Algoritmo

Naive Bayes

Jones Granatyr

Introdução

- Abordagem probabilística (Teorema de Bayes)
- Exemplos
  - Filtros de spam
  - Mineração de emoções
  - Separação de documentos

Base original

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Naive Bayes

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15000 3	>= 15000 <= 35000 4	> 35000 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

Classes / Risco de crédito	Atributos previsores: História do crédito			Atributos previsores: Dívida		Atributos previsores: Garantias		Atributos previsores: Renda anual		
	Quantidade por tipo									
	Boa	Desconhecida	Ruim							
	5	5	4							
Alto	1/6	2/6	3/6							
6/14	6 classes 'Alto' de 14									Total de 14
Moderado	1/3	1/3	1/3							
3/14	3 classes 'Moderado' de 14									
Baixo	3/5	2/5	0							
5/14	5 classes 'Baixo' de 14									

Em 6 dos casos onde a classe risco é 'Alta', apenas um tem histórico de credito 'Boa'

E assim com cada tipo de cada atributo previsor por cada tipo de classe

História do crédito	Risco
Ruim	Alto
Desconhecida	Alto
Desconhecida	Moderado
Desconhecida	Alto
Desconhecida	Baixo
Desconhecida	Baixo
Ruim	Alto
Ruim	Moderado
Boa	Baixo
Boa	Baixo
Boa	Alto
Boa	Moderado
Boa	Baixo
Ruim	Alto

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7					
Alto 6/14	1/6	2/6	3/6	4/6	2/6					
Moderado 3/14	1/3	1/3	1/3	1/3	2/3					
Baixo 5/14	3/5	2/5	0	2/5	3/5					

Dívida	Risco
Alta	Alto
Alta	Alto
Baixa	Moderado
Baixa	Alto
Baixa	Baixo
Baixa	Baixo
Baixa	Alto
Baixa	Moderado
Baixa	Baixo
Alta	Baixo
Alta	Alto
Alta	Moderado
Alta	Baixo
Alta	Alto

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3			
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0			
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3			
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5			

Garantias	Risco
Nenhuma	Alto
Nenhuma	Alto
Nenhuma	Moderado
Nenhuma	Alto
Nenhuma	Baixo
Adequada	Baixo
Nenhuma	Alto
Adequada	Moderado
Nenhuma	Baixo
Adequada	Baixo
Nenhuma	Alto
Nenhuma	Moderado
Nenhuma	Baixo
Nenhuma	Alto

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15 3	>= 15 <= 35 4	> 35 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

Renda anual	Risco
< 15.000	Alto
>= 15.000 a <= 35.000	Alto
>= 15.000 a <= 35.000	Moderado
> 35.000	Alto
> 35.000	Baixo
> 35.000	Baixo
< 15.000	Alto
> 35.000	Moderado
> 35.000	Baixo
> 35.000	Baixo
< 15.000	Alto
>= 15.000 a <= 35.000	Moderado
> 35.0000	Baixo
>= 15.000 a <= 35.000	Alto

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa	Desconhecida	Ruim	Alta	Baixa	Nenhuma	Adequada	< 15	>= 15 <= 35	> 35
	5	5	4	7	7	11	3	3	4	7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

Dados do novo cliente

História = Boa

Dívida = Alta

Garantias = Nenhuma

Renda = > 35

Soma: 0,0079 + 0,0052 + 0,0514 = **0,0645**

Somamos o resultado dos riscos (classes), considerando eles 100%, e depois com isso, para cada risco, dividimos pleo todo e multiplicamos por 100 para termos a %

P = Risco

Ou seja, ao multiplicar os risco de acordo com seus dados, vamos ter uma % para cada tipo (alto, moderado, baixo)

