uma base de dados que possui uma distribuição de classes desigual ou seja existem classes que aparecem mais que as outras.

2. Compare as métricas accuracy e f1 em bases de dados binárias indicando as suas vantagens em relação a outra.

R: A métrica accuracy mede a proporção de previsões corretas em relação ao total de previsões feitas pelo modelo. É uma métrica amplamente utilizada porque é fácil de entender e interpretar, mas pode ser enganosa em algumas situações, especialmente quando há classes desbalanceadas ou quando os erros de falsos positivos e falsos negativos têm custos diferentes. Em geral, a accuracy é útil quando as classes têm distribuição equilibrada e quando os erros de classificação têm o mesmo peso.

Já a métrica F1 é uma medida de precisão e recall combinados, que leva em conta tanto os falsos positivos quanto os falsos negativos. O F1 é a média harmônica da precisão e do recall, e é útil quando as classes têm distribuição desequilibrada e quando os erros de classificação têm custos diferentes. Em geral, o F1 é útil quando se deseja ter um equilíbrio entre a precisão e o recall.

Portanto, podemos dizer que a vantagem da accuracy é a sua simplicidade e facilidade de interpretação, especialmente em casos onde os custos dos erros são iguais e as classes são equilibradas. Já a vantagem do F1 é a sua capacidade de lidar com classes desequilibradas e erros com custos diferentes, além de ser uma métrica que considera tanto a precisão quanto o recall.

3. Explique porque há necessidade de usar os conceitos micro, macro e weighted em conjunto com as métricas precision, recall e f1? Porque esses conceitos não se aplicam a accuracy?

R: É necessário usar esses conceitos para se ter uma visão mais ampla, uma vez que essas medidas são calculadas especificamente para cada classe. Para ter uma visão global dos resultados dessas métricas para todas as classes é preciso fazer uma composição dos resultados obtidos em cada classe. Por isso se necessita fazer uma média dos resultados dessas classes. As médias podem ser macro (aritmética), micro (generaliza a fórmula para ser aplicada a todas as classes) e weighted (ponderada em função do número de exemplos de cada classe). Accuracy é uma métrica que leva em conta todas as classes e portanto já oferece uma visão global dos resultados.

Referências complementares

4. Qual a relação entre micro precision, micro recall, micro f1 e accuracy quando o conjunto de teste é perfeitamente balanceado? Essa relação se aplica a macro f1?

R: Independentemente da base ser ou não balanceada a micro precision e a micro recall são iguais a accuracy. Como a micro F1 é a média harmônica desses dois valores, que são iguais a accuracy, seu valor também é igual a accuracy. Quando a base é perfeitamente balanceada, as métricas micro e macro geram sempre os mesmos resultados. Logo, a macro F1 também será igual a accuracy.

Referência complementar (ver Notes)

5. Porque o método de amostragem por resubstituição não é indicado para avaliação de métodos de classificação? Apresente uma utilidade para este método de amostragem.

R: Porque usa o conjunto de treino para teste. Desta forma o classificador pode decorar todos os casos gerando overfitting e tendo um baixo desempenho quando se precisa generalizar o conhecimento. O método é útil apenas para testar a implementação do algoritmo ou para descartar o modelo aprendido. Se o classificador tem baixo desempenho usando resubstituição é certo que seu desempenho não será bom em qualquer outra situação.

6. Apresente as vantagens do método de validação cruzada estratificada em comparação ao método de divisão percentual.

R: Tal como a divisão percentual, o método de validação cruzada estratificada separa devidamente os conjuntos de treino e teste. Mas, em contraste a divisão percentual, possibilita usar todos os dados disponíveis tanto para treino quanto para teste. Outra vantagem é que os resultados não são obtidos por uma única divisão, permitindo lidar com a variabilidade nas escolha dos exemplos de treino e teste.

7. Quando o método de exclusão de 1 deve ser preferido em relação ao método de validação cruzada? Qual a maior dificuldade para aplicação desse método?

R: Quando a quantidade de instâncias for pequena. A maior dificuldade para aplicar esse método é a quantidade de repetições, prejudicando o desempenho em tempo, uma vez que, alguns algoritmos de aprendizado de máquina tem altas complexidades computacionais.

