Atividade 6 **Yan A. S. Duarte**
PEL 208 Tópicos Especiais em Aprendizagem
Prof. Reinaldo A. C. Bianchi Entrega: 27/11/2017

Introdução

Esse relatório tem como objetivo detalhar a teoria, a implementação, os resultados e a conclusão da sexta atividade do curso. A propósta do exercício é implementar um classificador Naive-Bayes.

Teoria

Classificador Naive Bayes

O classificador Naive Bayes é um classificador probabilístico baseado no teorema de Bayes. Ele recebe o título de *Naive*, ou seja, ingênuo, devido ao fato de que os atributos são condicionalmente independentes. Isso faz com que as informações de um evento não tenha ligação com outro.

Por ser um método simples de classificação, o mesmo possui um bom desempenho e necessita de uma quantidade de dados relativamente pequena para realizar a classificiação com uma precisão satisfatória.

Para melhor compreensão do classificador observemos o exemplo a seguir:

Utilizaremos dados de pessoas que estão sendo diagnosticadas com uma nova doença. Logo depois de observar os dados, de 100 pessoas, 20 tiveram resultado positivo para a doença (20%) e 80 se mantiveram saudáveis (80%), sabendo que das infectadas, 90% tiveram resultados positivos, e 30% de pessoas que eram saudáveis também receberam o resultado positivo.

Listando os dados, temos:

- 100 pessoas fizeram o exame.
- 20% das pessoas que fizeram o exame tiveram resultados positivos.
- 90% das pessoas que estavam doentes, receberam resultados positivos.
- 30% das pessoas que não estavam doentes, receberam resultados positivos.

A pergunta que o algoritmo busca responder é: Se uma nova pessoa realizar o exame e receber um resultado positivo, qual a probabilidade de ela estar doente?

O classificador Naive Bayes busca encontrar a probabilidade a posteriori (estar doente e receber exame com resultado positivo), multiplicando a probabilidade a priori (estar doente) pela probabilidade de *"receber um exame positivo e estar doente"*.

Deve-se anotar também a probabilidade a posteriori da negação (não estar doente e receber exame com resultado positivo). Ou seja:

- $P(doente|positivo) = 20\% * 90\%$
- $P(doente|positivo) = 0,2 * 0,9$
- $P(doente|positivo) = 0,18$
- $P(saudavel|positivo) = 80\% * 30\%$

- $P(saudavel|positivo) = 0,8 * 0,3$
- $P(saudavel|positivo) = 0,24$

Após isso realizamos a normalização dos dados, para que a soma das duas probabilidades resulte 1 (100%).

Para tal, dividimos o resultado pela soma das duas probabilidades. Exemplo:

- $P(doente|positivo) = 0,18/(0,18 + 0,24) = 0,4285$
- $P(saudvel|positivo) = 0,24/(0,18 + 0,24) = 0,5714$
- $0,4285 + 0,5714 \approx 1$.

Matematicamente, podemos definir o teorema de Bayes da seguinte maneira:

$$P(B|A) = \frac{P(B|A)P(A)}{P(B)}$$

Implementação

Para a elaboração do exercício foi utilizada a linguagem de programação C++. Foram criadas duas funções para executar a classificação e mais algumas funções auxiliares para obter os dados de entrada e impirmir na tela os dados de saída.

Função ocorrencias

Essa função retorna uma matriz com o número de ocorrências dos possíveis valores de cada atributo:

```
vector< vector< vector<double> > >
        ocorrencias(vector< vector<double> > entrada, double classe){


    vector< vector< vector<double> > > result;

    for (int j=0; j<entrada[0].size(); j++){

        vector< vector<double> > temp;
        for (int i=0; i<entrada.size(); i++){
            if (entrada[i][0] == classe){

                bool incluir = true;

                for (int k=0; k<temp.size(); k++){
                    if (entrada[i][j]==temp[k][0]){
                        temp[k][1]++;
                        incluir = false;
                    }
                }

                if (incluir == true){
                    vector<double> vetor_inclusao;

                    vetor_inclusao.push_back(entrada[i][j]);
                    vetor_inclusao.push_back(1);
```

```
                    temp.push_back(vetor_inclusao);
                }
            }
        }
        result.push_back(temp);
    }

    return result;
}
```

Função classificacao

Essa função realiza a classificação:

```
double classificacao(
        vector< double > teste,
        vector< vector< vector<double> > > ocorrencia,
        double total_casos){


    double result = ocorrencia[0][0][1] / total_casos;
    for (int i=0; i<teste.size(); i++){
        for (int j=0; j<ocorrencia[i+1].size(); j++){
            if (ocorrencia[i+1][j][0] == teste[i]){
                result *= (ocorrencia[i+1][j][1] / ocorrencia[0][0][1]);
            }
        }
    }

    return result;
}
```

Testes

Os datasets utilizados nessa atividade foram extraídos dos próprios slides. E são apresentados a seguir.

Mau pagador

O primeiro dataset possui algumas informações econômicas de dez pessoas diferentes. Essas informações são utilizadas para treinar o classificador.

Para testar o classificador, foi utilizada a seguinte amostra:

Tabela 1: Dataset Mau pagador

ID	Casa Própria	Estado Civil	Rendimento	Mau Pagador
1	S	Solteiro	Alto	N
2	N	Casado	Médio	N
3	N	Solteiro	Baixo	N
4	S	Casado	Alto	N
5	N	Divorciado	Médio	S
6	N	Casado	Baixo	N
7	S	Divorciado	Alto	N
8	N	Solteiro	Médio	S
9	N	Casado	Baixo	N
10	N	Solteiro	Médio	S

Casa Própria	Estado Civil	Rendimento
N	Divorciado	Médio

Celular

O segundo dataset informa qual o tipo de computador que dez diferentes pessoas possuem. Essas informações foram utilizadas para treinar o classificador.

Tabela 2: Dataset Celular

Nome	Laptop	Celular
Kate	PC	Android
Tom	PC	Android
Harry	PC	Android
Annika	Mac	iPhone
Naomi	Mac	Android
Joe	Mac	iPhone
Chakotay	Mac	iPhone
Neelix	Mac	Android
Kes	PC	iPhone
B'Elanna	Mac	iPhone

Para testar o classificador, foi utilizada a seguinte amostra:

Laptop
Mac

Resultados

Mau pagador

Para o dado de entrada:

Casa Própria	Estado Civil	Rendimento
N	Divorciado	Médio

o resultado obtido a partir do classificador foi: Mau pagador = Verdadeiro

Celular

Para o dado de entrada:

Laptop
Mac

o resultado obtido a partir do classificador foi: Celular = iPhone

Conclusão

O relatório propós a implementação do classificador Naive Bayes em liguagem c++. Para executar o treinamento e os testes do classificador, foram utilizados dois datasets presentes no material da disciplina.

Cada dataset continha dez amostras para treinamento e uma amostra para testes. O resultado da classificação foi satisfatório, demonstrando qual classe a amostra pertencia de acordo com o treinamento executado.

Referências

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer, 2001.*

[2] Naive Bayes classifier Disponível em (https://en.wikipedia.org/wiki/Naive_Bayes_classifier). Acesso em: 25 de nov. de 2017

[3] *Textbook Naive Bayes Classifier Disponível em (http://www.statsoft.com/textbook/naive-bayes-classifier). Acesso em: 24 de nov. de 2017*