Experimento de um Classificador de Padrões Baseado na Regra *Naive* de Bayes

WILIAN SOARES LACERDA[1]
ANTÔNIO DE PÁDUA BRAGA[2]

[1]UFLA - Universidade Federal de Lavras
DCC - Departamento de Ciência da Computação
Cx Postal 37 - CEP 37.200-000 Lavras (MG)
(lacerda)@ufla.br

[2]UFMG - Universidade Federal de Minas Gerais
DELT - Departamento de Engenharia Eletrônica
Av. Antônio Carlos, 6.627 CEP 31.270-901 Belo Horizonte (MG)
apbraga@cpdee.ufmg.br

Resumo. Este artigo apresenta um método estatístico de classificação de padrões comumente utilizado e baseado na Regra de Bayes. A apresentação de um exemplo de solução de problema de classificação/predição comprova a simplicidade e eficiência do método.

Palavras-Chave: Regra de Bayes, probabilidade, máquina de aprendizado

1 Introdução

Com o desenvolvimento crescente do uso de computadores e sensores de baixo custo para coleta de dados, há uma grande volume de dados sendo gerados por sistemas físicos, biológicos e sociais. Tais dados prontamente disponíveis podem ser usados para derivar modelos pela estimação de relações úteis entre variáveis do sistema (isto é, dependências entrada-saídas desconhecidas).

Um método de aprendizagem é um algoritmo (usualmente implementado em software) que estima um mapeamento desconhecido (dependência) entre entradas do sistema e saídas baseado nos dados disponíveis, isto é, nas amostras (entradas, saídas) conhecidas [1]. Uma vez que a dependência tem sido estimada, ela pode ser usada para a predição das saídas futuras do sistema baseado nos valores de entrada conhecidos.

Em estatística, a tarefa de aprendizado preditivo (baseado nas amostras) é chamado estimação estatística. Ela resulta das propriedades de estimação de alguma distribuição estatística (desconhecida) tendo como base as amostras conhecidas ou dados de treinamento. Informações contidas nos dados de treinamento (experiências passadas) podem ser usadas para responder questões sobre amostras futuras.

Neste artigo, é apresentado o método de classificação estatístico baseado na Regra de Bayes. É considerado que as entradas são independentes entre si, o que na maioria dos problemas práticos não é verdade (daí o nome *naive* ou ingênuo). Esta suposição simplifica a abordagem do problema de classificação sem comprometer significativamente a precisão do resultado. Um exemplo simples de aplicação do método é mostrado utilizando dados nominais.

2 Conceitos Básicos de Probabilidade

Para um melhor entendimento do conteúdo do artigo, é apresentado um resumo dos principais conceitos e métodos de análise das variáveis aleatórias.

2.1 Experimento aleatório

Um experimento aleatório é aquele no qual o resultado varia de modo imprevisível, quando é repetido nas mesmas condições [3].

2.2 Variável aleatória (v.a.)

É uma função que associa um número real $X(\xi)$ a cada aparecimento ξ no espaço de amostragem do experimento aleatório [3]. É comum representar uma variável aleatória por uma letra maiúscula (como X, Y, W) e qualquer valor particular da variável aleatória por uma letra minúscula (tal como x, y, w).

Uma variável aleatória pode ser considerada uma função que mapeia todos elementos do espaço de amostragem (coisas) nos pontos da linha real (números) ou alguma parte dela [3]. Mais de um ponto do espaço amostral pode ser mapeado em um mesmo valor da variável aleatória.

2.3 Probabilidade

Para definir probabilidade pode-se analisar o experimento de se jogar um dado (cubo com 6 faces numeradas) e observar o número que aparece em sua face superior. Existem seis números que podem ser o resultado. Pode-se assim definir dois conjuntos para este experimento: o conjunto de todos os possíveis resultados e o conjunto das possibilidades de ocorrência dos resultados. O conjunto de todos os possíveis resultados é chamado de espaço amostral, simbolizado como **S**. Todo experimento possui o seu espaço amostral.

