Aprendizado Probabilístico:

NAÏVE BAYES

Aprendizado Probabilístico

- Fundamenta-se na teoria das probabilidades
- Opera calculando as probabilidades para as hipóteses induzidas
- Aprendizado bayesiano
  - Maior representante deste paradigma
  - Simplicidade, elegância e, mais importante, bons resultados

Características

- Cada exemplo de treinamento pode incrementar ou decrementar a probabilidade de uma hipótese
  - Mais flexibilidade na classificação
- Conhecimento a priori pode ser combinado com os dados observados para determinar a probabilidade final de uma hipótese
- Pode-se acomodar hipóteses que fazem predições probabilísticas
  - Por exemplo, o paciente tem 93% de chance de cura

Um pouco de probabilidade ...

- Um evento é um conjunto (ou subconjunto) de possibilidades que tem uma probabilidade associada.
  - Por exemplo, quando jogamos uma moeda temos um de dois eventos possíveis: *cara* ou *coroa*.
- Outro aspecto importante é a relação entre dois eventos.
  - Por exemplo, a dependência entre chuva e a formação de nuvens.
    - A chuva somente ocorre se há a formação de nuvens.

Um pouco de probabilidade ...

- As relações entre dois eventos podem ser:
  - Disjuntos (ou exclusivos): se um não pode acontecer ao mesmo tempo que o outro.
  - Independentes : podem ocorrer ao mesmo tempo, mas a ocorrência de um não afeta a possibilidade de ocorrência do outro.
  - Dependentes: se a ocorrência de um afeta o outro.

Um pouco de probabilidade ...

- Se os eventos são independentes, a probabilidade de ocorrerem é dada por:

$$P(A \cap B) = P(A) * P(B)$$

  - Se temos 2 hd's, cada um tem probabilidade de falhar no próximo ano igual a 0,3 (30%). A probabilidade dos dois falharem no próximo ano é 0,09 (9%).
- Infelizmente as coisas não são tão simples quando os eventos são dependentes. É aqui que entra o Teorema de Bayes.

Teorema de Bayes

- Para dois evento A e B:

$$P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

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

- Se eventos $A_1 ... A_n$ são mutuamente exclusivos e suas probabilidades somam 1, então:

$$P(B) = \sum_{i=1}^{n} P(B|A) * P(A)$$

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

Classificar os seguintes valores:
X = (Idade <= 30, Renda = Media, Estudante = sim, Crédito = bom)
Y = Compra_Computador?

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

P(Y=sim) e P(Y=não)
Probabilidades:

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

P(Y=sim) e P(Y=não)
Probabilidades: P(Y=sim) = 9/14 = 0,643

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

P(Y=sim) e P(Y=não)
Probabilidades: P(Y=sim) = 9/14 = 0,643
P(Y=não) = 5/14 = 0,357 = 1-P(Y=sim)

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (**Idade <= 30**, Renda = Media, Estudante = sim, Crédito = bom)
Probabilidades:

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (**Idade <= 30**, Renda = Media, Estudante = sim, Crédito = bom)
Probabilidades: P[Idade <= 30 | Y = sim] = 2/9 = 0,222

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (**Idade <= 30**, Renda = Media, Estudante = sim, Crédito = bom)
Probabilidades: P[Idade <= 30 | Y = sim] = 2/9 = 0,222
P[Idade <= 30 | Y = não] = 3/5 = 0,6

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, **Renda = Media**, Estudante = sim, Crédito = bom)
Probabilidades:

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, **Renda = Media**, Estudante = sim, Crédito = bom)
Probabilidades: P[Renda = Media | Y = sim] = 4/9 = 0,444

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, **Renda = Media**, Estudante = sim, Crédito = bom)
Probabilidades: P[Renda = Media | Y = sim] = 4/9 = 0,444
P[Renda = Media | Y = não] = 2/5 = 0,4

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, **Estudante = sim**, Crédito = bom)
Probabilidades:

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, **Estudante = sim**, Crédito = bom)
Probabilidades: P[Estudante = sim | Y = sim] = 6/9 = 0,667

