Métodos estatísticos de aprendizagem

Baseada em:

RUSSEL, S.; PETER, N. “Inteligência Artificial” - cap 20.
MITCHELL, T. “Machine Learning” – cap. 6

Profa. Josiane M. Pinheiro
novembro/2008

Aprendizagem Estatística

- Dados são evidências – instanciações de alguma ou de todas as variáveis aleatórias do domínio
- Hipóteses são teorias probabilísticas de como o domínio funciona

Exemplo muito simples

- Suponha que nosso doce favorito está embalado em uma embalagem opaca e contém dois sabores:
  - Cereja (oba!)
  - Lima (argh!)
- O doce é vendido em sacos muito grandes, dos quais existem cinco tipos conhecidos:
  - 10% são h_1: 100% cereja
  - 20% são h_2: 75% cereja + 25% lima
  - 40% são h_3: 50% cereja + 50% lima
  - 20% são h_4: 25% cereja + 75% lima
  - 10% são h_5: 100% lima

Exemplo muito simples

- Dado um novo saco de doce, a variável aleatória H denota o tipo de saco, com valores possíveis de h_1 até h_5
- Não sabemos qual é o tipo de saco que temos, por isso H não é diretamente observável
- À medida que os doces são abertos, os dados são revelados: $D_1, D_2, ..., D_n$ onde cada D_i é uma variável aleatória com valores *cereja* e *lima*
- A tarefa básica é prever o sabor do próximo pedaço de doce

Aprendizagem Bayesiano

- Simplesmente calcula a probabilidade de cada hipótese, considerando os dados, e faz previsões de acordo com ela
- Previsões são feitas com uso de todas as hipóteses, ponderadas por suas probabilidades
- Seja **D** a representação de todos os dados, com valor observado **d**; então a probabilidade de cada hipótese é dada pela regra de Bayes:

$$P(h_i \mid \mathbf{d}) = \alpha \; P(\mathbf{d} \mid h_i) P(h_i)$$

Aprendizagem Bayesiano

- Previsões usam a média ponderada das probabilidades sobre a hipótese:

$$P(X \mid \mathbf{d}) = \sum_i P(X \mid \mathbf{d}, h_i) P(h_i \mid \mathbf{d})$$

$$P(X \mid \mathbf{d}) = \sum_i P(X \mid h_i) P(h_i \mid \mathbf{d})$$

- A probabilidade dos dados é calculada sob a suposição de que as observações são independentemente e identicamente distribuídas de forma que:

$$P(\mathbf{d} \mid h_i) = \prod_j P(d_j \mid h_j)$$

Cálculos para o exemplo

Probabilidades à priori da hipóteses :

$P(h_1) = 0{,}1 \quad P(h_2) = 0{,}2 \quad P(h_3) = 0{,}4 \quad P(h_4) = 0{,}2 \quad P(h_5) = 0{,}1$

Probabilidade do doce ser de lima, dado cada uma das hipóteses :

$P(d = l \mid h_1) = 0 \qquad P(d = l \mid h_2) = 0{,}25 \qquad P(d = l \mid h_3) = 0{,}5$

$P(d = l \mid h_4) = 0{,}75 \qquad P(d = l \mid h_5) = 1$

Probabilidade de cada uma das hipóteses, dado que o primeiro doce é de lima :

$P(h_i \mid \mathbf{d}) = \alpha\, P(\mathbf{d} \mid h_i) * P(h_i)$

$P(h_1 \mid d_1 = l) = \alpha\, P(d_1 = l \mid h_1) * P(h_1) = \alpha\, 0 * 0{,}1 = 0 / 0{,}5 = 0$

$P(h_2 \mid d_1 = l) = \alpha\, P(d_1 = l \mid h_2) * P(h_2) = \alpha\, 0{,}25 * 0{,}2 = 0{,}05 / 0{,}5 = 0{,}1$

$P(h_3 \mid d_1 = l) = \alpha\, P(d_1 = l \mid h_3) * P(h_3) = \alpha\, 0{,}5 * 0{,}4 = 0{,}2 / 0{,}5 = 0{,}4$

$P(h_4 \mid d_1 = l) = \alpha\, P(d_1 = l \mid h_4) * P(h_4) = \alpha\, 0{,}75 * 0{,}2 = 0{,}15 / 0{,}5 = 0{,}3$

$P(h_5 \mid d_1 = l) = \alpha\, P(d_1 = l \mid h_5) * P(h_5) = \alpha\, 1 * 0{,}1 = 0{,}1 / 0{,}5 = 0{,}2$

