Classificador Bayesiano: Notas de aula

Prof. Frederico Gualberto Ferreira Coelho

12 de novembro de 2018

Sumário

Capítulo 1

Probabilidade Condicional

Para entendermos melhor a regra de bayes que será explicada no capítulo 2 e consequentemente desenvolvermos a formulação do classificador Bayesiano temos que fazer uma pequena revisão em probabilidades condicionais.

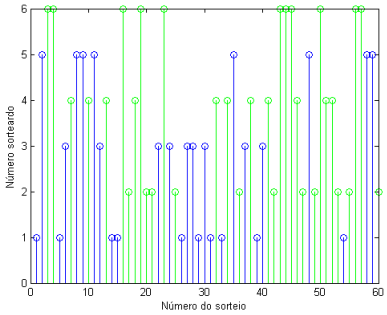
Uma boa maneira de iniciarmos o assunto é lançarmos mão do exemplo de um dado não viciado como mostrado em [Kay06]. Facilmente podemos calcular a probabilidade de ocorrer um número 1 ao se lançar o dado. Esta probabilidade será de $1/6$ visto que todos os seis números possuem a mesma chance de serem sorteados (o dado é não viciado). Contudo, qual será a probabilidade de ser sorteado o mesmo número 1 se agora sabemos de antemão que o número sorteado foi ímpar.

A probabilidade condicional se aplica exatamente este tipo de problema, onde se tem uma informação à priori sobre o possível resultado. Não interessa aqui discutir como esta informação foi obtida, mas simplesmente, queremos calcular a probabilidade de um evento A ocorrer dado que um segundo evento B ocorreu.

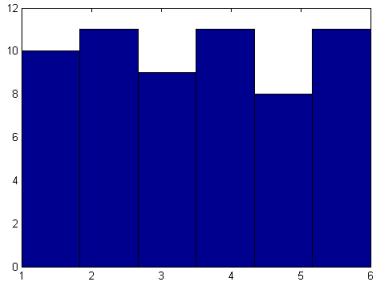
Continuando no exemplo do dado, imagine que tenhamos uma sequência de resultados de 60 lançamentos em série de um dado não viciado. Os resultados são mostrado na figura 1.1(a). Os eventos em verde são os números pares sorteados, e os azuis são os ímpares.

Do ponto de vista frequentista, podemos analisar os resultados obtidos na figura 1.1 levando-se em consideração o histograma dos resultados,mostrado na figura 1.1(b). De acordo com esta abordagem verificamos que o evento "sorteio do número 1" ocorreu exatamente[1] 10 vezes em 60 tentativas. Assim,

[1] foi dito "exatamente" pois o número de ocorrências do número 1 dividido por 60, neste



(a) Resultado dos lançamentos



(b) Histograma dos resultados

Figura 1.1: Problema do dado

podemos afirmar que a probabilidade de ocorrer um número 1 em qualquer lançamento é de aproximadamente $1/6$. Contudo, qual é a probabilidade de ocorrer um número 1 dado que sabemos de antemão que só foram sorteados números ímpares[2]?

Ou seja, qual a probabilidade de ocorrer o evento $A = \{1\}$ dado que o evento $B = \{1, 3, 5\}$ ocorreu? Com base nas informações contidas na figura 1.1 podemos afirmar que a probabilidade é de 10 dividido por 27 (que é o total de ocorrências dos números ímpares). Assim a probabilidade de A ocorrer dado que B ocorreu pode ser aproximada pela equação 1.1.

$$P(A|B) \approx \frac{N_A}{N_B} \approx \frac{10}{27} \approx \frac{1}{3} \tag{1.1}$$

Contudo o problema poderia ser diferente, por exemplo, definir a probabilidade de ocorrer o evento $A = \{1, 4\}$ dado que ocorreu o evento B definido anteriormente. Ora, como o evento B inclui apenas números ímpares, temos que tomar cuidado para levar em conta apenas os elementos em A que podem ocorrer. Assim, a equação 1.1 assume a forma da equação 1.3, onde N_S é o

exemplo construído aleatoriamente, dá exatamente $1/6$ que é a probabilidade de ocorrência de qualquer um dos números de um dado não viciado. Contudo, escolheu-se o número três de propósito para o exemplo ficar mais claro. A questão é que, para que todos os números, na abordagem frequentista, tenham exatamente a mesma frequencia, é necessário uma série de sorteios aleatórios muito grande, o que poluiria muito a figura 1.1 dificultando a visualização do problema.