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, **Estudante = sim**, Crédito = bom)
Probabilidades: P[Estudante = sim | Y = sim] = 6/9 = 0,667
P[Estudante = sim | Y = não] =1/5 = 0,2

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, Estudante = sim, **Crédito = bom**)
Probabilidades:

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, Estudante = sim, **Crédito = bom**)
Probabilidades: P[Credito = bom | Y = sim] = 6/9 = 0,667

Teorema de Bayes na Classificação

ID	Idade	Renda	Estudante	Crédito	Compra computador
1	<= 30	Alta	Não	Bom	Não
2	<= 30	Alta	Não	Bom	Não
3	31..40	Alta	Não	Bom	Sim
4	> 40	Média	Não	Bom	Sim
5	> 40	Baixa	Sim	Bom	Sim
6	> 40	Baixa	Sim	Excelente	Não
7	31..40	Baixa	Sim	Excelente	Sim
8	<= 30	Média	Não	Bom	Não
9	<= 30	Baixa	Sim	Bom	Sim
10	> 40	Média	Sim	Bom	Sim
11	<= 30	Média	Sim	Excelente	Sim
12	31..40	Média	Não	Excelente	Sim
13	31..40	Alta	Sim	Bom	Sim
14	> 40	Média	Não	Excelente	Não

X = (Idade <= 30, Renda = Media, Estudante = sim, **Crédito = bom**)
Probabilidades: P[Credito = bom | Y = sim] = 6/9 = 0,667
P[Credito = bom | Y = não] = 3/5 = 0,6

Teorema de Bayes na Classificação

- Calculamos isoladamente o valor da probabilidade condicional de cada atributo, mas para que eles sejam calculado de forma interseccionada, temos:

$$P(x_1, x_2, \dots, x_d|C) = P(x_1|C) * P(x_2|C) * \cdots * P(x_d|C)$$

- Com isso, é possível chegar a uma forma mais geral do Teorema de Bayes:

$$P(Y|X) = \frac{P(Y)\prod_{i=1}^{d} P(X_i|Y)}{P(X)}$$

Teorema de Bayes na Classificação

- Assim temos que:

P(Idade $\leq$ 30, Renda = Media, Estudante = sim, Crédito = bom | Y=sim) =
P(Idade $\leq$ 30 | Y = sim) * P(Renda = Media | Y = sim) *
P(Estudante = sim | Y = sim) * P(Crédito = bom | Y = sim) =
0,222 * 0,444 * 0,667 * 0,667 = 0,0438 $\therefore$ P(X|Y=*sim*) = **0,0438**

P(Idade $\leq$ 30, Renda = Media, Estudante = sim, Crédito = bom | Y=não) =
P(Idade $\leq$ 30 | Y = não) * P(Renda = Media | Y = não) *
P(Estudante = sim | Y = não) * P(Crédito = bom | Y=não) =
0,6 * 0,4 * 0,2 * 0,6 = 0,0288 $\therefore$ P(X|Y=*não*) = **0,0288**

Teorema de Bayes na Classificação

- Como calcular a probabilidade de alguém com o perfil X?
  - Pela lei da probabilidade total!

... ou seja, neste caso teríamos que:

P(X) = P(X|Y=*sim*) * P(Y=*sim*) + P(X|Y=*não*) * P(Y=*não*)

P(X) = 0,0438 * 0,643 + 0,0288 * 0,357 = **0,0384**

Teorema de Bayes na Classificação

- Aplicando o teorema de Bayes, calculamos então P(Y=*sim*|X) e P(Y=*não*|X)

P(Y=*sim*|X) = P(X|Y=*sim*) * P(Y=*sim*) / P(X)
= 0,0438 * 0,643 / 0,0384 = **0,7334**
P(Y=*não*|X) = P(X|Y=*não*) * P(Y=*não*) / P(X)
= 0,0288 * 0,357 / 0,0384 = **0,2676**

- Ou seja, como a probabilidade P(Y=*sim*|X) > P(Y=*não*|X) o classificador Bayesiano prediz que a tupla: X = (Idade $\leq$ 30, Renda = *Media*, Estudante = *sim*, Crédito = *bom*) é classificada na classe: **Compra-Computador = sim**

Classificador Naïve Bayes

- Um classificador Naïve Bayes estima a probabilidade de classe condicional P(X|Y) a partir de uma determinada amostra de dados predizendo a classe mais provável.

