

MAC425/5739 - Inteligência Artificial

Vítor Kei Taira Tamada - 8516250

Exercício-programa 4 - Relatório

Relatório

(a)

Teste		wine	car	ionosphere
Métrica	maxdepth			
Entropy	N	OK	OK	OK
	1	FAIL	FAIL	OK
	2	OK	OK	FAIL
	4	OK	OK	FAIL
	5	FAIL	FAIL	FAIL
Gini	N	OK	OK	OK
	1	FAIL	FAIL	OK
	2	OK	OK	FAIL
	4	OK	OK	FAIL
	5	FAIL	FAIL	FAIL
Error	N	FAIL	FAIL	FAIL
	1	FAIL	FAIL	FAIL
	2	FAIL	FAIL	FAIL
	4	FAIL	FAIL	FAIL
	5	OK	FAIL	FAIL
	6	FAIL	FAIL	FAIL
	7	FAIL	FAIL	FAIL

$N(t)$ = valor padrão de maxdepth para o teste t

$N(\text{wine}) = 3 \mid N(\text{car}) = 4 \mid N(\text{ionosphere}) = 1$

FAIL significa que a mensagem "Most frequent classifier is better" foi retornada para o teste

Error teve mais entradas para mostrar que o teste wine ter passado foi um caso isolado. Testes com valores ainda maiores foram executados, sem alteração nos resultados

(b) Tanto o critério de entropia (*entropy*) quanto o Gini tem resultados semelhantes. Isso se deve ao fato de ambos medirem algo parecido: enquanto Gini mede a probabilidade de um exemplo aleatório ser classificado corretamente se a classe a ser atribuída for escolhida aleatoriamente, a entropia mede o ganho de informação. Entretanto, há méritos em se utilizar Gini ao invés da entropia: por conta da forma como se calcula cada um deles, a entropia é naturalmente mais lenta devido à operação logarítmica.

A divisão baseada em erro de classificação, por outro lado, mede o erro de classificação de um nó, como o nome diz. Como é possível ver na tabela, este critério tende a errar mais, uma vez que as divisões de nós feitas são encerradas mais rapidamente em comparação com Gini e entropia.

(c) Questões teóricas

Questão 1 - Métricas

R) É importante dado que algumas métricas são mais precisas, enquanto outras são mais custosas de se calcular. Além disso, diferentes algoritmos utilizam diferentes métricas para determinar qual o melhor atributo para dividir o conjunto de dados em um determinado nó.

Questão 2 - Condição de parada

R) Verificar se o conjunto de dados sendo verificados é puro, ou seja, se todas as entradas tem a mesma classificação (*label*), se todas as entradas do conjunto de dados são iguais e se o conjunto de dados é vazio.

Questão 3 - Acurácia

R) Um dos motivos é que por mais que a árvore tenha um `self.maxdepth` maior, os nós folhas são alcançados antes mesmo de chegar nessa profundidade máxima; ou seja, a acurácia não aumenta nem diminui. Nesses casos, a condição de parada não é alcançar a profundidade máxima, mas o conjunto de dados ser puro ou todas as entradas serem iguais.

Por outro lado, se houver muitos atributos, o aumento da profundidade faz com que muitos deles sejam analisados - mais do que o necessário. Isso resulta no problema de *overfitting*, que reduz a acurácia da árvore de classificação.

Questão 4 - Comparação

R) Quando a árvore de decisão formada ou a sua construção for complexa ou custosa demais, ou quando o conjunto de dados de treinamento ainda precisa ser tratado (eliminação de entradas duplicadas ainda é necessária por exemplo). Ou seja, quando não vale a pena construir a árvore de decisão.