[2]sem se preocupar com o porque só foram sorteados números ímpares e, claro, considerando um dado não viciado

número de elementos do espaço de eventos $S = \{1, 2, 3, 4, 5, 6\}$.

$$\frac{N_{A \cap B}}{N_B} = \frac{\frac{N_{A \cap B}}{N_S}}{\frac{N_B}{N_S}} \approx \frac{P[A \cap B]}{P[B]} \tag{1.2}$$

$$P[A|B] = \frac{P[A \cap B]}{P[B]} \tag{1.3}$$

Em outras palavras, a probabilidade de A dado que B ocorreu é, aproximadamente, o número de vezes que ocorreu A e B dividido pelo número de vezes que ocorreu B. O termo $P[A \cap B]$ é chamado de probabilidade conjunta e o termo $P[B]$ é chamado de probabilidade marginal e sua função é normalizar a probabilidade conjunta. O termo $P[A|B]$ é chamado de probabilidade condicional. É óbvio que $P[B] \neq 0$, caso contrário toda esta formulação não teria sentido.

É importante ressaltar que a probabilidade condicional é de fato uma probabilidade para um evento B específico, pois os axiomas que definem uma probabilidade são respeitados [Pap91].

$$P[A|B] \geq 0 \tag{1.4}$$

O primeiro axioma, mostrado na equação 1.4 é facilmente satisfeita já que $P[A \cap B] \geq 0$ e $P[B] > 0$.

A equação 1.5 explicita o segundo axioma que pode ser derivado diretamente do fato de que se $B \subset A$ então $P[A|B] = 1$

$$P[S|B] = 1 \tag{1.5}$$

Para provar o terceiro axioma, definido na equação 1.7, observamos que se os eventos A e C são mutuamente exclusivos então os eventos $A \cap B$ e $C \cap B$ são mutuamente exclusivos também, o que nos leva à equação 1.6 que resulta na equação do terceiro axioma.

$$P[A \cup C|B] = \frac{P[(A \cup C)\, B]}{P[B]} = \frac{P[A \cap B] + P[C \cap B]}{P[B]} \tag{1.6}$$

$$P[A \cup C|B] = P[A|B] + P[C|B] \tag{1.7}$$

Assim podemos concluir que todos resultados envolvendo probabilidades também são válidos para probabilidades condicionais.

Capítulo 2

A regra de Bayes

A definição da probabilidade condicional no capítulo 1 nos permite deduzir uma importante fórmula, chamada regra ou teorema de Bayes, que pode ser utilizada para computar probabilidades condicionais.

A equação 1.3, também é conhecida como teorema da probabilidade total, e é reproduzida na equação 2.1, rearanjada, de forma a isolarmos o termo da probabilidade conjunta.

$$P[A \cap B] = P[A|B]\, P[B] \tag{2.1}$$

Da mesma forma como temos $P[A \cap B]$ calculada como na equação 2.1, ela também pode ser calculado como na equação 2.2.

$$P[A \cap B] = P[B|A]\, P[A] \tag{2.2}$$

Substituindo $P[A \cap B]$ da equação 2.2 na equação 2.1 e rearranjando os termos chegaremos na equação 2.3.

$$P[A|B] = \frac{P[B|A]\, P[A]}{P[B]} \tag{2.3}$$

A equação 2.3 é conhecida como o "teorema de Bayes". Os termos $P[A]$ e $P[B]$ são chamadas de probabilidades marginais ou evidências, e o termo $P[B|A]$ é conhecido como verossimilhança. Conhecendo as probabilidades marginais e uma probabilidade condicional (a verossimilhança), podemos inferir sobre a outra probabilidade condicional. Isto nos permite avaliar a validade de um evento uma vez que outro foi observado.

2.1 Um exemplo da aplicação do teorema de Bayes

Para um melhor entendimento do teorema de Bayes veja o seguinte exemplo retirado de [Pap91].

Uma pessoa pertence a uma população de 100.000 habitantes onde 2.000 deles têm o diagnóstico de câncer. Esta pessoa se submete a um teste específico cuja eficácia na identificação da doença é de 95% e seu resultado foi positivo. O que se pode concluir sobre a probabilidade desta pessoa ter câncer?

Nós não podemos concluir, simplesmente, que a probabilidade desta pessoa ser diagnosticada com câncer seja de 95%. A análise não pode ser feita assim.

A eficiência do teste é de 95%, o que significa que 95% dos casos realmente positivos serão classificados corretamente (como positivos) e o mesmo percentual dos casos negativos serão classificados corretamente (como negativos). Vamos então definir que o evento T ocorre quando o resultado do teste é positivo, e o evento N ocorre para os resultados negativos. Também representaremos o conjunto de pessoas saudáveis por H e o conjunto de pessoas com câncer de C.