$\alpha = 0 + 0{,}05 + 0{,}2 + 0{,}15 + 0{,}1 = 0{,}5$

Cálculos para o exemplo

Probabilidade do doce ser de lima, dado cada uma das hipóteses :

$P(d = l \mid h_1) = 0 \qquad P(d = l \mid h_2) = 0{,}25 \qquad P(d = l \mid h_3) = 0{,}5$

$P(d = l \mid h_4) = 0{,}75 \qquad P(d = l \mid h_5) = 1$

Probabilidade de cada uma das hipóteses, dado que o primeiro doce é de lima :

$P(h_1 \mid d_1 = l) = 0 \;\; P(h_2 \mid d_1 = l) = 0{,}1 \;\; P(h_3 \mid d_1 = l) = 0{,}4$

$P(h_4 \mid d_1 = l) = 0{,}3 \;\; P(h_5 \mid d_1 = l) = 0{,}2$

Possibilidade do segundo doce ser de lima :

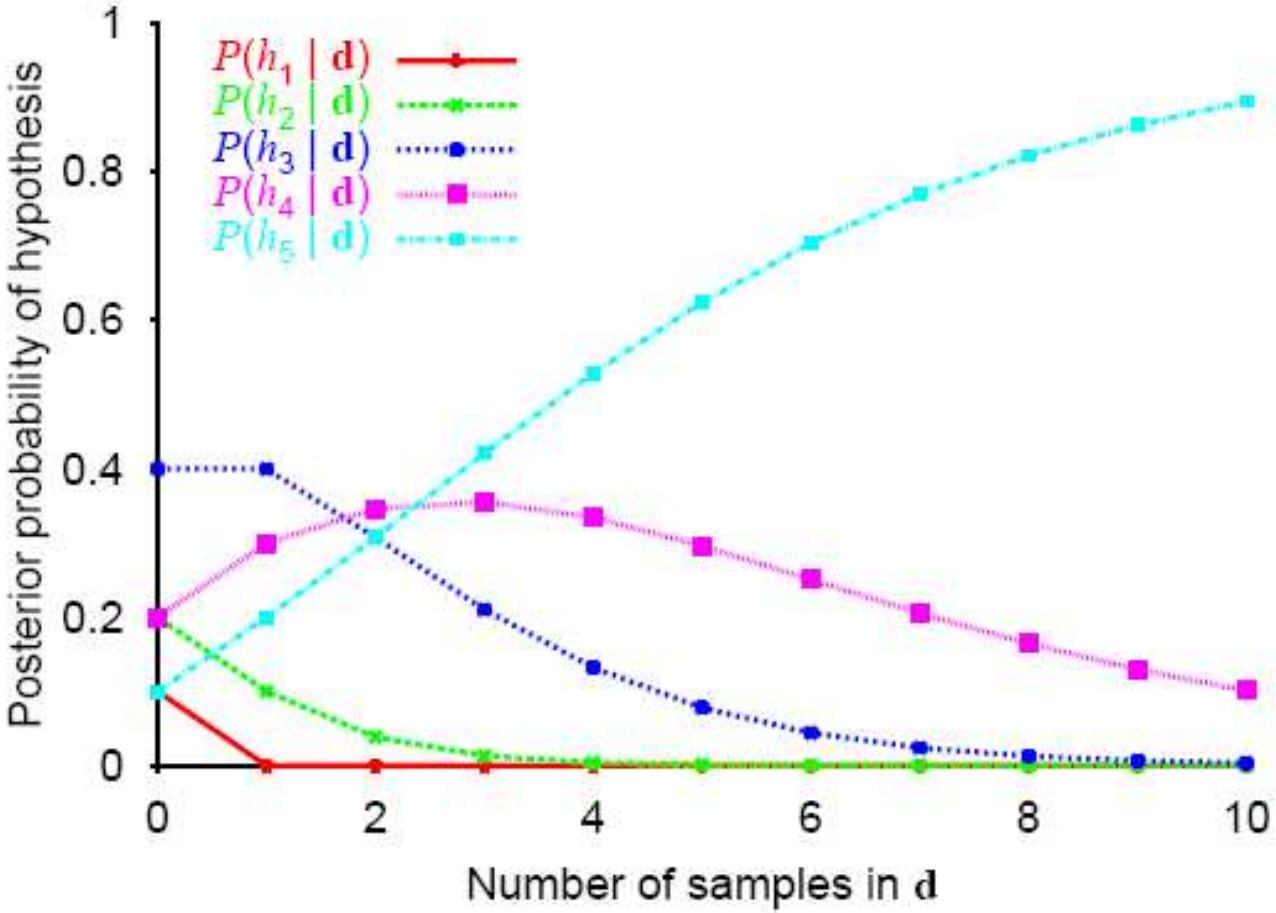
$$P(X \mid \mathbf{d}) = \sum_i P(X \mid h_i) P(h_i \mid \mathbf{d})$$

$$P(d_{proximo} = l \mid d_1 = l) = P(d_1 = l \mid h_1) P(h_1 \mid d_1 = l) + P(d_1 = l \mid h_2) P(h_2 \mid d_1 = l)$$