P(Alto) = 6/14 * 1/6 * 4/6 * 6/6 * 1/6

P(Alto) = 0,0079

P(Alto) = 0,0079 / 0,0645 * 100 = **12,24%**

P(Moderado) = 3/14 * 1/3 * 1/3 * 2/3 * 1/3

P(Moderado) = 0,0052

P(Moderado) = 0,0052 / 0,0645 * 100 = **8,06%**

P(Baixo) = 5/14 * 3/5 * 2/5 * 3/5 * 5/5

P(Baixo) = 0,0514

P(Baixo) = 0,0514 / 0,0645 * 100 = **79,68%**

Risco de crédito	História do crédito			Dívida		Garantias		Renda anual		
	Boa 5	Desconhecida 5	Ruim 4	Alta 7	Baixa 7	Nenhuma 11	Adequada 3	< 15 3	>= 15 <= 35 4	> 35 7
Alto 6/14	1/6	2/6	3/6	4/6	2/6	6/6	0	3/6	2/6	1/6
Moderado 3/14	1/3	1/3	1/3	1/3	2/3	2/3	1/3	0	2/3	1/3
Baixo 5/14	3/5	2/5	0	2/5	3/5	3/5	2/5	0	0	5/5

História = Ruim
Dívida = Alta
Garantias = Adequada
Renda = < 15

Correção Laplaciana

A correção laplaciana, seria como 'adicionar um novo registro' de 1/total+1 no atributo previsor que for 0, por exemplo

No ruim/baixo temos um 0, e então adicionamos um registro onde substituiria 1/5+1, e todos os outros atributos dependentes dessas informações aumentariam mais 1 também, para cada alteração de algum 0, ou seja, o Baixo ficaria 6/15, boa/baixo ficaria 3/6, desconhecida/baixo ficaria 2/6, e o Ruim ficaria 5...

P(Alto) = 6/14 * 3/6 * 4/6 * 0 * 3/6

P(Moderado) = 3/14 * 1/3 * 1/3 * 1/3 * 0

P(Baixo) = 5/14 * 0 * 2/5 * 2/5 * 0

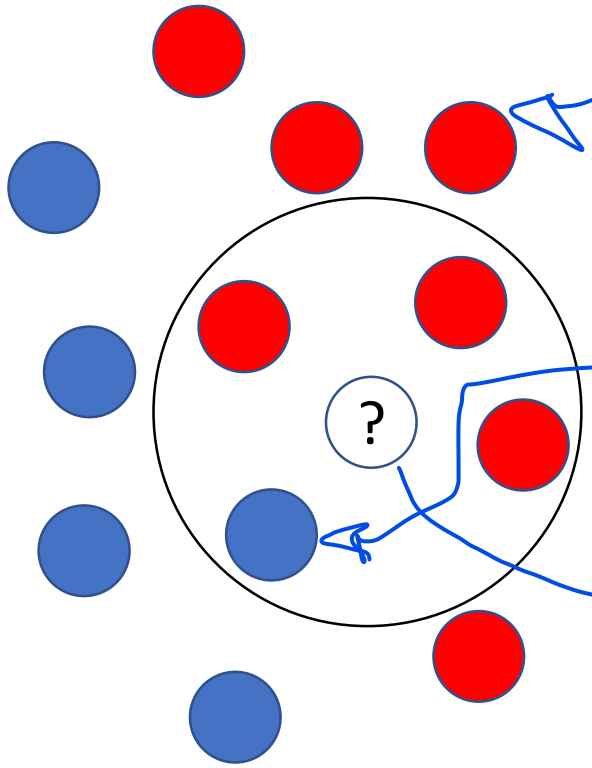
P(vermelho) = 7 / 12

Classe Vermelha

P(azul) = 5 / 12

Classe Azul

Probabilidades apriori



P’(vermelho) = 3 / 7

P’(azul) = 1 / 5

DESCOBRIR

Probabilidades posteori

P’’(vermelho) = 7 / 12 * 3 / 7 = 21 / 84 = **0,25**

P’’(azul) = 5 / 12 * 1 / 5 = 5 / 60 = **0,08**

As Chances da bolinha que queremos descobrir ser vemelha ou azul
Quanto maior o resultado, mais chance

Vantagens x desvantagens

- Vantagens
  - Rápido
  - Simplicidade de interpretação
  - Trabalha com altas dimensões
  - Boas previsões em bases pequenas → 300++

- Desvantagem
  - Combinação de características (atributos independentes) – cada par de características são independentes – nem sempre é verdade

Conclusão