Disciplina: ICC204 – Aprendizagem de Máquina e Mineração de Dados 31/05/2019
Prof. Rafael Giusti (rgiusti@icomp.ufam.edu.br)

Lista de Exercícios 4

A tabela abaixo pode ser utilizada para aproximar o valor de $-\log_2(a/b)$. Consulte a linha e a coluna correspondentes. Por exemplo, na linha 3, coluna, tem-se que $\log_2(^3/_5) = -0{,}736$ é aproximadamente $^{-8}/_{11} = -0{,}727$.

	B									
A	1	2	3	4	5	6	7	8	9	10
1	0	1	11/7	2/1	7/3	18/7	14/5	3/1	19/6	10/3
2		0	4/7	1/1	4/3	11/7	9/5	2/1	13/6	7/3
3			0	3/7	8/11	1/1	11/9	17/12	11/7	7/4
4				0	1/3	4/7	4/5	1/1	7/6	4/3
5					0	3/11	1/2	2/3	6/7	1/1
6						0	2/9	5/12	4/7	8/11
7							0	1/5	4/11	1/2
8								0	1/6	1/3
9									0	1/7
10										0

1. Considere a seguinte base de exemplos (# é o número da instância, não um atributo):

#	X1	X2	X3	Classe
1	A	1	V	+
2	A	1	F	+
3	A	2	V	-
4	B	1	F	+
5	B	2	F	+
6	B	1	V	-
7	B	2	V	-
8	B	3	V	-
9	B	1	F	-

   a) Induza uma árvore de decisão utilizando ganho de informação como critério de divisão.

   b) Utilizando o algoritmo de cobertura visto em aula, induza regras de conhecimento.

   c) Considere um *ensemble* de voto majoritário simples no qual os classificadores-base são os dois modelos que você induziu nos itens e um classificador 1-NN. Como esse *ensemble* irá classificar o exemplo <B, 3, F>? Não desenvolva os cálculos das raízes se não houver necessidade.

2. No contexto de *ensembles*, o que é diversidade? De que maneiras podemos introduzir diversidade em um *ensemble*?

3. Considere o seguinte esquema de ensembles. Utilizamos um único conjunto de treinamento D definimos várias $y_1(x)$, $y_2(x)$, $y_3(x)$ etc. tais que cada $y_i(x)$ é um classificador 1-NN utilizando a distância Minkowski com p=i. Isto é, $y_1(x)$ é o classificador 1-NN utilizando distância de Manhattan, $y_2(x)$ é o 1-NN com distância euclidiana e assim por diante. O *ensemble* classificará um novo exemplo como o voto majoritário ponderado de cada $y_i(x)$.

   Considere que a função de distância de Minkowski é $d_{Mink}^{(p)}(x, y) = \sqrt[p]{\Sigma| xi - yi |^p}$.

   a) Esse esquema introduz diversidade ao *ensemble*? Se sim, de qual tipo? Se não, que pequenas alterações poderíamos fazer à estrutura do *ensemble* para obter diversidade?

   b) Em vez de utilizar o mesmo conjunto D para todos os classificadores-base, poderíamos empregar AdaBoost para gerar um *ensemble* de classificadores 1-NN, todos com distância euclidiana? Qual seria o procedimento?

   c) O *ensemble* descrito no item b) teria um conjunto diverso de classificadores-base? Seria possível obter um *ensemble* eficaz (mais eficaz que um único classificador 1-NN induzido sobre o conjunto original D) através desse procedimento? Sob quais condições?

4. O classificador R0 (*zero* rule), também conhecido como classificador majoritário, é aquele que sempre classifica os exemplos com a classe majoritária do conjunto de treinamento. Por exemplo, para os dados do Exercício 1, o R0 seria o modelo y(x) = -, que é a classe majoritária daquele conjunto.

   Utilizando os dados da questão 1, gere um *ensemble* bagging com M=4 classificadores-base, cada um sendo o R0. Use a seguinte sequência de números pseudoaleatórios para efetuar a amostragem dos conjuntos (use aritmética modular caso o número pseudoaleatório seja maior que o número de exemplos): 14, 16, 12, 2, 8, 4, 13, 17, 17, 7, 10, 4, 7, 15, 9, 14, 10, 9, 13, 5, 16, 3, 2, 6, 8, 5, 11, 1, 12, 1, 11, 15, 3, 6.

   Você acredita que esse *ensemble* pode ser adequado? Qual é o erro empírico desse *ensemble*, considerando como referência o conjunto de dados original da questão 1?

5. Bentina gostaria de ficar milionária. Para isso, ela desenvolveu um *ensemble* para classificar análise de risco de crédito. A base de dados de Bentina possui registros de 20 milhões de empréstimos cedidos por um grande operador financeiro no Brasil. Os registros contêm dados sócio-econômicos dos clientes, do empréstimo e do desfecho do empréstimo (dívida aberta, sanada dentro do prazo ou sanada com atraso).

   Que estratégia de *ensemble* você poderia sugerir para Bentina?