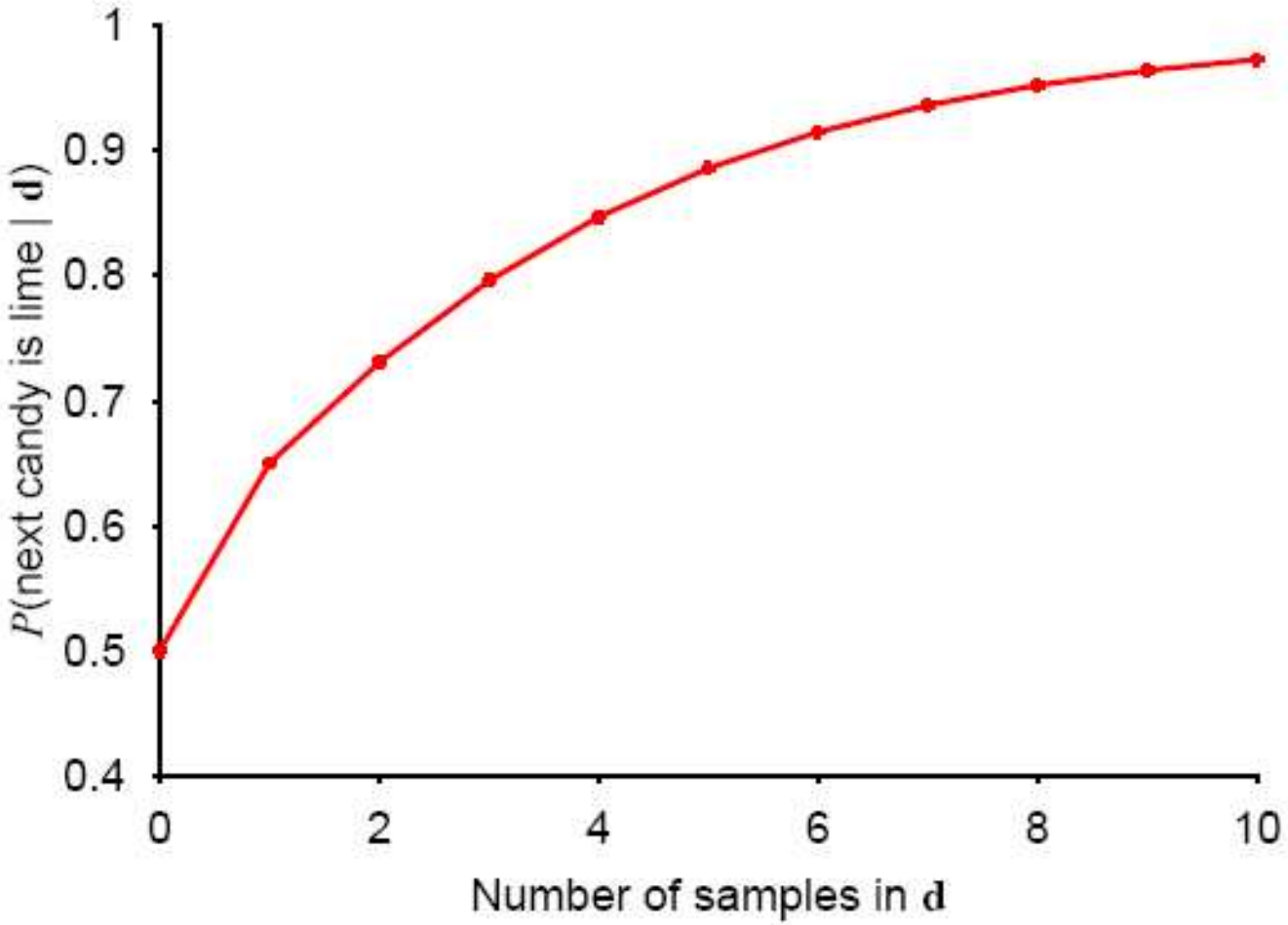
$$+ P(d_3 = l \mid h_1) P(h_3 \mid d_1 = l) + P(d_1 = l \mid h_4) P(h_4 \mid d_1 = l) + P(d_1 = l \mid h_5) P(h_5 \mid d_1 = l)$$

$$P(d_{proximo} = l \mid d_1 = l) = 0 + 0{,}25 * 0{,}1 + 0{,}5 * 0{,}4 + 0{,}75 * 0{,}3 + 1 * 0{,}2 = 0{,}65$$