Desta forma temos que:

$$P[T|C] = 0.95$$
$$P[T|H] = 0.05$$

e que:

$$P[N|C] = 0.05$$
$$P[N|H] = 0.95$$

Na falta de outras informações, uma pessoa escolhida ao acaso nesta população tem a probabilidade de $98.000/100.00 = 0.98$ de ser saudável e, consequentemente, de 0.02 de sofrer de câncer. Assim, $P[H] = 0.98$ e $P[C] = 0.02$.

Agora podemos utilizar o teorema de Bayes para interpretar corretamente o resultado do teste de câncer aplicado ao paciente. A probabilidade do

paciente sofrer de câncer, dado que o seu teste retornou positivo é dado pela equação 2.4.

$$P\left[C|T\right] = \frac{P\left[T|C\right]P\left[C\right]}{P\left[T\right]} \tag{2.4}$$

Pelo teorema da probabilidade total temos que $P\left[T\right] = P\left[T|C\right]P\left[C\right] + P\left[T|H\right]P\left[H\right]$. Substituindo esta equação em 2.4 e substituindo os valores, chegamos ao resultado da equação 2.5.

$$P\left[C|T\right] = \frac{P\left[T|C\right]P\left[C\right]}{P\left[T|C\right]P\left[C\right] + P\left[T|H\right]P\left[H\right]} = \frac{0.95 \times 0.02}{0.95 \times 0.02 + 0.05 \times 0.98} = 0.278 \tag{2.5}$$

Podemos concluir então, que, se uma pessoa, que não sabe se tem ou não câncer, fizer o teste e ele der positivo, ela terá uma probabilidade de 27,8% de ter realmente a doença. Contudo, se for sabido que ela tem câncer, então a probabilidade do teste realmente indicar a doença será de 95%.

Capítulo 3

Classificador Bayesiano

3.1 Definição do problema

Imagine o seguinte problema: definir a qual classe pertence uma determinada nova observação (amostra) baseado nas características que podem ser extraídas dele e de outras observações previamente classificadas.

Neste trabalho vamos lidar com o problema de classificação em duas classes de uma observação com duas características.

Assim, os dados que temos para os padrões já classificados são:

- um vetor de duas dimensões para cada amostra ou "padrão" contendo a informação das suas duas características observadas ou medidas.
- um vetor contendo a classificação dos padrões já conhecidos, ou seja, a informação sobre qual classe pertence o padrão.

Enquanto que para os padrões novos, que se quer classificar, só temos o vetor com suas características. As figuras 3.1 e 3.1 exemplificam bem o problema.

A figura 3.1 mostra os dados sobre os quais já se conhece previamente sua classificação. Eles compõem as classes "bolinha" (o) e "cruzinha" ($+$). O problema então é definir a quais destas duas classes pertencem os dados novos que não se conhece sua classificação, baseado nas informações do vetor de características $X_{i,j} = [X_{i,1}X_{i,2}]$ de cada padrão i já conhecido e nas informações do vetor de características dos próprios padrões que se quer classificar. Espera-se que estas informações sejam suficientes para discriminar os padrões. Estes padrões que se quer classificar são mostrados na figura 3.1.

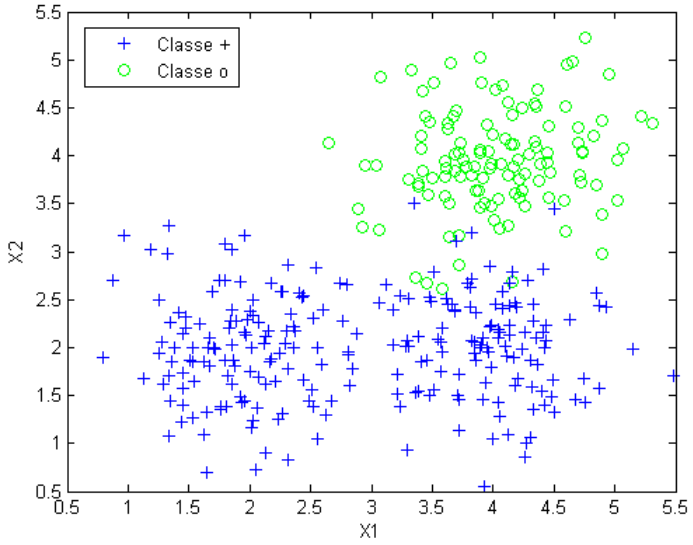


Figura 3.1: Dados classificados (rotulados) do problema, ou seja, estes são os dados dos quais já se conhece sua classificação (classe *o* ou classe +)