- Pré-considerações:
  - Assume-se que os atributos são condicionalmente independentes (*Naïve* – ingênuo ou simples);
  - As probabilidades condicionais são estimadas para os atributos de acordo com a sua classificação:
    - Discretos (categórico);
    - Contínuo.

Atributos Condicionalmente Independentes

- Atributos que apresentam independência estatística entre si:
  - Dois eventos são estatisticamente independentes se a probabilidade da ocorrência de um evento não é afetada pela ocorrência do outro evento.
- Exemplo:
  - Tamanho do braço x Habilidades de Leitura
    - Considerando a Idade, a dependência não ocorre.

Atributos Discretos ou Categóricos

- É aquele atributo para o qual é possível estabelecer um conjunto de valores finito.
- Exemplo:
  - Sexo: {Masculino, Feminino}
  - Cor da Pele: {Branca, Marrom, Amarela, Preta}

Atributos Contínuos

- São considerados contínuos os atributos que possuem muitos ou infinitos valores possíveis
  - Exemplo:
    - Idade: pertence aos Naturais (N)
    - Peso: pertence aos Reais (R)
- Existem duas formas de estimar a probabilidade de classe condicional para atributos contínuos:
  - Discretização dos atributos;
  - Distribuição Gaussiana.

Atributos Contínuos

- Discretização de atributos contínuos:
  - Os atributos contínuos são divididos em intervalos discretos, que substituem os valores desses atributos.
  - Esta abordagem transforma os atributos contínuos em atributos ordinais.
- A transformação dos atributos contínuos em atributos discretos permite que sejam tratados como atributos discretos.

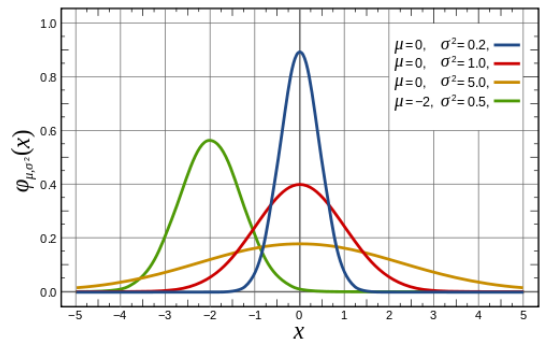
Atributos Contínuos

- Distribuição Gaussiana:
  - Assume uma certa forma de distribuição de probabilidade para variáveis contínuas, e estima os parâmetros da distribuição usando os dados de treinamento.

$$P(X = x|Y = y) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Caracterizada por dois parâmetros:
  - Média (μ): $\mu = \frac{\sum y}{n}$
  - Variância (σ^2) da amostra:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{(n-1)}$$



Classificador Naïve Bayes

- Suponha o seguinte conjunto de treinamento:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Classificador Naïve Bayes

- Cálculo dos atributos discretos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

P(CasaProp=*Sim*|Inad=*Não*) = 3/7
P(CasaProp=*Não*|Inad=*Não*) = 4/7
P(CasaProp=*Sim*|Inad=*Sim*) = 0
P(CasaProp=*Não*|Inad=*Sim*) = 1

Classificador Naïve Bayes

- Cálculo dos atributos discretos:

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

P(CasaProp=*Sim*|Inad=*Não*) = 3/7
P(CasaProp=*Não*|Inad=*Não*) = 4/7
P(CasaProp=*Sim*|Inad=*Sim*) = 0
P(CasaProp=*Não*|Inad=*Sim*) = 1

P(EstCivil=*Solteir*|Inad=*Não*) = 2/7
P(EstCivil=*Divor*|Inad=*Não*) = 1/7
P(EstCivil=*Casad*|Inad=*Não*) = 4/7
P(EstCivil=*Solteir*|Inad=*Sim*) = 2/3
P(EstCivil=*Divor*|Inad=*Sim*) = 1/3
P(EstCivil=*Casad*|Inad=*Sim*) = 0