Probabilidade Posteriores para 10 observações de doce de lima



Previsão Bayesiana para o sabor do próximo doce



Características da aprendizagem Bayesiana

- Para qualquer probabilidade **a priori** fixa que não ignora a hipótese verdadeira, a probabilidade posterior de qualquer hipótese falsa irá desaparecer
  - A probabilidade de gerar dados não-característicos indefinidamente é muitíssimo pequena
- A previsão bayesiana é otima, para qualquer tamanho de conjunto, pequeno ou grande
  - Para problemas reais de aprendizagem o espaço de hipóteses é em geral muito grande, ou infinito
  - Em muitos casos, para calcular o somatório da previsão de forma tratável, devemos recorrer a métodos aproximados

Hipótese de Máximo a Posteriori (MAP)

- Uma aproximação muito comum para calcular a previsão é fazê-la com base em uma única hipótese mais provável
  - Uma h_i que maximize $P(h_i|\mathrm{d})$ – hipótese de máximo a posteriori ou MAP
  - As previsões feitas desta forma são aproximadamente bayesianas até o ponto em que $P(X|\mathrm{d})$ é aproximadamente igual a $P(X|h_{\mathrm{MAP}})$
- No exemplo dos doces $h_{\mathrm{MAP}} = h_5$ após $d_1=d_2=d_3=lima$
  - Neste ponto a $P(d_4=lima) = 1$, enquanto a previsão bayesiana é mais prudente $P(d_4=lima) = 0{,}8$
  - À medida que chegam mais dados as previsões de MAP e Bayes se tornam cada vez mais próximas, porque as concorrentes da hipótese MAP se tornam cada vez menos prováveis

Hipótese de Máxima probabilidade (MP)

- Supõe uma probabilidade a priori **uniforme** sobre o espaço de hipóteses
  - Quando não existe nenhuma razão para preferir uma hipótese sobre outra a priori
- A aprendizagem MAP se reduz à escolha de um h_i que maximize $P(\mathbf{d}|H_i)$
- Proporciona uma boa aproximação para a aprendizagem bayesiana e de MAP quando o conjunto de dados é grande

Modelo de Bayes Ótimo

- Dada uma nova instância X, qual é a sua *classificação* mais provável? *cereja* ou *lima*?
  - Vimos que $h_{MAP}(X)$ nem sempre é a classificação mais provável
- Considere a probabilidade das três hipóteses:
  - $P(h_1/D) = 0$, $P(h_2|D) = 0.05$, $P(h_3|D) = 0.2$, $P(h_4|D) = 0.35$ e $P(h_5|D) = 0.4$
- E as probabilidades de lima dada cada hipótese:
  - $P(\text{lima}|h_1) = 0$, $P(\text{lima}|h_2) = 0{,}25$, $P(\text{lima}|h_3) = 0.5$, $P(\text{lima}|h_4) = 0.75$ e $P(\text{lima}|h_5) = 1$
- Qual é a classificação mais provável de x? *lima* ou *cereja*?

Modelo de Bayes Ótimo

Classificação do próximo doce após os dados abservados $\mathbf{D}$:

$$P(X \mid \mathbf{D}) = \sum_i P(X \mid h_i) P(h_i \mid \mathbf{D})$$

$$P(l \mid \mathbf{D}) = P(l \mid h_1) P(h_1 \mid \mathbf{D}) + P(l \mid h_2) P(h_2 \mid \mathbf{D}) + P(l \mid h_3) P(h_3 \mid \mathbf{D})$$
$$+ P(l \mid h_4) P(h_4 \mid \mathbf{D}) + P(l \mid h_5) P(h_5 \mid \mathbf{D}) = 0{,}8$$

$$P(c \mid \mathbf{D}) = P(c \mid h_1) P(h_1 \mid \mathbf{D}) + P(c \mid h_2) P(h_2 \mid \mathbf{D}) + P(c \mid h_3) P(h_3 \mid \mathbf{D})$$
$$+ P(c \mid h_4) P(h_4 \mid \mathbf{D}) + P(c \mid h_5) P(h_5 \mid \mathbf{D}) = 0{,}2$$

Desta forma, a classificação do próximo doce será = lima $(0{,}8 > 0{,}2)$

Modelo de Bayes Ótimo

Se a possível classificação do novo exemplo pode ser qualquer valor $v_j \in V$, a probabilidade de que a classificação correta seja v_j :

$$P(v_j \mid \mathbf{D}) = \sum_i P(v_j \mid h_i) * P(h_i \mid \mathbf{D})$$

e a classificação Bayesiana ótima será :