8. Os métodos mais robustos de classificação supervisionada contém vários parâmetros. De que forma pode ser determinado os valores desses hiperparâmetros? Qual é o método mais usado e porque ele é o preferido?

R: Podemos buscar esses parâmetros testando cada um deles, normalmente é utilizada uma busca em grade que faz todas combinações possíveis entre conjuntos de valores desses hiperparâmetros e testa no classificador escolhendo apenas a melhor.

9. Qual a forma mais indicada de se evitar superajuste dos métodos. Explique como ela funciona. Considere o uso do classificador K Vizinhos Mais Próximos com o método indicado e explique porque nesse caso não é possível indicar o valor de K encontrado.

R: É interessante usar a validação cruzada aninhada para evitar super ajustes. Esse método utiliza dois loops de validação cruzada - um interno onde se faz uma busca em grade pelos melhores valores de hiperparâmetro e um externo no qual se obtém a estimativa de desempenho do método. Neste caso não dá para indicar o k encontrado pois eles variam de acordo com o fold de teste. Para aplicar o modelo na prática, deve-se fazer a validação cruzada simples nos valores que obtiveram melhor desempenho em cada fold de teste da validação aninhada e escolher o melhor (seria a melhor aposta, embora o valor obtido não seria generalizável -> para isso deve-se usar o valor obtido na validação cruzada aninhada).

10. Porque é importante usar testes estatísticos na comparação de desempenho entre métodos de classificação? Qual a diferença entre usar um método paramétrico e não paramétrico de testes de hipóteses? Explique porque é necessário ter 30 valores para aplicar o teste T. Indique em qual dos testes é mais fácil mostrar a diferença entre os métodos. Explique porque isso ocorre.

R: É importante para dar um fundamento estatístico a crença na diferença de desempenho entre os métodos. Sem testes estatísticos, há uma chance maior de que a diferença ocorra por conta do acaso na divisão dos exemplos. Métodos paramétricos assumem que os dados obedecem a distribuições estatísticas (normalmente a normal), já o não paramétrico não pressupõe isso.

O Teste t precisa de 30 experimentos, pois na estatística acredita-se que quando se repete um experimento 30 vezes ele se aproximará da distribuição normal.

O teste T, que é paramétrico, é mais específico, por causa de assumir uma distribuição nos dados, que o Wilcoxon, que é não paramétrico, isso faz com que o test t seja mais fácil para

identificar diferenças significativas (ou seja, p-value do wilconxon > p-value do t test em geral).

11. Explique o que é o problema do bias de similaridade apresentado no artigo "An Experimental Methodology to Evaluate Machine Learning Methods for Fault Diagnosis based on Vibration Signals" e apresente o contexto em que ele pode ocorrer.

R: A divisão de um sinal utilizada como técnica no artigo, produz sub sinais que vão para tanto para o conjunto de treino quanto de teste, criando assim um modelo que quando testado já pode ser visto em outros pedaços do sinal testado. Isso implica em criar um modelo superajustado que pode não funcionar em situações reais.

12. Considere a seguinte matriz de confusão:

Classe	A (Preditas)	B (Preditas)	C (Preditas)
A (Verdadeiras)	6	2	0
B (Verdadeiras)	1	4	1
C (Verdadeiras)	1	0	5

Calcule as seguintes métricas:

a) acurácia (accuracy)

b) precisão (precision) da classe B

c) revocação macro (macro recall)

R:

a) acurácia = 15/20 = 0.75

b) precisão(B) = 4/6 = 0.66

c) revocação macro = (6/8 + 4/6 + 5/6)/3 = 0,75

13. Explique como funciona o método de reamostragem de validação cruzada aninhada. Apresente a vantagem desse método em relação ao método de validação cruzada simples.

R: O método de reamostragem de validação cruzada aninhada consiste na realização de dois ciclos aninhados de validação cruzada. O ciclo interno realiza uma busca (normalmente, em grade) para ajustar os valores dos hiperparâmetros do método de classificação. O ciclo externo é usado para medir o desempenho do classificador. O número de folds (e até a própria divisão de folds) do ciclo interno podem ser diferentes (normalmente são) dos do ciclo externo. A vantagem desse método em relação a validação cruzada simples é que reduz a chance de acontecer superajuste e, por conseguinte, uma previsão de desempenho otimista que não seja generalizável.