Um evento é definido como um subconjunto do espaço amostral. No exemplo de se jogar um dado, pode-se definir o evento "resultar em um número ímpar". Este evento é um conjunto com três elementos.

Para cada evento definido em um espaço amostral **S**, deseja-se atribuir um número não negativo chamado probabilidade. Probabilidade é uma função dos eventos definidos. A notação adotada é $P(A)$, para "a probabilidade de ocorrência do evento A". Dois axiomas importantes dizem que: $P(A) \geq 0$ e $P(\mathbf{S}) = 1$, ou seja, a probabilidade de ocorrência de qualquer evento é sempre maior que zero (e menor que 1) e a probabilidade de ocorrência de um evento definido como sendo o espaço amostral **S** é sempre 1 [3]. Em termos gerais, a probabilidade de ocorrência de um evento A, $P(A)$, será igual ao número de ocorrências do evento A (n_A) dividido pelo número de ocorrências total (N), do espaço amostral (Equação 1) [5].

$$P(A) = \frac{n_A}{N} \tag{1}$$

2.4 Probabilidade condicional

Define-se a probabilidade condicional de um evento A tendo ocorrido um evento B (com probabilidade diferente de zero) como sendo [3] mostrado pela Equação 2.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2}$$

onde $P(A \cap B)$ é a probabilidade de ocorrência simultânea dos eventos A e B, isto é, probabilidade conjunta de A e B (também descrita simplesmente como $P(A, B)$).

Para $P(A) \neq 0$, pode-se escrever também:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{3}$$

Combinando as Equações 2 e 3 tem-se a principal forma do teorema de Bayes mostrado na Equação 4.

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \tag{4}$$

2.5 Independência estatística

Dois eventos (A e B) são estatisticamente independentes se a probabilidade da ocorrência de um evento não é afetada pela ocorrência do outro evento. Matematicamente, isto é descrito pelas Equações 5 e 6.

$$P(B|A) = P(B) \tag{5}$$

$$P(A|B) = P(A) \tag{6}$$

Independência também significa que a probabilidade da ocorrência conjunta (interseção) de dois eventos deve ser igual ao produto das probabilidades dos dois eventos (Equação 7).

$$P(A \cap B) = P(A)P(B) \tag{7}$$

2.6 Função de distribuição de probabilidade

Função distribuição de probabilidade acumulativa (também chamada de cdf, por sua abreviação, do inglês *cumulative probability distribution function*) de uma variável aleatória X, denotada por $F_X(x)$, é definida como sendo a probabilidade $P\{X \leq x\}$ (Equação 8). A função distribuição é uma função de x [3].

$$F_X(x) = P\{X \geq x\} \tag{8}$$

O argumento x pode ser qualquer número variando de $-\infty$ até $+\infty$.

2.7 Função de densidade de probabilidade

Função densidade de probabilidade (também chamada de pdf, por sua abreviação, do inglês *probability density function*) de uma variável aleatória X, denotada por $f_X(x)$, é definida como sendo a derivada da função distribuição [3] (Equação 9).

$$f_X(x) = \frac{dF_X(x)}{dx} \tag{9}$$

2.8 Valor médio de uma variável aleatória

O valor esperado (E) ou média $(\bar{X})$ de uma variável aleatória X é definida como mostrado pela Equação 10. A média aritmética de um grande número de observações independentes de uma variável aleatória X tenderá a convergir para $E(X)$.

$$E(X) = \int_{-\infty}^{+\infty} \xi f_X(\xi) d\xi \qquad (10)$$

3 Sistemas de Aprendizado

As tarefas específicas de aprendizagem são:

- Classificação: reconhecimento de padrões ou estimação de fronteiras (limites) de decisão de classe.
- Regressão: estimação de funções contínuas desconhecidas de dados ruidosos.
- Estimação: de densidade de probabilidade das amostras.