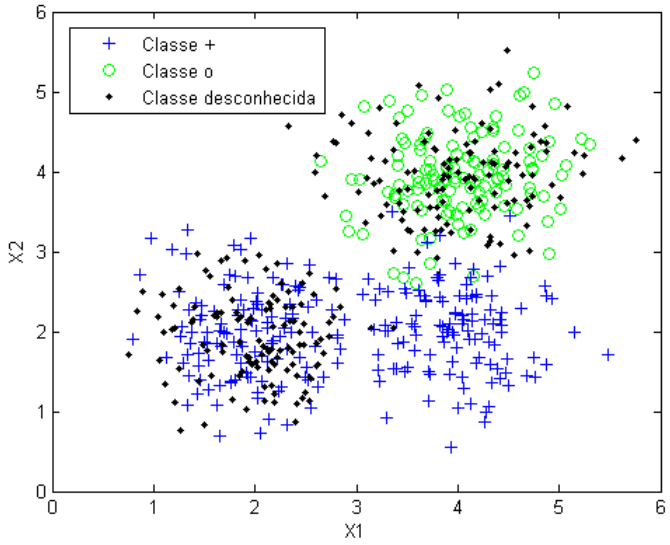


Figura 3.2: Dados rotulados e não rotulados do problema

Para uma quantidade suficientemente grande de dados seremos capazes de determinar as funções de densidade de probabilidade que vão nos ajudar na discriminação dos dados novos, ou seja, na solução do problema.

3.2 Probabilidades à priori

As probabilidades à priori são funções de densidade de probabilidade que podemos extrair do conjunto de dados existente, na tentativa de se estabelecer uma maneira de classificar os novos padrões. Quando não se tem uma regra clara associando o vetor de características X à classe correspondente, são estas probabilidades que nos ajudarão a classificá-los, ou seja, a decisão de qual classe devemos associar X só poderá ser tomada baseada em informações à priori sobre a esperança de cada classe.

O conhecimento à priori sobre a probabilidade de ocorrer elementos das classes C_1 (o) e C_2 $(+)$ são obtidos pelas probabilidades $P[C_1]$ e $P[C_2]$ respectivamente. Elas podem ser entendidas como sendo a chance de cada elemento pertencer a uma classe ou outra sem conhecermos a distribuição do vetor de características X. Desta forma, estas probabilidades são calculadas como na equação 3.2 onde n_1 e n_2 são a quantidade de padrões das classes 1 e 2 respectivamente.

$$P[C_1] = \frac{n_1}{n_1 + n_2} \tag{3.1}$$

$$P[C_2] = \frac{n_2}{n_1 + n_2} \tag{3.2}$$

3.3 Um classificador ingênuo

Um classificador ingênuo (do termo em inglês "naive classifier") pode ser construído sem se saber nada a respeito de X e de seu relacionamento com as duas classes, apenas associando o novo padrão à classe de maior probabilidade à priori.

Suponha, novamente, o exemplo do ítem 2.1, onde se quer definir a probabilidade do paciente estar realmente doente dado que o teste para câncer de acurácia de 95% deu positivo. Se não se conhece nada à priori, nem a acurácia do teste, mas apenas a quantidade da população que está doente

ou não, simplesmente, decidiríamos classificar o paciente como sadio, pois estaríamos cometendo, no máximo, um erro de 2%.

Do ponto de vista estatístico este tipo de classificação não é tão ruim, ainda mais se as classes forem muito desbalanceadas, contudo, se o problema impõem altos custos para erros de classificação este tipo de classificador se torna pobre. No caso do nosso exemplo, um erro de classificação significa a morte do paciente, assim, não podemos aplicar, pura e simplesmente, um classificador ingênuo. Mas será necessário adicionarmos mais informação ao modelo para podermos melhorar a classificação.

3.4 Verossimilhanças

Se informações de X forem fornecidas ao classificador, as probabilidades à priori $P[C_1]$ e $P[C_2]$ poderão ser alteradas pelas probabilidades condicionais $P[x|C_1]$ e $P[x|C_2]$, podendo levar a melhores classificações do novo padrão. Estas duas probabilidades condicionais são as mais importantes na regra de Bayes. Elas são as verossimilhanças de X.

De forma análoga às probabilidades à priori, estas verossimilhanças podem ser determinadas a partir dos dados. Elas podem ser obtidas estimando as funções de densidade de probabilidade de X separadamente para cada padrão em C_1 e C_2. Uma vez determinadas estas quantidades, já podemos aplicar a regra de Bayes.

A evidência de X ($P[X]$) pode ser calculada, uma vez que X pode ocorrer em qualquer uma das duas classes. Para o nosso exemplo de duas classes $P[X]$ é dado pela equação 3.3, que nada mais é que a equação da probabilidade total descrita anteriormente pela equação 1.3. Todavia, não é necessário calcular este valor para projetarmos um classificador Bayesiano, como será explicado no ítem 3.5

$$P[X] = P[X|C_1]\,P[C_1] + P[X|C_2]\,P[C_2] \tag{3.3}$$

3.5 O classificador Bayesiano

O decisor de Bayes, que minimiza o risco de classificação global [Bra09], define que um padrão, cujo vetor de características é X, será associado à classe C_j de maior probabilidade a posteriori $P[C_1|x]$.

