Disciplina: ICC204 – Aprendizagem de Máquina e Mineração de Dados 27/05/2019
Prof. Rafael Giusti (rgiusti@icomp.ufam.edu.br)

Lista de Exercícios 1

Resolução

1. Dados um conjunto de exemplos de treinamento $X = \{x_1, x_2, \dots, x_n\}$ e um conjunto de classes $C = \{C_1, C_2, \dots, C_k\}$, um problema de classificação pode ser dividido em duas etapas: fase de inferência e fase de decisão.

Na fase de inferência, os dados de treinamento são utilizados para a definição de probabilidades a posteriori $p(C_k \mid x)$ e, na fase de decisão, os valores de probabilidade a posteriori são utilizados para fazer atribuições de classes para instâncias. Diante desse contexto, essas duas fases podem ser aprendidas por meio de duas abordagens:

(1) separadamente via método bayesiano; ou
(2) conjuntamente por meio de aprendizagem de uma função que mapeia as instâncias de entrada diretamente às classes do problema, ou seja, a decisões.

a) Explique com suas palavras como cada uma dessas duas abordagens funciona.

Na primeira abordagem, primiero resolve-se o problema de determinar as probabilidades condicionais $p(x \mid C_k)$ para cada classe. Além disso, deve-se determinar a probabilidade a priori $p(C_k)$ de cada classe. Com base nesses valores, as probabilidades posteriores $p(C_k \mid x)$ podem ser calculadas. Com base nas probabilidades posteriores, pode-se utilizar teoria de decisão para decidir qual é a classe mais provável para um exemplo.

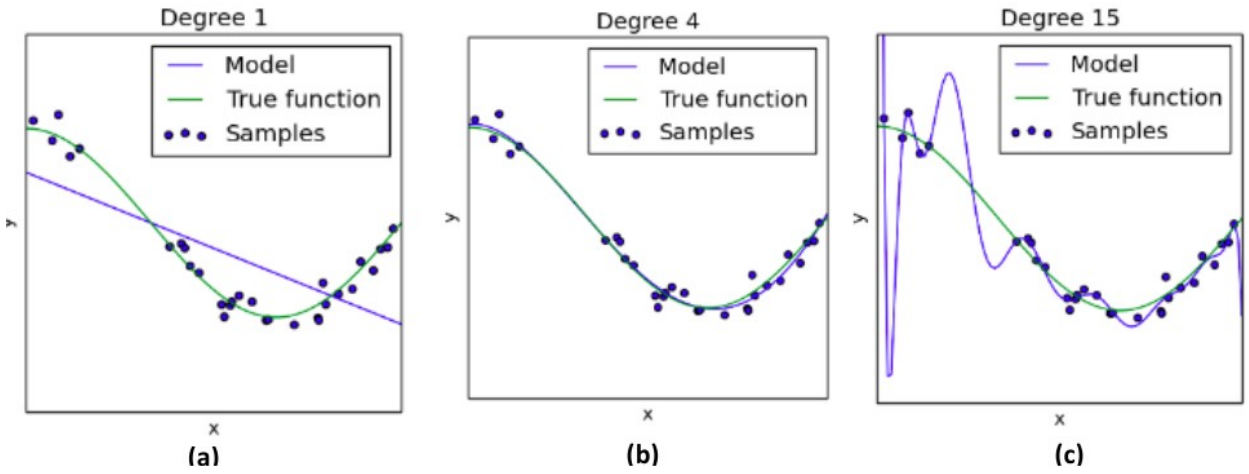
Na segunda abordagem, busca-se uma função $f(x)$, chamada função discriminante, que tenta mapear cada instância x diretamente a uma classe C_k.

b) Descreva vantagens e desvantagens de cada abordagem.

A primeira abordagem tem a vanagem de permitir a detecção de outliers, uma vez que a distribuição dos dados é conhecida. Por outro lado, pode ser necessário uma quantidade excessiva de dados para que as probabilidades sejam devidamente calculadas. Além disso, existe o risco de que a inferência seja feita incorretamente, impondo estruturas inadequadas sobre os dados. Essas estruturas podem ter pouca influência sobre as probabilidades posteriores $p(C_k \mid x)$ ou podem ter influência negativa.

Uma vantagem da segunda abordagem é que não é necessário impor uma estrutura aos dados. Entretanto, o fato de não haver cálculo de probabilidades pode impedir a detecção de novidade nos dados. Além disso, pode ser necessário o treinamento com bases de dados muito grandes para que as características de generalização do problema sejam aprendidas.

Para as questões a seguir, considere o exemplo de aproximação de função ilustrado nas imagens abaixo, as quais foram obtidas ao variar-se o grau do polinômio da função de aproximação.