Classificador Naïve Bayes

- O atributo Renda Anual é **contínuo**

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Para a classe: não

- Média:

  $\mu = (125 + 100 + 70 + 120 + 60 + 220 + 75) / 7 =$ **110**

- Variância:

  $\sigma^2 = (125-110)^2 + (100-110)^2 + (70-110)^2 + (120-110)^2 + (60-110)^2 + (220-110)^2 + (75-110)^2 / 6 =$ **2975**

Classificador Naïve Bayes

- O atributo Renda Anual é **contínuo**

Registro	Casa Própria	Estado Civil	Renda Anual	Inadimplente
1	Sim	Solteiro	125K	Não
2	Não	Casado	100K	Não
3	Não	Solteiro	70K	Não
4	Sim	Casado	120K	Não
5	Não	Divorciado	95K	Sim
6	Não	Casado	60K	Não
7	Sim	Divorciado	220K	Não
8	Não	Solteiro	85K	Sim
9	Não	Casado	75K	Não
10	Não	Solteiro	90K	Sim

Para a classe: sim

- Média:

  $\mu = (95 + 85 + 90) / 3 =$ **90**

- Variância:

  $\sigma^2 = (95-90)^2 + (85-90)^2 + (90-90)^2 / 2 =$ **25**

Classificador Naïve Bayes

- Dado o conjunto de treinamento anterior, qual seria a classificação para um indivíduo com o seguinte registro:
  - Casa Própria=*Não*, Estado Civil=*Casado* e Renda Anual=*120K*
- Devemos avaliar qual a maior probabilidade entre as probabilidades posteriores, ou seja, calcular:

  P(Inadimplente=*Não*|X); e

  P(Inadimplente=*Sim*|X)

Classificador Naïve Bayes

- Para calcular as probabilidades posteriores de nosso interesse, isto é, os valores para P(Inadimplente=*Não*|X) e P(Inadimplente=*Sim*|X) é necessário:
  - Calcular as classes condicionais:

    P(X|Inadimplente=*Não*) e P(X|Inadimplente=*Sim*)
- P(X|Inadimplente=*Não*):

  P(CasaProp=*Não*|Inad=*Não*) * P(EstCivil=*Casad*|Inad=*Não*) * P(Renda Anual=*120K*|Inad=*Não*) = 4/7 * 4/7 * 0,0072 = **0,0024**
- P(X|Inadimplente=*Sim*):

  P(CasaProp=*Não*|Inad=*Sim*) * P(EstCivil=*Casad*|Inad=*Sim*) * P(Renda Anual=*120K*|Inad=*Sim*) = 1 * 0 * $1{,}2 \times 10^{-9}$ = **0**

Classificador Naïve Bayes

- Para calcular as probabilidades posteriores de nosso interesse, isto é, os valores para P(Inadimplente=*Não*|X) e P(Inadimplente=*Sim*|X) é necessário:
  - Calcular P(X|In

$$P(Renda = 120|Y = Não) = \frac{1}{\sqrt{2\pi 2975}} exp\left(-\frac{(120-110)^2}{2*2975}\right)$$

- P(X|Inadimplente=*Não*):
  P(CasaProp=*Não*|Inad=*Não*) * P(EstCivil=*Casad*|Inad=*Não*) * P(Renda Anual=*120K*|Inad=*Não*) = 4/7 * 4/7 * 0,0072 = **0,0024**
- P(X|Inadimplente=*Sim*):
  P(CasaProp=*Não*|Inad=*Sim*) * P(EstCivil=*Casad*|Inad=*Sim*) * P(Renda Anual=*120K*|Inad=*Sim*) = 1 * 0 * 1,2x10^{-9} = **0**

Classificador Naïve Bayes

- Para calcular as probabilidades posteriores de nosso interesse, isto é, os valores para P(Inadimplente=*Não*|X) e P(Inadimplente=*Sim*|X) é necessário:
  - Calcular P(X|In

$$P(Renda = 120|Y = Não) = \frac{1}{\sqrt{2\pi 2975}} exp\left(-\frac{(120-110)^2}{2*2975}\right)$$

$$P(Renda = 120|Y = Sim) = \frac{1}{\sqrt{2\pi 25}} exp\left(-\frac{(120-90)^2}{2*25}\right)$$