Contudo, o classificador de Bayes pode ser simplificado. Ao invés de se calcular a probabilidade a posteriori de x pertencer à C_1, usando as equações 3.4 e 3.5, podemos prescindir do cálculo de $P[X]$ dividindo uma probabilidade à posteriori pela outra.

$$P[C_1|x] = \frac{P[x|C_1]\,P[C_1]}{P[x]} \tag{3.4}$$

$$P[C_2|x] = \frac{P[x|C_2]\,P[C_2]}{P[x]} \tag{3.5}$$

Ou seja, para o nosso caso de duas classes, a regra geral do classificador de Bayes pode ser simplificada para associar X à classe C_1 se $\frac{P[C_1|x]}{P[C_2|x]} \geq 1$, caso contrário será associado à classe C_2.

Capítulo 4

Implementação

4.1 Dados

Como pode ser implementado o classificador de Bayes na prática, para o problema de classificação proposto na seção 3.1? Vamos explicar a implementação diretamente com um exemplo numérico para maior clareza das explicações.

O que temos inicialmente são os dados do vetor característica X que podemos dispor em uma matriz $X_{i,j} = \{X_{i,1}, X_{i,2}\}$ para cada um dos n_1 e n_2 padrões, tanto da classe C_1 ou (o) como para a classe C_2 ou $(+)$. Onde i e j são os índices do padrão em questão e da característica respectivamente. Obviamente conhecemos também, para estes padrões, qual a sua classificação. A tabela 4.1 exemplifica os dados para os 5 primeiros padrões da classe C_1 identificada como -1 na coluna com título "Classe" e da classe C_2 identificado como $+1$.

A classe 1 é composta de 240 padrões (n_1) e a classe 2 é composta por 120 padrões (n_2). A figura 3.1 mostra estes pontos graficamente.

Os dados a serem classificados são ao todo 240 (n_u) e apenas possuimos os valores de X_1 e X_2, que para não causar confusão definiremos, para o conjunto não rotulado, como sendo o vetor de caractrísticas $X_{u_{z,j}} = \left\{X_{u_{z,1}}, X_{u_{z,2}}\right\}$. O índice z indica o padrão não rotulado em questão. O que queremos é determinar a Classe de cada um dos padrões não rotulados. A tabela 4.2 exemplifica o conjunto de dados não rotulados e a figura 3.1 mostra estes pontos graficamente.

Tabela 4.1: Exemplo parcial dos dados dos padrões rotulados

Padrão	X_1	X_2	Classe
1	2,00	1,74	+1
2	1,82	2,20	+1
3	2,60	2,40	+1
4	0,96	3,16	+1
5	2,23	1,25	+1
⋮	⋮	⋮	⋮
241	3,50	4,75	-1
242	4,07	4,72	-1
243	3,92	3,49	-1
244	3,36	2,73	-1
245	4,65	4,98	-1
⋮	⋮	⋮	⋮

Tabela 4.2: Exemplo parcial dos dados dos padrões não rotulados

Padrão	X_1	X_2	Classe
1	1,66	2,24	**?**
2	2,37	2,31	**?**
3	2,01	1,61	**?**
4	2,58	2,45	**?**
5	1,26	0,76	**?**
⋮	⋮	⋮	⋮

4.2 Algoritmo

De posse dos dados descritos na seção 4.1, podemos iniciar a explicação do algoritmo e os cálculos.

1º passo - calcular as evidências $P[C_1]$ e $P[C_2]$

Estas probabilidades são facilmente calculadas. Para tanto só necessitamos saber o número de elementos rotulados de cada classe. As equações 4.1

e 4.2 demonstram estes cálculos.

$$P\left[C_1\right] = \frac{n_1}{n_1 + n_2} = \frac{240}{240 + 120} = 0.667 \tag{4.1}$$

$$P\left[C_2\right] = \frac{n_2}{n_1 + n_2} = \frac{120}{240 + 120} = 0.333 \tag{4.2}$$

2º passo - preparação para o cálculo das verossimilhanças

Aqui é necessário fazer uma pequena observação para prosseguirmos com os cálculos. Queremos determinar, para cada elemento $X_{u_{z,j}}$, a sua probabilidade considerando que a sua classe seja C_1, e depois considerando que seja C_2, de acordo com a distribuição dos dados rotulados, para podermos classificar este elemento.

