Aprendizado Bayesiano

Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação
Departamento de Ciências de Computação
Rodrigo Fernandes de Mello
mello@icmc.usp.br

Anteriormente...

- Conceito de **Probabilidade Condicional**
  - É a probabilidade de um evento A dada a ocorrência de um evento B

$$P(A|B)$$

  - Formulamos:

$$P(A \mid B) = \frac{P(B \cap A)}{P(B)}$$

  - Ou:

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

Por exemplo

- Conceito de Probabilidade Condicional:
  - Quando há dependência:
    - Em Aprendizado Bayesiano assumimos que um valor de atributo ocorre em função de outro
  - Por exemplo:

Dia	Panorama	Temperatura	Umidade	Vento	Jogar Tênis
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Intermediária	Alta	Fraco	Sim
D5	Chuvoso	Fria	Normal	Fraco	Sim
D6	Chuvoso	Fria	Normal	Forte	Não
D7	Nublado	Fria	Normal	Forte	Sim
D8	Ensolarado	Intermediária	Alta	Fraco	Não
D9	Ensolarado	Fria	Normal	Fraco	Sim
D10	Chuvoso	Intermediária	Normal	Fraco	Sim
D11	Ensolarado	Intermediária	Normal	Forte	Sim
D12	Nublado	Intermediária	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Intermediária	Alta	Forte	Não

Por exemplo

- Conceito de Probabilidade Condicional:
  - Qual a probabilidade de um evento A dado que B ocorreu?

$$P(A \mid B) = \frac{P(B \cap A)}{P(B)}$$

  - Seja B igual a (Umidade = Alta)

Dia	Panorama	Temperatura	Umidade	Vento	Jogar Tênis
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Intermediária	Alta	Fraco	Sim
D5	Chuvoso	Fria	Normal	Fraco	Sim
D6	Chuvoso	Fria	Normal	Forte	Não
D7	Nublado	Fria	Normal	Forte	Sim
D8	Ensolarado	Intermediária	Alta	Fraco	Não
D9	Ensolarado	Fria	Normal	Fraco	Sim
D10	Chuvoso	Intermediária	Normal	Fraco	Sim
D11	Ensolarado	Intermediária	Normal	Forte	Sim
D12	Nublado	Intermediária	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Intermediária	Alta	Forte	Não

Por exemplo

- Conceito de Probabilidade Condicional:
  - Dois possíveis valores podem ser calculados para A:
    - Jogar Tênis = Sim
    - Jogar Tênis = Não

$$P(A \mid B) = \frac{P(B \cap A)}{P(B)}$$

  - Portanto:

$$P(Jogar\ Tênis = Sim|Umidade = Alta) = \frac{P(Jogar\ Tênis = Sim \cap Umidade = Alta)}{P(Umidade = Alta)}$$

$$P(Jogar\ Tênis = Não|Umidade = Alta) = \frac{P(Jogar\ Tênis = Não \cap Umidade = Alta)}{P(Umidade = Alta)}$$

Por exemplo

- Conceito de Probabilidade Condicional:
  - Quais as probabilidades?

$$P(Umidade = Alta) = \frac{7}{14} = 0.5$$

$$P(Jogar\ Tênis = Sim \cap Umidade = Alta) = \frac{3}{14} = 0.214$$

$$P(Jogar\ Tênis = Não \cap Umidade = Alta) = \frac{4}{14} = 0.286$$

  - Logo:

$$P(Jogar\ Tênis = Sim|Umidade = Alta) = \frac{\frac{3}{14}}{\frac{7}{14}} = 0.428$$

$$P(Jogar\ Tênis = Não|Umidade = Alta) = \frac{\frac{4}{14}}{\frac{7}{14}} = 0.571$$

Por exemplo

- Conceito de Probabilidade Condicional:
  - Conclusão:
    - Tendo certeza sobre a condição de Umidade Alta, podemos inferir que a probabilidade de:
      - Jogar Tênis = Sim é de 42.8%
      - Jogar Tênis = Não é de 57.1%

Teorema de Bayes

Teorema de Bayes

Teorema de Bayes

- Teorema de Bayes:

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

$$\text{Assim: } P(B \cap A) = P(A|B) \cdot P(B)$$

$$\text{Como } P(B \cap A) = P(A \cap B) \text{ logo:}$$

$$P(A|B) \cdot P(B) = P(B|A) \cdot P(A)$$

$$\text{e chegamos ao Teorema de Bayes: } P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Teorema de Bayes