- P(X|Inadimplente=*Não*):
  P(CasaProp=*Não*|Inad=*Não*) * P(EstCivil=*Casad*|Inad=*Não*) * P(Renda Anual=*120K*|Inad=*Não*) = 4/7 * 4/7 * 0,0072 = **0,0024**
- P(X|Inadimplente=*Sim*):
  P(CasaProp=*Não*|Inad=*Sim*) * P(EstCivil=*Casad*|Inad=*Sim*) * P(Renda Anual=*120K*|Inad=*Sim*) = 1 * 0 * 1,2x10^{-9} = **0**

Classificador Naïve Bayes

- Juntando os termos:
- P(Inadimplente=*Não*|X) = α * P(Não) * P(X|Não)
  = α * 7/10 * 0,0024
  = 0,0016α

  ... onde α = 1/P(X) → Termo constante!

- Como a probabilidade condicional de P(X|Inadimplente=*Sim*) é zero, P(Inadimplente=*Sim*|X) também é zero 0

- Logo, P(Não|X) > P(Sim|X), o registro X é classificado como **Não**

CLASSIFICAÇÃO DE TEXTOS

Aplicações de Classificação de Textos

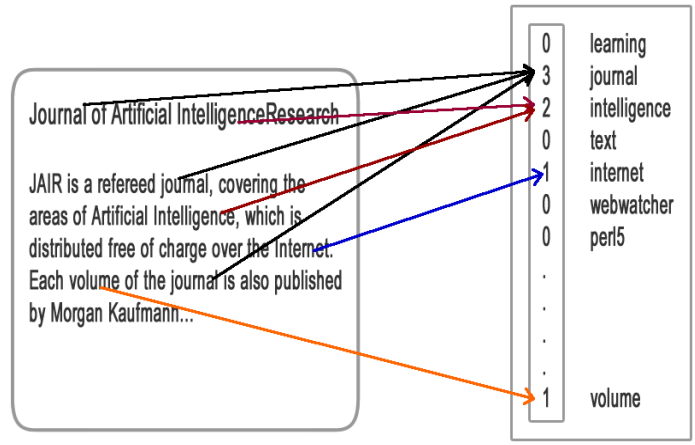
- Páginas web
  - Recomendação
  - Classificação em tópicos (ex.: hierarquia do Yahoo)
- Mensagens de fóruns/blogs
  - Recomendação
  - Filtragem de spam
  - Análise de sentimentos (em relação a produtos)
- Mensagens de e-mail
  - Colocação de anúncios (Gmail)
  - Separação em pastas
  - Filtragem de spam
  - Priorização
- Artigos de jornal
  - Personalização

Probabilidade Condicional

- Suponha a seguinte situação hipotética:
  - Dentre 74 e-mails recebidos, 30 eram spam
  - Dentre 74 e-mails recebidos, 54 continham a palavra “promoção”
  - 20 e-mails contendo a palavra “promoção” foram marcados como spam.
- Eis a questão: qual a probabilidade de que o último e-mail recebido seja spam, dado que ele contém a palavra “promoção”?

Representação de Textos

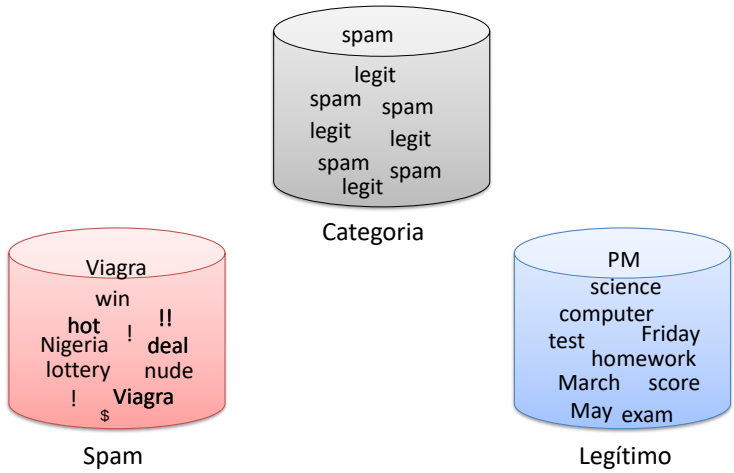
- Modelo mais comum é o *Bag-of-Words*
  - A ordem em que as palavras aparecem é desconsiderada
  - Um atributo por palavra, podendo ser
    - Booleano = indica a presença da palavra
    - Numérico = indica a frequência
  - Palavras sem significado (chamadas de *stopwords*) são removidas.
    - Ex.: artigos, pronomes