2. Qual a relação entre o valor do grau do polinômio e os conceitos de *overfitting* e *underfiting*?

   Quanto maior o grau do polinômio, mais complexa será a função de aproximação e, portanto, maior a capacidade do modelo. Dessa forma, um grau muito pequeno fará uma aproximação pobre dos dados, causando *underfitting*. Por outro lado, se o grau for muito elevado, a aproximação será complexa demais, ajustando-se excessivamente aos dados de treinamento e causando *overfitting*.

3. O ajuste do grau do polinômio e a busca pela função de aproximação mais adequada estão relacionados ao teorema *No Free Lunch*?

   Sim, pois o teorema do NFL determina que não existe nenhum algoritmo de AM que é universalmente superior a qualquer outro, de modo que é necessário ajustar corretamente os hiperparâmetros para uma aplicação específica, tomando-se com base os dados do problema.

4. Esse algoritmo tenta minimizar o Erro Empírico? Por quê?

   Sim, pois o algoritmo procura tornar a curva o mais próximo possível dos dados, minimizando o erro médio (ou erro quadrático médio) para os exemplos do conjunto de treinamento.

5. Sobre erro de generalização, responda:

   a) Qual dos três modelos finais apresentará menor erro de generalização? Por quê?

      Assumindo-se que as amostras de teste serão dadas com distribuição próxima à da função real (curva verde ou coincidente nas três figuras), o modelo *b* é claramente superior, visto que aproxima muito melhor a função real do que os outros dois. O modelo *a* é muito pobre e apresenta *overfitting*, enquanto o modelo *c* está superajustado aos dados de treinamento e sofre de *overfitting*.

b) Qual a relação entre erro de generalização e capacidade do modelo?

A capacidade do modelo está relacionada à complexidade do modelo. Se a capacidade for muito baixa, então o modelo será pouco complexo e incapaz de generalizar a função-conceito, portanto sofrerá *underfitting*. Por outro lado, se a capacidade for muito elevada, o modelo será complexo demais e provavelmente será superajustado aos exemplos de treinamento, causando *overfitting*. Em ambos os casos, o erro de generalização será elevado.

c) Qual a relação entre erro de generalização e dilema viés-variância?

O viés de um modelo está relacionado a uma "resistência" em se ajustar aos dados. Se o viés for muito elevado, então a aproximação da função-conceito será pobre, causando *underfitting* e alto erro de generalização. Portanto, durante o treinamento deve-se minimizar o viés do modelo.

Por outro lado, ao minizarmos o viés, tendemos a aumentar sua variância. Se a variância for muito elevada, o modelo se ajustará demasiadamente aos dados de treinamento, causando *overfitting* e alto erro de generalização.

Assim, deve existir um equilíbrio entre o viés e a variância do modelo para que seja possível generalizar bem os dados.

Nos próximos exercícios, considere a seguinte relação:

A1	A2	A3	A4	Classe
F	F	V	F	F
V	V	F	F	F
F	V	V	F	F
V	F	F	F	V
F	V	V	V	F
V	V	F	V	F
F	F	F	V	V

6. Construa uma árvore de decisão, usando o algoritmo ID3, calculando a ganho de informação para cada nó. Inclua todos os passos do cálculo na resposta.

Calculamos a informação do espaço original e, em seguida, a informação média do particionamento promovido por cada atributo. Observe que os valores finais de ganho podem conter erros de arredondamento.

Para a raiz

Informação original: $info([2, 5]) = -(^{2}/_{7})\log(^{2}/_{7}) - (^{5}/_{7})\log(^{5}/_{7}) = 0{,}8631$

Atributo A1:

$info([1, 3], [1, 2]) = (^{4}/_{7})[\ -(^{1}/_{4})\log(^{1}/_{4}) - (^{3}/_{4})\log(^{3}/_{4})\ ] + (^{3}/_{7})[\ -(^{1}/_{3})\log(^{1}/_{3}) - (^{2}/_{3})\log(^{2}/_{3})\ ]$
$= 0{,}8571$

$ganho(A1) = info([2, 5]) - info([1, 3], [1, 2]) = 0{,}006$

Atributo A2:
info([2, 1], [0, 4]) = $(^3/_7)$[$-(^2/_3)\log(^2/_3) - (^1/_3)\log(^1/_3)$ + 0] = 0,3935
ganho(A2) = info([2, 5]) - info([2, 1], [0, 4]) = 0,4695

Atributo A3:
info([2,2], [0, 3]) = $(^4/_7)$[$-(^2/_4)\log(^2/_4) - (^2/_4)\log(^2/_4)$ + 0] = 0,5714
ganho(A3) = info([2, 5]) - info([2,2], [0, 3]) = 0,292

Atributo A4:
info([1, 3], [1, 2]): $(^4/_7)$[$-(^3/_4)\log(^3/_4) - (^1/_4)\log(^1/_4)$] + $(^3/_7)$[$-(^2/_3)\log(^2/_3) - (^1/_3)\log(^1/_3)$]
= 0,8571
ganho(A4) = info([2, 5]) - info([1, 3], [1, 2]) = 0,006

O melhor atributo para a raiz é o atributo A2. Observe que todos os exemplos para os quais A2=V pertencem à classe F. Portanto consideramos agora o sub-espaço para os quais A2=F.