- Em aprendizado de máquina queremos:
  - A melhor hipótese h_{MAP} de um espaço H dado que observamos um conjunto de treinamento D
    - Logo substituindo em:

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

    - Logo para uma hipótese h qualquer em H temos:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Teorema de Bayes

- No entanto, para obtermos h_{MAP} computamos:

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$

- Denominamos h_{MAP} como a hipótese com máxima probabilidade a posteriori
  - Ou seja, aquela que gera melhores resultados para conjuntos de dados nunca vistos dado o treinamento ocorrido sobre o conjunto D

Teorema de Bayes

Vejamos um Exemplo

Teorema de Bayes: Exemplo

- Considere o problema de diagnóstico médico em que há duas alternativas:
  - O paciente tem a doença W
  - O paciente não tem a doença W
- Considere que o paciente foi submetido a um exame
- Sabemos que:
  - Dada a população mundial, apenas 0.008 desta apresenta tal doença
  - O teste laboratorial pode apresentar:
    - O teste apresenta um correto positivo (verdadeiro positivo) para a presença da doença em 98% dos casos
    - O teste apresenta um correto negativo (verdadeiro negativo) para a ausência da doença em 97% dos casos

Teorema de Bayes: Exemplo

- Podemos resumir o cenário em:

$$P(W) = 0.008 \quad P(\neg W) = 0.992$$
$$P(\oplus|W) = 0.98 \quad P(\ominus|W) = 0.02$$
$$P(\oplus|\neg W) = 0.03 \quad P(\ominus|\neg W) = 0.97$$

- Suponha o exame de um paciente deu positivo:
  - Devemos diagnosticá-lo como portador da doença W?

Teorema de Bayes: Exemplo

- Nossa hipótese nesse caso é:
  - "O paciente é portador de W?"
- Buscamos, então pela máxima probabilidade a posteriori
  - Ou seja, a partir dos dados que temos, qual a probabilidade máxima de acertarmos tal hipótese sobre dados (exame deste paciente) nunca vistos?
- Conhecemos:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad \begin{array}{ll} P(W) = 0.008 & P(\neg W) = 0.992 \\ P(\oplus|W) = 0.98 & P(\ominus|W) = 0.02 \\ P(\oplus|\neg W) = 0.03 & P(\ominus|\neg W) = 0.97 \end{array}$$

Teorema de Bayes: Exemplo

- Sendo:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Necessitamos calcular:

$$P(W|\oplus) = \frac{P(\oplus|W)P(W)}{P(\oplus)}$$

O paciente tem a doença sabendo que o exame deu positivo?

$$P(\neg W|\oplus) = \frac{P(\oplus|\neg W)P(\neg W)}{P(\oplus)}$$

O paciente não tem a doença sabendo que o exame deu positivo?

Teorema de Bayes: Exemplo

- Nossa hipótese nesse caso é:
  - "O paciente é portador de W?"
- Buscamos, então pela máxima probabilidade a posteriori
  - Ou seja, a partir dos dados que temos, qual a probabilidade máxima de acertarmos tal hipótese sobre dados (exame deste paciente) nunca vistos?
- Conhecemos:

$$\begin{array}{ll} P(W) = 0.008 & P(\neg W) = 0.992 \\ P(\oplus|W) = 0.98 & P(\ominus|W) = 0.02 \\ P(\oplus|\neg W) = 0.03 & P(\ominus|\neg W) = 0.97 \end{array}$$

- Logo:

$$P(\oplus|W)P(W) = 0.98 \cdot 0.008 = 0.0078$$
$$P(\oplus|\neg W)P(\neg W) = 0.03 \cdot 0.992 = 0.0298$$

Teorema de Bayes: Exemplo

- Não temos $P(\oplus)$, mas como a probabilidade de ambos totalizam 1 (só há duas possibilidades), podemos normalizar:

$$P(\oplus|W)P(W) = 0.98 \cdot 0.008 = 0.0078$$
$$P(\oplus|\neg W)P(\neg W) = 0.03 \cdot 0.992 = 0.0298$$

- Temos:

$$P(W|\oplus) = \frac{0.0078}{0.0078 + 0.0298} = 0.207$$
$$P(\neg W|\oplus) = \frac{0.0298}{0.0078 + 0.0298} = 0.793$$

A maior probabilidade é que o paciente não é portador de W, apesar do teste ser positivo!