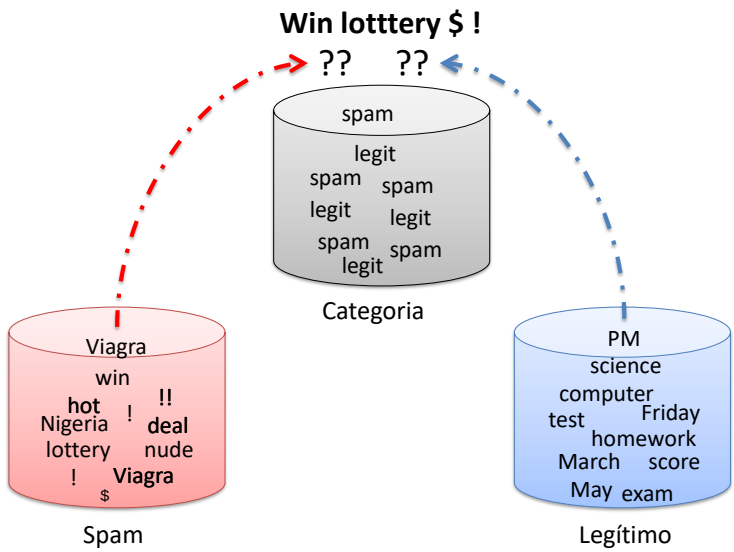
Métodos de Classificação de Textos

- Representações de texto tem alta dimensão.
  - Um atributo por palavra.
- Vetores são esparsos porque muitas palavras são raras.
  - Lei de Zipf
- Algoritmos com alto viés que previnem super-ajuste em altas dimensões são os melhores.
- Para a maioria dos problemas de classificação de textos, há muitos atributos relevantes.
- Métodos que somam evidências de muitos atributos (como Naïve Bayes, kNN, Rede neural, SVM) funcionam melhor do que os que isolam alguns atributos relevantes (árvore de decisão ou indução de regras).

Modelo Naïve Bayes para textos



Modelo Naïve Bayes para textos



Naive Bayes para Textos: Treinamento

Seja *D* um conjunto de documentos
Seja *V* o vocabulário de todas as palavras nos documentos de *D*
Para cada classe *ci* $\in$ *C*
  Seja *Di* o subconjuntos de documentos em *D* que pertencem à categoria *ci*
  P(*ci*) = |*Di* | / |*D*|
 Seja *Ti* a concatenação de todos os documentos em *Di*
  Seja *ni* o número total de ocorrências de palavras em *Ti*
  Para cada palavra *wj* $\in$ *V*
   Seja *nij* o número de ocorrências de *wj* em *Ti*
    Let P(*wij* | *ci*) = (*nij* + 1) / (*ni* + |*V*|)

Naive Bayes para Textos: Teste

- Dado um documento de teste X
- Seja *n* o número de ocorrências de palavras em X, retorne a classe:

$$argmax_{c_i \in C}\ P(c_i) \prod_{i=1}^{n} P(a_i|c_i)$$

... onde a_i é a palavra que ocorre na *i*-ésima posição de X.

Prevenção de *Underflow*

- Multiplicar muitas probabilidades, que estão entre 0 e 1, pode resultar num *underflow* de ponto flutuante.
- Como $\log(xy) = \log(x) + \log(y)$, é melhor fazer todos os cálculos somando logs de probabilidades ao invés de multiplicar probabilidades.
- Classe com maior valor de log-probabilidade é também a mais provável na escala normal.