Em sistemas de aprendizado, existem os seguintes estágios de operação:

- aprendizagem/estimação (baseado nas amostras de treinamento).
- operação/predição quando predições são feitas para futuro ou amostras de teste.

Aprendizado supervisionado é usado para estimar um mapeamento (entrada/saída) desconhecido baseado em amostras (entrada/saída) conhecidas. Classificação e regressão são exemplos de tarefas deste tipo. O termo supervisionado denota o fato que valores de saída para amostras de treinamento são conhecidos.

Em aprendizado não supervisionado, apenas amostras de entrada são dados ao sistema de aprendizado, e não há noção da saída durante aprendizado. O objetivo do aprendizado não supervisionado pode ser estimar a distribuição de probabilidade das entradas ou descobrir uma estrutura natural (isto é, agrupamentos) nos dados de entrada.

A distinção entre aprendizado supervisionado e não supervisionado está no nível da declaração apenas do problema. Isto não implica que métodos originalmente desenvolvidos para aprendizado supervisionado não possa ser usado (com pequenas modificações) para tarefas de aprendizado não supervionado, e vice-versa.

O problema de aprendizagem/estimação de dependência dos dados é apenas parte do procedimento experimental geral para tirar conclusões dos dados. O procedimento experimental geral adotado em estatística clássica envolve as seguintes etapas:

1. Declaração do problema.
2. Formulação de hipóteses: especifica uma dependência desconhecida a qual é para ser estimada dos dados experimentais.
3. Geração de dados/Projeto do experimento: artificial, natural, observado, controlado.
4. Coleta de dados e pré-processamento: escalonamento, codificação.
5. Estimação do modelo: predição precisa, capacidade de generalização.
6. Interpretação do modelo/esboço das conclusões.

O modelo resultante não poderá ser válido se os dados não são informativos ou a formulação do problema não é estatisticamente significativa.

3.1 Dependência Estatística

Inferência estatística e sistemas de aprendizado estão interessados em estimação das dependências não conhecidas escondidas nos dados como mostra a Figura 1 [1].

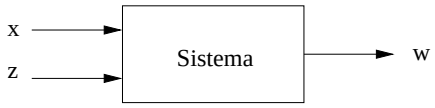

Figura 1: Um sistema com entradas-saída

O objetivo do aprendizado preditivo é estimar dependências não conhecidas entre as variáveis de entrada $(\mathbf{x})$ e a saída (w), de um conjunto de observações passadas de valores $(\mathbf{x}, w)$. O outro conjunto de variáveis rotuladas $\mathbf{z}$ denotam todos outros fatores que afetam a saída mas cujos valores não são observados ou controlados. Portanto, o conhecimento de valores de entrada observáveis $(\mathbf{x})$ não especificam unicamente as saídas (w). Esta incerteza nas saídas refletem a falta de conhecimento dos fatores não observados (z), e isto resulta em dependência estatística entre os dados observados e saídas. O efeito de entradas não observáveis (z) pode ser caracterizado pela distribuição de probabilidade condicional $p(w|x)$, o qual denota a probabilidade que w ocorrerá dado a entrada x.

3.2 Aprendizado adaptativo

Com amostras finitas, é sempre melhor resolver diretamente um exemplo particular do problema de aprendizado do que tentar resolver um problema mais geral

(e muito mais difícil) de estimação de densidade conjunta (entrada, saída) [1]. Os métodos clássicos podem não ser apropriados para muitas aplicações porque modelamento paramétrico (com amostras finitas) impõem muitas suposições rígidas sobre a dependência desconhecida; isto especifica sua forma paramétrica. Isto tende a introduzir grande polarização no modelamento, isto é, a discrepância do modelo paramétrico assumido e a (desconhecida) verdade.

Iqualmente, métodos não paramétricos clássicos trabalham apenas em casos assintóticos (tamanho das amostras muito grande).