Uma boa aproximação, baseado no Teorema do limite central [Kay06, Pap91], é considerar o dados iid e a função de densidade de probabilidade das classes gaussiana.

Para compreendermos melhor a definição e o cálculo desta grandeza, vamos considerar um pequeno exemplo de uma variável. Imagine dois grupos de pessoas, cuja distribuição de idades em cada um deles seja ditado por uma função gaussiana, cada uma com média e variância diferentes, tal qual a figura 4.1. O grupo G_1 de pesoas tem média de idade igual a 16 anos e o grupo G_2 tem média igual a 20 anos.

Tomamos aleatóriamente um indivíduo que pode pertencer a qualquer um destes dois grupos. Suponha que a idade deste indivíduo seja de 19 anos. Assim, a verossimilhança deste indivíduo, **se** a classe dele for G_1, será de 2,4% (identificado por um $\times$ na figura 4.1). Já se considerarmos que a classe do elemento seja a G_2 então sua verossimilhança será de 26% (ponto identificado por um círculo vermelho no gráfico).

Então, para o cálculo destas verossimilhanças vamos considerar uma função de densidade de probabilidade gaussiana para duas variáveis, descrita na equação 4.3 onde μ_1 e μ_2 são as médias dos dados de cada uma das variáveis X_1 e X_2, σ_1 e σ_2 são seus desvios padrão e ρ é o coeficiente de correlação entre as variáveis.

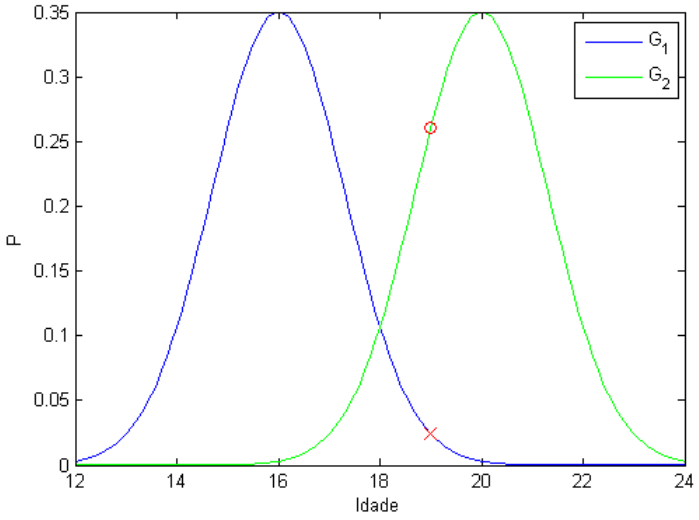


Figura 4.1: Exemplo cálculo da verossimilhança

$$PDF_{C_1}\left(X_{u_1}, X_{u_2}\right) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{X_{u_1}-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{X_{u_1}-\mu_1}{\sigma_1}\right)\left(\frac{X_{u_2}-\mu_2}{\sigma_2}\right)+\left(\frac{X_{u_2}-\mu_2}{\sigma_2}\right)^2\right]} \tag{4.3}$$

O coeficiente de correlação é calculado conforme a equação 4.4. Onde *cov* é a covariância entre as variáveis.

$$\rho = \frac{cov\left(X_{u_1}, X_{u_2}\right)}{\sqrt{\sigma_1^2\sigma_2^2}} \tag{4.4}$$

Calculando então para C_1:

$$\mu_{1_{C_1}} = \overline{X}_{1_{C_1}} = 2.93 \tag{4.5}$$

$$\mu_{2_{C_1}} = \overline{X}_{2_{C_1}} = 2.00 \tag{4.6}$$

$$\sigma_{1_{C_1}} = desviopadrao\left(X_{1_{C_1}}\right) = 1.12 \tag{4.7}$$

$$\sigma_{2_{C_1}} = desviopadrao\left(X_{2_{C_1}}\right) = 0.54 \tag{4.8}$$

$$cov_{C_1}\left(X_{1_{C_1}}, X_{2_{C_1}}\right) = 0.007 \tag{4.9}$$

$$\rho_{C_1} = \frac{0.007}{\sqrt{1.12^2 0.54^2}} = 0.012 \tag{4.10}$$

Calculando então para C_2:

$$\mu_{1_{C_2}} = \overline{X}_{1_{C_2}} = 4.03 \tag{4.11}$$
$$\mu_{2_{C_2}} = \overline{X}_{2_{C_2}} = 3.95 \tag{4.12}$$
$$\sigma_{1_{C_2}} = desviopadrao\left(X_{1_{C_2}}\right) = 0.54 \tag{4.13}$$
$$\sigma_{2_{C_2}} = desviopadrao\left(X_{2_{C_2}}\right) = 0.53 \tag{4.14}$$
$$cov_{C_2}\left(X_{1_{C_2}}, X_{2_{C_2}}\right) = 0.042 \tag{4.15}$$
$$\rho_{C_2} = \frac{0.007}{\sqrt{1.12^2 0.54^2}} = 0.15 \tag{4.16}$$

3º passo - seleção de um elemento não rotulado

Depois de ter calculado os dados para se determinar a verossimilhança de um elemento não rotulado, nosso próximo passo é selecionar ao acaso um elemento z do conjunto de dados a serem classificados.