Métricas de Similaridade de Texto

- Medir a similaridade de textos é um problema bastante estudado.
  - Métricas são baseadas no modelo "*bag of words*".
- Normalmente é feito um pré-processamento: "*stop words*" são removidas e as palavras são reduzidas à sua raiz morfológica.
- Modelo vetorial de Recuperação de Informação (IR) é a abordagem padrão.

O modelo vetorial

- Supõe-se que *t* termos distintos restam após o pré-processamento; chamados de termos do vocabulário.
- Estes termos "ortogonais" formam um espaço vetorial.

  Dimensão = *t* = |vocabulário|
- Cada termo, *i*, num documento ou consulta, *j*, tem um peso dado por um número real, $w_{ij.}$
- Tanto documentos quando consultas são representados por vetores *t*-dimensionais:

$$d_j = (w_{1j}, w_{2j}, \ldots, w_{tj})$$

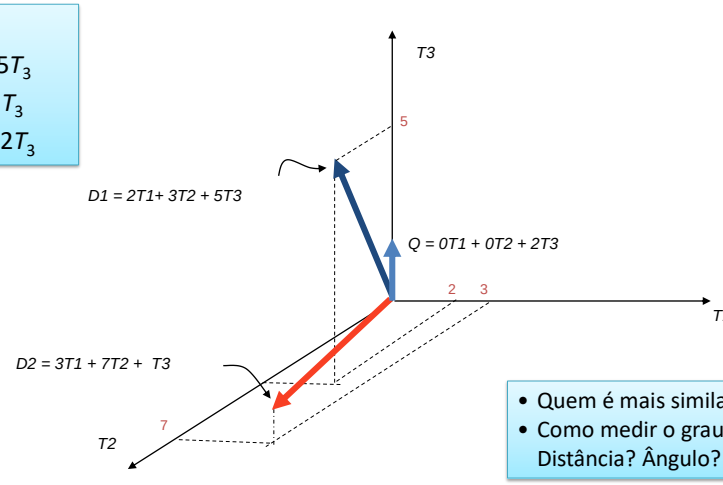
Representação gráfica

Exemplo:

$D_1 = 2T_1 + 3T_2 + 5T_3$

$D_2 = 3T_1 + 7T_2 + T_3$

$Q = 0T_1 + 0T_2 + 2T_3$



- Quem é mais similar a Q? *D1* ou *D2*?
- Como medir o grau de similaridade? Distância? Ângulo? Projeção?

Coleção de Documentos

- Uma coleção de *n* documentos pode ser representada no modelo vetorial por uma matriz.
- Uma entrada na matriz corresponde ao "peso" do termo no documento; zero indica que o termo não é significativo no documento ou simplesmente não existe no documento.

	T_1	T_2	….	T_t
D_1	w_{11}	w_{21}	…	w_{t1}
D_2	w_{12}	w_{22}	…	w_{t2}
:	:	:		:
:	:	:		:
D_n	w_{1n}	w_{2n}	…	w_{tn}

Pesos: Frequência dos Termos

- Termos frequentes em um documento são mais importantes, i.e. mais indicativos do tópico do documento.

  f_{ij} = frequência do termo *i* no documento *j*

- Podemos obter a *frequência do termo* (*tf*) dividindo *f* pela frequência do termo mais comum no documento:

  $tf_{ij} = f_{ij} / max_i\{f_{ij}\}$

Pesos: Frequência Inversa dos Documentos

- Termos que aparecem em muitos documentos *diferentes* são *menos* significativos.

  df_i = frequência em documentos do termo *i*

  = número de documentos contendo o termo *i*

  idf_i = frequência inversa em documentos do termo *i,*

  = $\log_2 (N/ df_i)$

  (N: número total de documentos)
- É uma indicação do *poder de discriminação* do termo.
- Log é usado para diminuir o efeito em relação a *tf*.