Teorema de Bayes: Exemplo

- Agora podemos calcular $P(\oplus)$:

$$P(W|\oplus) = \frac{P(\oplus|W)P(W)}{P(\oplus)}$$

$$P(\oplus) = \frac{P(\oplus|W)P(W)}{P(W|\oplus)}$$

  - A qual resume a probabilidade dos dados coletados serem positivos em todo o conjunto de dados

$$P(\oplus) = \frac{P(\oplus|W)P(W)}{P(W|\oplus)} = \frac{0.98 \cdot 0.008}{0.207} = \frac{0.03 \cdot 0.992}{0.793} = 0.038$$

Teorema de Bayes: Exemplo

- Conclusões:
  - A doença é tão rara que precisamos de um teste com maior taxa de acerto, por exemplo:
    - Considere um outro teste que apresenta um correto positivo (verdadeiro positivo) para a presença da doença em 99,5% dos casos
    - O teste apresenta um correto negativo (verdadeiro negativo) para a ausência da doença em 99,9% dos casos
  - Nessa situação teríamos:

$$P(W) = 0.008 \quad P(\neg W) = 0.992$$
$$P(\oplus|W) = 0.995 \quad P(\ominus|W) = 0.005$$
$$P(\oplus|\neg W) = 0.001 \quad P(\ominus|\neg W) = 0.999$$

Agora sim!

$$P(\oplus|W)P(W) = 0.995 \cdot 0.008 = 0.00796$$
$$P(\oplus|\neg W)P(\neg W) = 0.001 \cdot 0.992 = 0.000992$$

$$P(W|\oplus) = 0.89$$
$$P(\neg W|\oplus) = 0.11$$

Teorema de Bayes

Classificador Ótimo de Bayes

Classificador Ótimo de Bayes

- Considere um espaço de hipóteses H contendo somente h_1, h_2 e h_3
  - Seja a probabilidade a posteriori dessas hipóteses dada por:
    - $P(h_1|D) = 0.4$
    - $P(h_2|D) = 0.3$
    - $P(h_3|D) = 0.3$
  - Neste caso $h_{MAP} = h_1$
    - Ou seja, é a hipótese que apresenta máxima probabilidade a posteriori

Classificador Ótimo de Bayes

- Agora considere que um novo exemplo foi classificado como positivo por h_1 e negativo tanto por h_2 quanto h_3
  - Neste caso considerando todas hipóteses, temos que:
    - A probabilidade do exemplo ser positivo é 0.4
    - A probabilidade desse exemplo ser negativo é 0.6
  - Neste caso, a classificação dada pela hipótese mais provável é diferente da classificação dada considerando todas hipóteses
    - Nem sempre h_{MAP} acerta!
    - Muitas vezes votação usando hipóteses é uma boa opção
      - Ensembles de classificadores

Classificador Ótimo de Bayes

- Assim, podemos compor a classificação de cada hipótese do espaço em função de suas respectivas probabilidades:
  - Assim aumentamos a chance de acerto:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

  - A **Classificação Ótima de Bayes** busca pela opinião comum máxima entre as hipóteses:
    - Ótima pois considera TODAS hipóteses do espaço H

$$\arg\max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

Classificador Ótimo de Bayes

- Seja um exemplo classificado como positivo ou negativo segundo uma das hipóteses abaixo:

$$P(h_1|D) = 0.4, P(\ominus|h_1) = 0, P(\oplus|h_1) = 1$$
$$P(h_2|D) = 0.3, P(\ominus|h_2) = 1, P(\oplus|h_2) = 0$$
$$P(h_3|D) = 0.3, P(\ominus|h_3) = 1, P(\oplus|h_3) = 0$$

Resultados de classificação

Classificador Ótimo de Bayes

- Seja um exemplo classificado como positivo ou negativo segundo uma das hipóteses abaixo:

Foi classificado como positivo pela hipótese h1

$$P(h_1|D) = 0.4, P(\ominus|h_1) = 0, P(\oplus|h_1) = 1$$
$$P(h_2|D) = 0.3, P(\ominus|h_2) = 1, P(\oplus|h_2) = 0$$
$$P(h_3|D) = 0.3, P(\ominus|h_3) = 1, P(\oplus|h_3) = 0$$