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_i P(v_j \mid h_i) * P(h_i \mid \mathbf{D})$$

Modelo de Bayes Ingênuo (Naive Bayes)

- Suponha uma função de classificação f: $X \rightarrow$ V, onde cada instância X é descrita pelos atributos {a_1, …, a_n}

- O valor mais provável de f(x) é:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j \mid a_1, ..., a_n)$$

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} \frac{P(a_1, ..., a_n \mid v_j) * P(v_j)}{P(a_1, ..., a_n)}$$

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1, ..., a_n \mid v_j) * P(v_j)$$

Modelo de Bayes Ingênuo (Naive Bayes)

- Calcular $P(v_j)$ a partir dos dados de treinamento é fácil, o problema é calcular a probabilidade $P(a_1,..., a_n \mid v_j)$

- Suposição Bayesiana Ingênua - as variáveis $a_1,..., a_n$ são independentes :

$$P(a_1,...,a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

- Classificador Bayesiano Ingênuo:

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(a_1,...,a_n \mid v_j) * P(v_j)$$

$$V_{MAP} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) * \prod_i P(a_i \mid v_j)$$

Modelo de Bayes Ingênuo (Naive Bayes)

Dia	Tempo	Temp.	Humid.	Vento	Jogar
D1	Sol	Quente	Alta	Fraco	Não
D2	Sol	Quente	Alta	Forte	Não
D3	Coberto	Quente	Alta	Fraco	Sim
D4	Chuva	Normal	Alta	Fraco	Sim
D5	Chuva	Frio	Normal	Fraco	Não
D6	Chuva	Frio	Normal	Forte	Não
D7	Coberto	Frio	Normal	Forte	Sim
D8	Sol	Normal	Alta	Fraco	Não
D9	Sol	Frio	Normal	Fraco	Sim
D10	Chuva	Normal	Normal	Fraco	Sim
D11	Sol	Frio	Alta	Forte	?

P(Sim) = 5/10 = 0.5
P(Não) = 5/10 = 0.5

P(Sol|Sim) = 1/5 = 0.2
P(Sol|Não) = 3/5 = 0.6

P(Frio|Sim) = 2/5 = 0.4
P(Frio|Não) = 2/5 = 0.4

P(Alta|Sim) = 2/5 = 0.4
P(Alta|Não) = 3/5 = 0.6

P(Forte|Sim) = 1/5 = 0.2
P(Forte|Não) = 2/5 = 0.4

P(Sim)*P(Sol|Sim)*P(Frio|Sim)*P(Alta|Sim)* P(Forte|Sim) = 0.0032
P(Não)*P(Sol|Não)*P(Frio|Não)*P(Alta|Não)*P(Forte|Não) = 0.0288
⇒ Jogar_Tenis (D11) = Não

Modelo de Bayes Ingênuo (Naive Bayes)

Dia	Tempo	Temp.	Humid.	Vento	Jogar
D1	Sol	Quente	Alta	Fraco	Não
D2	Sol	Quente	Alta	Forte	Não
D3	Coberto	Quente	Alta	Fraco	Sim
D4	Chuva	Normal	Alta	Fraco	Sim
D5	Chuva	Frio	Normal	Fraco	Não
D6	Chuva	Frio	Normal	Forte	Não
D7	Coberto	Frio	Normal	Forte	Sim
D8	Sol	Normal	Alta	Fraco	Não
D9	Sol	Frio	Normal	Fraco	Sim
D10	Chuva	Normal	Normal	Fraco	Sim
D12	Coberto	Normal	Normal	Forte	?

Exercício: Calcule o classificação do exemplo D12 utilizando os 10 exemplos de treinamento acima.

Modelo de Bayes Ingênuo (Naive Bayes)

- Suposição de independência condicional frequentemente violada
- Mas funciona surpreendentemente bem
- Não é necessário calcular a probabilidade posterior $P(v_j|X)$, somente o valor máximo para cada v_j

$$\underset{v_j \in V}{\operatorname{argmax}} P(a_1,...,a_n \mid v_j) * P(v_j)$$

Naive Bayes para classificação de documentos

- Classificar documentos em duas classes:
  - {interesse, não-interesse}

- Variáveis $a_1,...a_n$ são palavras de um vocabulário e $P(a_i|v_j)$ é a freqüência com que cada palavra a_i aparece entre os documentos da classe v_j

- $P(v_j)$ = (nº de doc da classe v_j)/(nº total de doc)