As limitações da abordagem clássica providenciam motivação para métodos adaptativos (ou flexíveis). Métodos adaptativos conseguem maior flexibilidade por especificando uma larga classe de funções de aproximação (do que métodos paramétricos). O modelo de predição é então selecionado desta larga classe de funções. O principal problema torna-se escolher o modelo de complexidade ótima (flexibilidade) para os dados finitos a disposição.

Aprendizado é o processo de estimar uma dependência (entrada, saída) ou estrutura desconhecida de um sistema usando um limitado número de observações. O cenário geral de aprendizado envolve três componentes: um gerador de vetores de entrada aleatório, um sistema que retorna uma saída para dado vetor de entrada, e a máquina de aprendizado a qual estima um mapeamento (entrada, saída) desconhecido do sistema baseado nas amostras (entrada, saída) observadas (veja Figura 2).

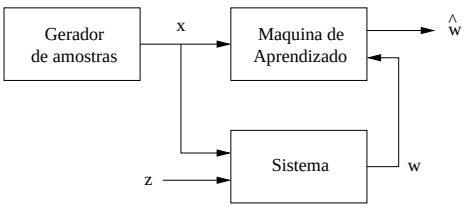

Figura 2: Um sistema de aprendizado com entradas-saída

O gerador produz vetores aleatórios $x \in R^d$ tirados independentemente de uma densidade de probabilidade fixa $p(x)$, a qual é desconhecida. Em geral, o modelador (máquina de aprendizado) não tem controle sobre quais vetores de entrada são fornecidos ao sistema.

O sistema produz um valor de saída w para todo vetor de entrada x de acordo com uma densidade de probabilidade condicional $p(w|x)$, o qual é também desconhecido. Esta descrição inclui o caso específico de um sistema determinístico onde $w = f(x)$, e também o caso da formulação de regressão de $w = f(x) + \epsilon$, onde ϵ é um ruído aleatório com média zero. Sistemas reais raramente tem saídas aleatórias verdadeiras; entretanto eles geralmente tem entradas não medidas (z). Estatisticamente, o efeito da variação das entradas não observáveis (z) na saída do sistema pode ser caracterizado como aleatório e representado como uma distribuição de probabilidade.

No caso mais geral, a máquina de aprendizado é capaz de implementar um conjunto de funções $f(\mathbf{x}, w)$, $w \in W$, onde W é um conjunto de parâmetros abstratos usados apenas para indexar o conjunto de funções. Na formulação o conjunto de funções implementadas pela máquina de aprendizado pode ser qualquer conjunto de funções, escolhidos a priori, antes do processo de inferência formal (aprendizado) ser iniciado.

O problema encontrado pela máquina de aprendizado é selecionar uma função (do conjunto de funções que ele suporta) que melhor aproxima a resposta do sistema. A máquina de aprendizado é limitado a observar um número finito de exemplos (n) em ordem para produzir esta seleção. Estes dados de treinamento como produzidos pelo gerador e sistema serão independentes e identicamente distribuídos de acordo a função de densidade de probabilidade conjunta (pdf):

$$p(\mathbf{x}, w) = p(\mathbf{x})p(w|\mathbf{x})$$

As amostras finitas (dados de treinamento) desta distribuição é indicada por:

$$f(\mathbf{x_i}, w_i), (i = 1, ..., n)$$

4 Regra de Bayes

Suponha que se conheça a probabilidade prévia (*a priori*) $P(w_j)$ e a densidade condicional $p(x|w_j)$ para $j = 1, 2$. A densidade de probabilidade conjunta de se encontrar um padrão que é da categoria w_j e possui característica de valor x (ou seja: $p(w_j, x)$), pode ser escrito de duas maneiras, mostradas na Equação 11 [2].

$$p(w_j, x) = P(w_j|x).p(x) = p(x|w_j).P(w_j) \quad (11)$$

Rearranjando a Equação 11, obtem-se a chamada fórmula de Bayes (Equação 12).