Foi classificado como negativo tanto pela hipótese h2 quanto por h3

$$\sum_{h_i \in H} P(\oplus|h_i)P(h_i|D) = 0.4$$
$$\sum_{h_i \in H} P(\ominus|h_i)P(h_i|D) = 0.6$$

$$\arg \max_{v_j \in \{\oplus,\ominus\}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = \ominus$$

Classificador Ótimo de Bayes

- Qualquer sistema que classifique exemplos conforme a equação

$$\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

é dito ser um Classificador Ótimo Bayesiano

- Nenhum classificador, usando o mesmo espaço H, pode superar, na média, um classificador Ótimo Bayesiano
  - Funciona como um esquema de votação considerando TODAS AS POSSÍVEIS HIPÓTESES!!!

Classificador Ótimo de Bayes

- Podemos voltar no problema anterior de realizar testes para dada doença e concluir que:
  - Poderíamos ter diversos testes diferentes
    - Cada um representa um hipótese distinta
  - Podemos compor os resultados desses testes e ponderá-los por suas probabilidades
    - A fim de diagnosticarmos um paciente

Teorema de Bayes

Classificador Naive Bayes

Classificador Naive Bayes

- É uma alternativa ao Classificador Ótimo de Bayes
  - O Ótimo apresenta alto custo quando o número de hipóteses em H é alto
  - Ou impossível quando número de hipóteses é infinito
    - Pois considera todas as possíveis hipóteses!!!
- O Classificador Naive Bayes é aplicável quando:
  - Temos um conjunto de atributos que representa cada exemplo
  - Cada um desses exemplos tem uma classe
  - Este classificador é solicitado para produzir a classe de um exemplo nunca visto com base em exemplos de treinamento

Classificador Naive Bayes

- Segundo o Teorema de Bayes:
  - Buscamos classificar o novo exemplo segunda sua classe mais provável, dado seu conjunto de atributos <a_1, a_2, ..., a_n>:

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)}$$

- Agora precisamos estimar, a partir do conjunto de treinamento:
  - A probabilidade $P(v_j)$ que é simples de ser estimada
  - No entanto, assumindo que o conjunto de treinamento tem tamanho limitado:
    - Torna-se difícil estimar $P(a_1, a_2, \ldots, a_n | v_j)$, pois há possivelmente poucas ou nenhuma ocorrência idêntica no conjunto de treinamento (devido a seu tamanho, i.e., número de exemplos)
    - Esta segunda probabilidade poderia ser estimada, somente se o conjunto de treinamento fosse muito grande

Classificador Naive Bayes

- O Classificador naïve Bayes faz uma simplificação:
  - Assume que os valores de atributos são independentes
  - Em outras palavras, a probabilidade de observar $a_1, a_2, \ldots, a_n$ é, justamente, o produto das probabilidades de cada atributo individual:

$$P(a_1, a_2, \ldots, a_n | v_j) = \prod_i P(a_i | v_j)$$

  - Assim, o Classificador Naïve Bayes é uma simplificação, a qual é dada por:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

    - Ao usar naïve Bayes perceberemos que não há a busca explícita por uma hipótese
      - Hipótese é formada simplesmente pela contagem de frequências

Classificador Naive Bayes: Exemplo

- Considere:

Dia	Panorama	Temperatura	Umidade	Vento	Jogar Tênis
D1	Ensolarado	Quente	Alta	Fraco	Não
D2	Ensolarado	Quente	Alta	Forte	Não
D3	Nublado	Quente	Alta	Fraco	Sim
D4	Chuvoso	Intermediária	Alta	Fraco	Sim
D5	Chuvoso	Fria	Normal	Fraco	Sim
D6	Chuvoso	Fria	Normal	Forte	Não
D7	Nublado	Fria	Normal	Forte	Sim
D8	Ensolarado	Intermediária	Alta	Fraco	Não
D9	Ensolarado	Fria	Normal	Fraco	Sim
D10	Chuvoso	Intermediária	Normal	Fraco	Sim
D11	Ensolarado	Intermediária	Normal	Forte	Sim
D12	Nublado	Intermediária	Alta	Forte	Sim
D13	Nublado	Quente	Normal	Fraco	Sim
D14	Chuvoso	Intermediária	Alta	Forte	Não

- Suponha o novo exemplo:

<Panorama=Ensolarado, Temperatura=Fria, Umidade=Alta, Vento=Forte>

- Nossa tarefa é predizer Sim ou Não para a atividade Jogar Tênis

Classificador Naive Bayes: Exemplo

- Neste caso temos:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$$

- Logo:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \\ P(Panorama = Ensolarado | v_j) P(Temperatura = Fria | v_j) \\ P(Umidade = Alta | v_j) P(Vento = Forte | v_j)$$

- Computando:

$$P(Vento = Forte | Jogar\ Tênis = Sim) = 3/9 \\ P(Vento = Forte | Jogar\ Tênis = Não) = 3/5 \\ P(Umidade = Alta | Jogar\ Tênis = Sim) = 3/9 \\ P(Umidade = Alta | Jogar\ Tênis = Não) = 4/5 \\ P(Temperatura = Fria | Jogar\ Tênis = Sim) = 3/9 \\ P(Temperatura = Fria | Jogar\ Tênis = Não) = 1/5 \\ P(Panorama = Ensolarado | Jogar\ Tênis = Sim) = 2/9 \\ P(Panorama = Ensolarado | Jogar\ Tênis = Não) = 3/5$$

Classificador Naive Bayes: Exemplo

- Sendo:

$$P(v_j = Sim) = 9/14 \\ P(v_j = Não) = 5/14$$

- Logo:

$$P(Sim) P(Ensolarado | Sim) P(Fria | Sim) P(Alta | Sim) P(Forte | Sim) = 0.0053 \\ P(Não) P(Ensolarado | Não) P(Fria | Não) P(Alta | Não) P(Forte | Não) = 0.0206$$

- Normalizando temos que a probabilidade de não Jogar Tênis é de 0.795, ou seja, 79,5% de chance de não Jogar Tênis
  - Considera conjunto de dados discreto!!!

Classificador Naive Bayes: Exemplo

- Vamos implementar o Naive Bayes...

Classificador Naive Bayes: Estimação de Probabilidades

- Estimar probabilidades com base em conjuntos de treinamento não é uma tarefa simples
  - Por exemplo, o que ocorre se precisamos estimar usando um conjunto de treinamento pequeno, por exemplo com n=5 exemplos:

$$P(Vento = Forte|Jogar\ Tênis = Não)$$

  - Agora suponha que sabemos que:

$$P(Vento = Forte|Jogar\ Tênis = Não) = 0.08$$

  - Tendo apenas 5 exemplos de treinamento, logo temos 0.08 * 5 = 0.4 exemplo no conjunto de treinamento?
    - Ou seja, não temos nenhum exemplo capaz de representar esta situação
    - Como resolvemos este problema?

Classificador Naive Bayes: Estimação de Probabilidades

- Podemos usar a m-estimativa de probabilidade na forma:

$$\frac{n_c + mp}{n + m}$$

- Ou seja, anteriormente estimamos a probabilidade de eventos, simplesmente pela contagem, por exemplo, para o conjunto de treinamento com 14 exemplos temos:

$$P(Vento = Forte|Jogar\ Tênis = Não) = \frac{n_c}{n} = \frac{3}{5}$$

Classificador Naive Bayes: Estimação de Probabilidades

- Podemos usar a m-estimativa de probabilidade na forma:

$$\frac{n_c + mp}{n + m}$$

  - Em que p é uma estimativa da probabilidade e m é uma constante denominada de equivalente ao tamanho do conjunto de treinamento
  - Uma forma de escolher p é assumir uma distribuição uniforme para todos seus k valores possíveis, assim:

$$p = \frac{1}{k}$$

  - Por exemplo, o atributo Vento pode assumir dois valores, logo p=0.5
  - Perceba que se m=0 então: $\frac{n_c + mp}{n + m} = \frac{n_c}{n}$

Classificador Naive Bayes: Classificando Textos

- Considere que temos uma coleção de textos ou documentos e desejamos recuperar:
  - "Textos sobre o tópico A"
- Neste caso, desejamos recuperar os textos mais relevantes para determinada consulta feita por um usuário
- Seja:
  - Uma coleção X de documentos
  - Cada documento contém palavras de diferentes tamanhos
  - Um conjunto de treinamento está disponível em que conhecemos a classe ou tópico que cada texto está relacionado
  - V é o conjunto finito que contém todas as classes ou tópicos possíveis para os textos da coleção

Classificador Naive Bayes: Classificando Textos

- Para simplificar considere que:
  - Cada texto está classificado em duas classes por um dado usuário:
    - Classe interessante
    - Classe desinteressante
- Precisamos, primeiramente, organizar os textos para utilizá-los a fim de aprender o conceito interessante/desinteressante