Ponderação TF-IDF

- Uma ponderação tipicamente utilizada é:

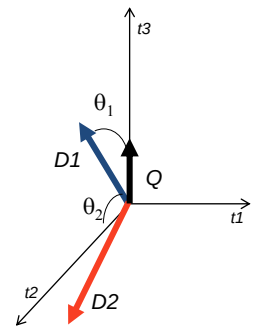
$$w_{ij} = tf_{ij}\, idf_i = tf_{ij} \log_2 (N/ df_i)$$

- Um termo que ocorre com frequência no documento mas raramente no resto da coleção tem peso maior.
- Muitas outras formas de ponderação foram propostas.
- Experimentalmente, determinou-se que a ponderação *tf-idf* funciona bem.

Medida de Similaridade de Cosseno

- Mede o cosseno do ângulo entre dois vetores.
- Produto interno normalizado pelo comprimento dos vetores.

$$CosSim(d_j, q) = \frac{\vec{d_j} \cdot \vec{q}}{|\vec{d_j}| \cdot |\vec{q}|} = \frac{\sum_{i=1}^{t}(w_{ij} \cdot w_{iq})}{\sqrt{\sum_{i=1}^{t} {w_{ij}}^2 \cdot \sum_{i=1}^{t} {w_{iq}}^2}}$$



D1 = 2T1 + 3T2 + 5T3 CosSim(*D1* , *Q*) = 10 / $\sqrt{(4+9+25)(0+0+4)}$ = 0.81
D2 = 3T1 + 7T2 + 1T3 CosSim(*D2* , *Q*) = 2 / $\sqrt{(9+49+1)(0+0+4)}$ = 0.13
Q = 0T1 + 0T2 + 2T3

D1 é 6 vezes melhor que *D2* usando similaridade de cosseno mas só 5 vezes melhor usando produto interno.

k-NN para Textos

Treinamento:

Para cada exemplo de treinamento <*x*, *c*(*x*)> ∈ *D*

Calcule o vetor TF-IDF correspondente, **d*x***, para o documento *x*

Exemplo de teste *y*:

Calcule o vetor TF-IDF **d** para o documento *y*

Para cada <*x*, *c*(*x*)> ∈ *D*

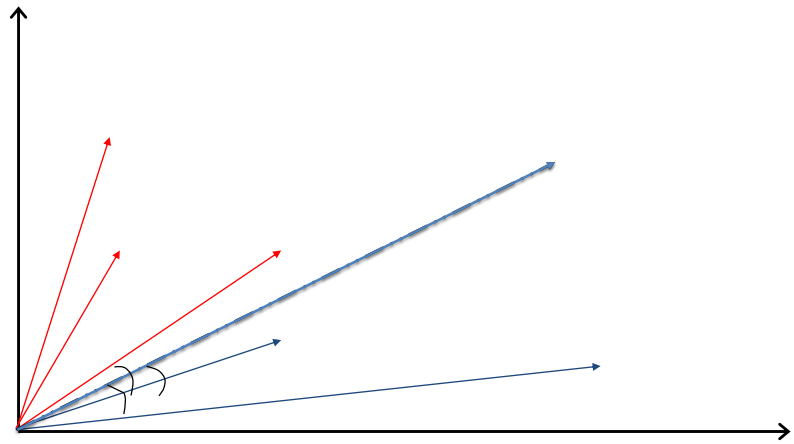
Seja ***sx*** = cosSim(**d**, **d*x***)

Ordene os exemplos, *x*, em *D* por valor decrescente de ***sx***

Seja *N* o conjunto dos primeiros *k* exemplos de D.

Retorne a classe majoritária dos exemplos em *N.*

Exemplo: 3-NN para Textos



Índice Invertido

- Busca linear na base de treinamento não é escalável.
- Índice invertido: estrutura mapeando palavras a documentos.
- Quando as *stopwords* são removidas, as palavras que sobram são raras, então um índice invertido ajuda a eliminar boa parte dos documentos que não tem muitas palavras em comum com o documento de teste.

Conclusões

- Existem muitas aplicações importantes da classificação de textos.
- Requer uma técnica que lide bem com vetores esparsos de muitos atributos, porque tipicamente cada palavra é um atributo e a maioria das palavras é rara.
  - Naïve Bayes
  - kNN com similaridade de cosseno
  - SVMs