Sub-espaço A2=F

info([2, 1]) = $-(^2/_3)\log(^2/_3) - (^1/_3)\log(^1/_3)$ = 0,9183

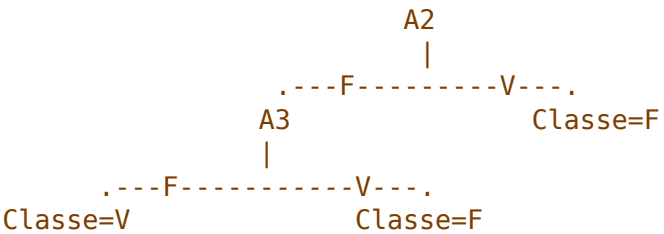
ganho(A1) = info([2, 1]) - info([1, 1], [1, 0])
= [$-(^2/_3)\log(^2/_3) - (^1/_3)\log(^1/_3)$] - $(^2/_3)$[$-(^1/_2)\log(^1/_2) - (^1/_2)\log(^1/_2)$] - 0
= 0,2516

ganho(A3) = info([2, 1]) - info([2, 0], [0, 1]) = info([2, 1]) - 0 = 0,9183

ganho(A4) = info([2, 1]) - info([1, 1], [1, 0]) = ganho(A1) = 0,2516

O melhor atributo para esse sub-espaço é, portanto, A3.

A árvore resultante será:



7. Considerando os atributos A1 a A4, aplique NaiveBayes como um algoritmo de aprendizado probabilístico e crie uma tabela com frequências e probabilidades para a coleção. Use a técnica de suavização de Laplace (ou seja, some 1 a todas as frequências) para evitar probabilidades 0.

P(Classe=F) = 0.714286
P(Classe=V) = 0.285714

P(A1=F | Classe=F) = 4/7 = 0.571429
P(A1=V | Classe=F) = 3/7 = 0.428571
P(A1=F | Classe=V) = 2/4 = 0.5
P(A1=V | Classe=V) = 2/4 = 0.5

P(A2=F | Classe=F) = 2/7 = 0.285714
P(A2=V | Classe=F) = 5/7 = 0.714286
P(A2=F | Classe=V) = 3/4 = 0.75
P(A2=V | Classe=V) = 1/4 = 0.25

P(A3=F | Classe=F) = 3/7 = 0.428571
P(A3=V | Classe=F) = 4/7 = 0.571429
P(A3=F | Classe=V) = 3/4 = 0.75
P(A3=V | Classe=V) = 1/4 = 0.25

P(A4=F | Classe=F) = 4/7 = 0.571429
P(A4=V | Classe=F) = 3/7 = 0.428571
P(A4=F | Classe=V) = 2/4 = 0.5
P(A4=V | Classe=V) = 2/4 = 0.5

8. Como o kNN classificaria o caso de teste t1 = {A1 = V, A2 = V, A3 = V, A4 = V} considerando os atributos A1 a A4 usando 5 vizinhos (k = 5)? Assuma que a distância entre atributos simbólicos é 0 se eles têm os mesmos valores e 1, caso contrário. Use uma métrica de distância Euclidiana e calcule a classe sem ponderação (votação simples).

   Para o exemplo <V, V, V, V, ?>, as distâncias aos protótipos são

Protótipo	Instância	Classe	Distância Euc²	Distância Euc
1	<F, F, V, F>	F	3,00	1,73
2	<V, V, F, F>	F	2,00	1,41
3	<F, V, V, F>	F	2,00	1,41
4	<V, F, F, F>	V	3,00	1,73
5	<F, V, V, V>	F	1,00	1,00
6	<V, V, F, V>	F	1,00	1,00
7	<F, F, V, V>	V	3,00	1,73

   Observe que minimizar o quadrado da distância euclidiana equivale a minimizar a distância euclidiana. Portanto, poderíamos ter calculado apenas a terceira coluna.

   Dos protótipos, quatro vizinhos podem ser definidos sem empates--i.e., os protótipos 5, 6, 2 e 3. O quinto vizinho poderia ser o protótipo 1, o 4 ou o 7. Escolhendo-se o 1 como quinto vizinho, temos 5 votos para a classe F. Escolhendo-se o 4 ou o 7 como quinto vizinho, temos 4 votos para a classe F e 1 voto para a classe V. Em ambos os casos, o exemplo de teste será classificado como F.