Classificador Naive Bayes: Classificando Textos

- Assuma que temos:
  - Um conjunto de treinamento com 700 textos classificados como interessantes e 300 como desinteressantes
- Ao usar Naïve Bayes percebemos que:
  - Ele não considera a posição dos atributos, ou seja, ao invés de:

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n|v_j)P(v_j)$$

  - Ele considera:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$$

Classificador Naive Bayes: Classificando Textos

- Há um problema com a estimativa:
  - Se uma palavra ocorre 10 vezes em um texto de 100 palavras sua probabilidade é de 10/100 = 0.1
- Isso se dá pois diversas outras palavras seriam possíveis em cada texto, logo, ao invés de estimar as probabilidades usando:

$$\frac{n_c}{n}$$

- Usaremos:

$$\frac{n_c + mp}{n + m}$$

  - Em que m representa o número de palavras possíveis e p a probabilidade de cada palavra ocorrer
  - Considerando a Língua Inglesa podemos assumir m = 50.000 e p=1/50.000

Classificador Naive Bayes: Classificando Textos

- Dessa maneira, estimamos as probabilidades na forma:

$$\frac{n_c + 1}{n + 50.000}$$

- Assim temos um melhor panorama da probabilidade de ocorrer uma palavra em dado texto
  - Uma vez que consideramos um universo de possíveis valores ou palavras que poderiam ocorrer no mesmo texto
- Ao invés de assumirmos 50.000 como tamanho de nosso vocabulário de palavras, podemos melhor estimá-lo realizando a contagem de todas as diferentes palavras que ocorrem em TODOS os textos da coleção X
  - Assim teremos um cálculo mais preciso de probabilidades para o problema que estamos abordando

Classificador Naive Bayes: Classificando Textos

- Logo, o algoritmo de aprendizado é dado por:

Entrada:
Coleção X de textos
V é o conjunto de classes possíveis para cada texto
Saída:
Probabilidades

Algoritmo de Aprendizado:
1. Coletar todas as palavras que ocorrem em todos textos da coleção X
   Vocabulário ← conjunto de todas palavras distintas
2. Calcular a probabilidade $P(v_j)$ e $P(w_k|v_j)$
   Para cada valor v_j em V faça:
   $docs_j$ ← subconjunto de textos que apresentam a classe v_j
   $P(v_j)$ ← $|docs_j| / |X|$
   $Texto_j$ ← Texto único que concatena todos os membros de $docs_j$
   n ← número total de palavras distintas em $Texto_j$
   Para cada palavra w_k em Vocabulário
   n_k ← número de vezes que a palavra w_k ocorre no $Texto_j$
   $P(w_k|v_j)$ ← (n_k +1) / (n + |Vocabulário|)

Classificador Naive Bayes: Classificando Textos

- Logo, o algoritmo de classificação naïve Bayes é dado por:

Entrada:
Um texto
Saída:
A classe estimada para dado texto

Algoritmo de Classificação:
atributos ← lista de palavras encontradas no novo texto
Retornar:

$$v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{a_i \in atributos} P(a_i|v_j)$$

Classificador Naive Bayes: Classificando Textos

- Implementação
  - Implementar os algoritmos anteriores
  - Estar a classificação de textos em datasets como:
    - 20 Newsgroups Dataset
      - http://archive.ics.uci.edu/ml/datasets/Twenty+Newsgroups
    - Reuters-21578 Dataset
      - http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

Outras Questões

- Há situações em que atributos podem apresentar algum nível de dependência
  - Por exemplo, em um texto sobre “machine learning” a probabilidade de aparecer a palavra learning após machine é maior
  - Para isso pode-se utilizar Bayesian Belief Networks:
    - Considera as dependências entre atributos

Sugestões

- Instale R:
  - Instale os pacotes tm e Snowball
  - Implemente Naive Bayes para classificar o conjunto de texto Reuters-21578 Dataset
  - Funções importantes:
    - stopwords()
    - removeNumbers()
    - removePunctuation()
    - removeWords()
    - StemDocument()
    - stripWhitespace()
    - termFreq()
    - PlainTextDocument()

Referências

- Tom Mitchell, Machine Learning, 1994
- Wikipedia, Conditional Probability, http://en.wikipedia.org/wiki/Conditional_probability