4º passo - calcular as verossimilhanças

Para este elemento z selecionado vamos aplicar a fórmula 4.3 com seus respectivos parâmetros calculados no passo 2 para cada classe.

Suponha que $z = 2$ com X_{u_1} e X_{u_2} definidos como na tabela 4.2. Assim, suas verossimilhanças calculadas ficam como mostram as equações 4.17 e 4.18.

$$P\left[X_{u_{2,j}}|C_1\right] = 0.1967 \tag{4.17}$$

$$P\left[X_{u_{2,j}}|C_2\right] = 1.45x10^{-4} \tag{4.18}$$

5º passo - determinar a classe

Agora é simples! Para classificar o elemento $X_{u_{2,j}}$ basta dividirmos uma verossimilhança por outra (ponderadas pelas respectivas evidências) e verificar o resultado, como mostrado na equação 4.19.

$$K = \frac{P\left[C_1|X_{u_{2,j}}\right]}{P\left[C_2|X_{u_{2,j}}\right]} = \frac{P\left[X_{u_{2,j}}|C_1\right]P\left[C_1\right]}{P\left[X_{u_{2,j}}|C_2\right]P\left[C_2\right]} = \frac{0.1967 \times 0.667}{1.45x10^{-4} \times 0.333} \approx 2698 \geq 1 \tag{4.19}$$

Assim, pela equação 4.19 temos que o padrão (elemento) $z = 2$ apresenta $K \geq 1$ significando que ele tem uma probabilidade maior de pertencer à classe 1, sendo então classificado como pertencendo a ela.

6º passo - Próximo elemento

Depois de classificado o elemento $z = 2$ retorna-se ao passo 3 e sorteia-se outro elemento repetindo todo o processo até que todos padrões tenham sido classificados.

Capítulo 5

Resultados

A figura 5.1 mostra como fica a classificação dos pontos utilizando um classificador bayesiano conforme explicado nos capítulos anteriores. Os pontos azuis na figura representam os dados da classe 1, sendo os dados existentes aqueles marcados com um (+) e os novos classificados, marcados com um (.). Os pontos em verde são os da classe 2. Da mesma forma os pontos novos, que foram classificados com a regra de Bayes, estão marcados com um (.) verde. A linha vermelha representa o limite entre as duas classes, definido pela aplicação da regra de Bayes. Entretanto esta separação depende muito das informações à priori e isto será discutido nas conclusões.

A título de ilustração apresentamos na figura 5.3 o resultado da classificação dos pontos não rotulados de um problema semelhante, porém, com número de padrões diferente, mostrado na figura 5.3. Veja como fica a superfície de separação agora que número de padrões da classe 1 está menor que o número de pontos da classe 2.