$$P(w_j|x) = \frac{p(x|w_j).P(w_j)}{p(x)} \quad (12)$$

onde para o caso de duas categorias:

$$p(x) = \sum_{j=1}^{2} p(x|w_j).P(w_j) \quad (13)$$

A fórmula de Bayes pode ser expressa informalmente em palavras como:

$$posterior = \frac{verossimilhanca \times prévio}{evidência} \quad (14)$$

A fórmula de Bayes mostra que observando o valor de x pode-se converter a probabilidade *a priori* $P(w_j)$ para a probabilidade *a posteriori* $P(w_j|x)$ - a probabilidade do estado natural de w_j, dado que o valor x da característica tem sido medido. Chama-se $p(x|w_j)$ a verossimilhança de w_j com respeito a x, um termo escolhido para indicar a categoria w_j para qual $p(x, w_j)$ é maior e mais parecida para ser a categoria verdadeira. O fator evidência, $p(x)$, pode ser visto meramente como fator de escala que garante que a probabilidade posterior soma para 1.

Tendo-se uma observação x para qual $P(w_1|x)$ é maior que $P(w_2|x)$, poderia-se naturalmente ser inclinado a decidir que o estado natural real é w_1. Quando observa-se um x particular, a probabilidade do erro de classificação é dada pela Equação 15.

$$P(erro|x) = \begin{cases} P(w_1|x) & \text{se decidir por } w_2 \\ P(w_2|x) & \text{se decidir por } w_1 \end{cases} \quad (15)$$

Para minimizar a probabilidade do erro, a regra de decisão de Bayes torna-se:

- "Decida w_1 se $P(w_1|x) > P(w_2|x)$, senão decida w_2."

Então:

$$P(erro|x) = min[P(w_1|x), P(w_2|x)] \quad (16)$$

Ou a seguinte regra de decisão equivalente:

- "Decida w_1 se $p(x|w_1).P(w_1) > p(x|w_2).P(w_2)$ caso contrário decida w_2."

4.1 Regra *Naive* de Bayes

Quando as relações de dependência entre os dados de entrada utilizadas por um classificador são desconhecidas, geralmente procede-se por tomar a simples suposição de que os dados são condicionalmente independentes dado a categoria. Assim, dados:

$$x = (x_1, ..., x_d)^T$$

então:

$$p(x|w_j) = \prod_{i=1}^{d} p(x_i|w_j) \quad (17)$$

Esta fórmula aplicada à Equação 12 é a tão chamada regra *naive* de Bayes. Na prática, geralmente apresenta bons resultados, como é exemplificado a seguir.

4.2 Exemplo de emprego da Regra de Bayes

A Tabela 1 mostra distribuições de frequência para a afinidade entre as características e a classe no conjunto de dados do Jogo de Tênis [4]. Desta tabela, é fácil calcular as probabilidades necessárias para aplicar a Regra *Naive* de Bayes.

As probabilidades para características nominais são estimadas usando contagem de frequência calculadas dos dados de treinamento. Qualquer frequência zero são recolocadas por $\frac{0.5}{m}$ como a probabilidade, onde m é o número de exemplos de treinamento. Assim, obtém-se as Tabelas 2 à 5 a partir da Tabela 1.

Imagine acordar de manhã e desejar determinar se o dia é apropriado para um Jogo de Tênis. Notando que o tempo está ensolarado, a temperatura está quente, a humidade é normal e o vento é fraco, aplica-se as Equações 17 e 12 e calcula-se a probabilidade posterior de cada classe (ignorando o fator evidência), usando probabilidades derivadas das Tabelas 2 à 5:

$$\begin{aligned} p(\text{não}jogar|ensolarado, quente, normal, fraco) = \\ p(\text{não}jogar) \times p(ensolarado|\text{não}jogar) \times \\ p(quente|\text{não}jogar) \times p(normal|\text{não}jogar) \times \\ p(fraco|\text{não}jogar) = \\ \frac{5}{14} \times \frac{3}{5} \times \frac{2}{5} \times \frac{1}{5} \times \frac{2}{5} = \\ 0.0069 \end{aligned}$$

$$\begin{aligned} p(jogar|ensolarado, quente, normal, fraco) = \\ p(jogar) \times p(ensolarado|jogar) \times \\ p(quente|jogar) \times p(normal|jogar) \times \\ p(fraco|jogar) = \\ \frac{9}{14} \times \frac{2}{9} \times \frac{2}{9} \times \frac{6}{9} \times \frac{6}{9} = \\ 0.0141 \end{aligned}$$

Assim, neste dia é recomendável jogar tênis.

