Trabalho prático de Sistemas de Apoio à Decisão

Classificadores Baysianos, de k-vizinhos, árvores de decisão, e redes neuronais
Escola Naval

O problema da Companhia de Seguros

Imagine que a companhia de seguros "Sinto&Such Pensórios" contratou a Marinha, dada a nossa inestimável experiência em sistemas de apoio à decisão, para desenvolver uma aplicação informática que permita aos vendedores ter uma estimativa do "valor" de um cliente. Um cliente é tão mais valioso quanto menos probabilidade tiver de se ver envolvido num acidente durante o ano seguinte.

Depois de ter perdido muito tempo a pensar sobre o que é que aumenta ou diminui a probabilidade de uma dada pessoa ter um acidente de automóvel, descobre que a companhia de seguros já tem uma base de dados relativa aos seus clientes, onde para além de uma série de dados quanto às suas características, tem um campo que indica se tiveram ou não um acidente em que a companhia de seguros teve despesas.

Para cada cliente, a companhia tem na sua base de dados, para além da idade, os seguintes campos binários:

m35	1 se o cliente tem mais de 35 anos (0 em caso contrário)
m65	1 se o cliente tem mais de 65 anos
cas	1 se o cliente é casado
tf	1 se o cliente tem filhos
tc5	1 se o cliente tem carta há mais de 5 anos
ts3	1 se o cliente tem seguro nesta companhia há mais de 3 anos
tsr	1 se o cliente tem seguro contra roubo
tst	1 se o cliente tem seguro contra todos os riscos
sm	1 se o cliente é do sexo masculino
tcs	1 se o cliente tem curso superior
est	1 se o cliente é estudante
tcp	1 se o cliente tem casa própria
tmt	1 se o cliente tem múltiplos telemóveis
fum	1 se o cliente fuma
ta	1 se o cliente teve um acidente

Pretende-se que o programa permita ao vendedor introduzir rapidamente as informações que dispõe sobre a pessoa a quem está a tentar vender uma apólice, e que o programa, usando a base de dados da empresa, preveja se essa pessoa vai ou não dar prejuízo e, se possível, qual a probabilidade de isso acontecer.

Numa primeira fase, a companhia de seguros prefere (por razões culturais…) que use apenas classificadores estatísticos "clássicos", nomeadamente classificadores de máxima verosimilhança (ML), *maximum a posterioi* (MAP), bayesianos, ou "naive Bayes", e quer que isso seja feito usando MS-Excel, para perceber ponto por ponto tudo que está a acontecer. Nunca segunda fase, está interessada em usar outros classificadores, implementados em WEKA, nomeadamente k-vizinhos, árvores de decisão, e redes neuronais. Nos casos em que precisa de saber os custos de decisões

erradas, a companhia informa-o que o custo (em lucros perdidos) por não tentar vender uma apólice a uma pessoa que seria um bom cliente é de 500, enquanto o custo de vender uma apólice a uma pessoa que é um mau cliente é de 600.

Para testar a sua capacidade, a companhia de seguros facultou-lhe uma base de dados com 1000 clientes (chamada "seguros"), e outra com 20 (chamada "prova"), onde ocultou o campo "ta".

Questões:

1) Se aparecer um cliente que é casado, deve ou não tentar vender uma apólice ? Para tomar essa decisão use um classificador ML, MAP, e Bayesiano com custos[1]. Um classificador naive de Bayes faz sentido neste caso ?

2) Aparece um cliente que tem claramente mais de 35 anos, mas que usa aliança (é casado), e vem com 2 telemóveis na mão. Deve tentar vender-lhe uma apólice ?

3) Aparece um senhor que preenche a ficha de inscrição, e através dela fica a saber que ele tem 36 anos, é casado, tem filhos, tem a carta há menos de 5 anos, não tem nenhum seguro, tem curso superior e já não estuda, não tem casa própria mas tem múltiplos telemóveis, e não fuma. Decida se lhe deve ou não vender uma apólice, usando:

    a. Um classificador MAP (sem naive Bayes)
    b. Um classificador MAP (com naive Bayes)

4) Se usasse um classificador de k-vizinhos (com k=11), qual seria a taxa de erro esperada ? Qual seria a previsão de um sistema desses para o cliente da questão 3 ?

5) Se usasse uma árvore de decisão do tipo C4.5 (a que o Weka chama J48), qual seria a taxa de erro esperada ? Qual seria a previsão de um sistema desses para o cliente da questão 3 ?

6) Se usasse uma rede neuronal multicamada, qual seria a taxa de erro esperada ? Qual seria a previsão de um sistema desses para o cliente da questão 3 ?

Bom trabalho !

[1] Não se esqueça que, como normalmente não calcula p(x), está a calcular um "score" proporcional a p(C|x) e não p(C|x) propriamente dito. Assim sendo tem que calcular o "score" para "dá prejuízo" e o score para "não dar prejuízo" (pois a soma dos dois não é 1).