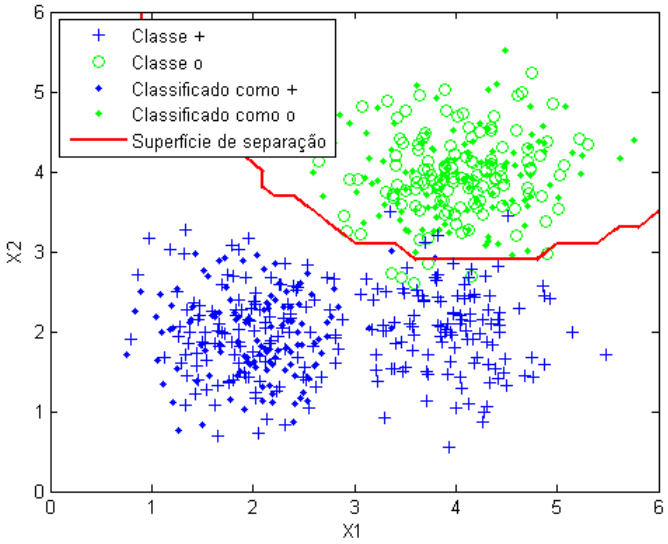


Figura 5.1: Resultado da classificação

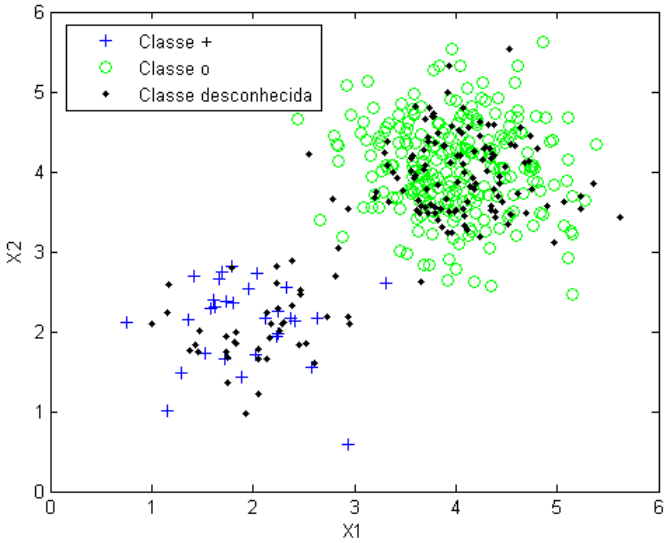


Figura 5.2: Outro problema de classificação

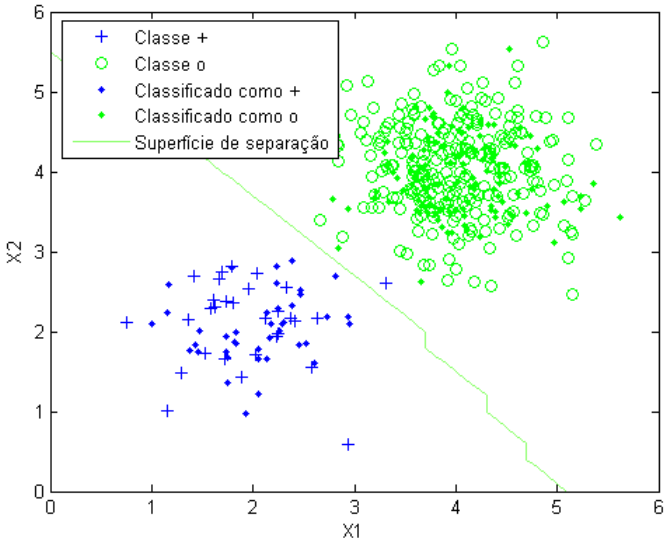


Figura 5.3: Resultado do outro problema de classificação

Capítulo 6

Conclusões

Hoje na literatura existem diversos trabalhos baseados na regra de Bayes e que foram fonte de pesquisa no desenvolvimento deste trabalho da disciplina de Introdução aos Processos Estocásticos. Citamos alguns, como por exemplo, o trabalho de Baback [MJP00] aplica o teorema de Bayes no reconhecimento facial e o trabalho de Bishop [Bis99] que desenvolve uma ferramenta PCA (principal component analysis) Bayesiano.

Mas apesar de ser aplicado nos mais variados problemas, este classificador tem algumas limitações que devem ser levadas em conta. Primeiro é o fato de que sem informação à priori, ele não tem como ser utilizado. Ainda nesta linha de raciocínio, quanto mais dados sobre as classes existir, melhores serão as classificações obtidas.

Entretanto há um problema que deve ser levado em conta. Este classificador é fortemente influenciado pelo balanceamento das classes. Dependendo do número de elementos de cada classe o resultado pode ser diferente. Basta comparar os gráficos de resultados dos dois exemplos apresentados no ítem 5, nas figuras 5.1 e 5.3. A superfície de separação tende a "envolver" e se aproximar da classe com menor número de padrões. Isso pode inserir mais erros de classificação na região de separação das classes, e até mesmo aumentar o erro de classificação da classe menor. Na figura 5.1 a superfície de separação tende a envolver a classe verde enquanto que na figura 5.3 a superficie de separação tende a "envolver" a classe azul.

Desta forma se faz necessário uma criteriosa avaliação da aplicabilidade deste método aos problemas de classificação, uma vez que ele é extremamente dependente dos dados e informações à priori.

Referências Bibliográficas

[Bis99] Christopher M. Bishop. Variational principal components. In *In Proceedings Ninth International Conference on Artificial Neural Networks, ICANN'99*, pages 509–514, 1999.

[Bra09] Antônio Braga. Notas de aula da disciplina reconhecimento de padrões, 2009.

[Kay06] Steven Kay. *Intuitive probability and random processes using MATLAB*. Kluwer Academis Publishers, 2006.

[MJP00] Baback Moghaddam, Tony Jebara, and Alex Pentland. Bayesian face recognition. *Pattern Recognition*, 33:1771–1782, 2000.

[Pap91] Athanasios Papoulis. *Probability, Random Variables and Stochastic Processes*. McGraw-Hill Companies, February 1991.