Devido a suposição que os valores das características são independentes dentro da classe, o classificador *naive* de Bayes apresenta uma performance de predição desfavoravelmente afetada pela presença de atributos redundantes nos dados de treinamento. Por exemplo, se há uma característica X que é perfeitamente correlacionada com uma segunda característica Y, então tratando elas como significados diferentes de X ou Y, tem o dobro do efeito na Equação 12 do que ela poderia ter. Assim, moderadas dependências entre as características resultarão em imprecisão na estimação da probabilidade, mas as probabilidades não são tão fortes para resultar no incremento do erro de classificação.

Tabela 1: Dados do Jogo de Tênis

	Características				
Exemplo	Tempo	Temperatura	Humidade	Vento	Classe
1	ensolarado	quente	alta	fraco	não jogar
2	ensolarado	quente	alta	forte	não jogar
3	nublado	quente	alta	fraco	jogar
4	chuva	média	alta	fraco	jogar
5	chuva	frio	normal	fraco	jogar
6	chuva	frio	normal	forte	não jogar
7	nublado	frio	normal	forte	jogar
8	ensolarado	média	alta	fraco	não jogar
9	ensolarado	frio	normal	fraco	jogar
10	chuva	média	normal	fraco	jogar
11	ensolarado	média	normal	forte	jogar
12	nublado	média	alta	forte	jogar
13	nublado	quente	normal	fraco	jogar
14	chuva	média	alta	forte	não jogar

Tabela 2: Ocorrências da característica Tempo

	Jogar	Não Jogar	
ensolarado	2	3	5
nublado	4	0	4
chuva	3	2	5
	9	5	14

Tabela 3: Ocorrências da característica Temperatura

	Jogar	Não Jogar	
quente	2	2	4
média	4	2	6
frio	3	1	4
	9	5	14

Tabela 4: Ocorrências da característica Humidade

	Jogar	Não Jogar	
alta	3	4	7
normal	6	1	7
	9	5	14

Tabela 5: Ocorrências da característica Vento

	Jogar	Não Jogar	
forte	3	3	6
fraco	6	2	8
	9	5	14

5 Conclusão

Foram apresentados os conceitos básicos de Sistemas de Aprendizado, voltado para os métodos baseados no conjunto de dados conhecidos. As principais definições utilizadas em estatística e probabilidades foram mostradas para entendimento da Regra de Bayes. A Regra *Naive* de Bayes foi utilizada na solução de um problema exemplo, mostrando a eficiência e simplicidade do método.

Referências

[1] CHERKASSKG, Vladimir & MULIER, Filip. Learning from Data: Concepts, Theory, and Methods. New York: John Wiley & Sons, 1998, 441p.

[2] DUDA, Richard O., HART, Peter E., STORK, David G. Pattern Classification. New York: John Wiley & Sons, 2000, 654 p.

[3] GARCIA, Alberto Leon. Probability and Random processes for Electrical Engineering. New York: Addison-Wesley Publishing Company, 1989, 583 p.

[4] HALL, Mark A. Correlation-based Feature Selection for Machine Learning. Department of Computer Science - University of Waikato, Hamilton - New Zealand, April, 1999, 178p. (PhD Thesis)

[5] MAGALHÃES, Marcos Nascimento & LIMA, Antônio Carlos Pedroso. Noções de Probabilidade e Estatística. São Paulo: Editora da Universidade de São Paulo, 4a edição, 2002, 416 p.