Reconhecimento de Padrões em Imagens
Aprendizagem Bayesiana

Dainf - UTFPR

Profa. Leyza Baldo Dorini

Teoria de decisão Bayesiana

É uma abordagem estatística para o problema de reconhecimento de padrões. Baseia-se na quantificação do custo/benefício entre várias decisões de classificação baseada nos custos e probabilidades associados a elas.

Suposições:

- O problema é posto em termos probabilísticos.
- Todas as probabilidades relevantes são conhecidas. Essa situação raramente ocorre na prática, mas permite obter o classificador (Bayesiano) ótimo e comparar com outros classificadores.

Exemplo: classificação do peixe

Considere novamente o problema de classificação entre salmão e robalo.

Vamos definir uma variável aleatória, ω, para cada classe:

$$\omega = \omega_1 \quad \text{para o robalo}$$

$$\omega = \omega_2 \quad \text{para o salmão}$$

Motivação

Treinamento

- Suponha que seja extraído um vetor de características, $\mathbf{x}$, a partir de um conjunto de treinamento associado a uma determinada classe, ω_i.
- A probabilidade condicionada à classe (*likelihood*), denotada por $p(\mathbf{x}|\omega_i)$, nos diz quão frequentemente amostras da classe ω_i possuem as características $\mathbf{x}$.

Teste

- Na classificação buscamos a probabilidade *a posteriori*, $p(\omega_i|\mathbf{x})$, ou seja, dado que uma nova amostra tem as características $\mathbf{x}$, qual a probabilidade de que ela pertença à classe ω_i?
- Para essa estimativa, a probabilidade *a priori* também é considerada, a qual representa o conhecimento sobre um determinado domínio.

Probabilidades *a priori* (*priors*)

Prior: conhecimento de quão provável é a ocorrência de uma amostra de uma determinada classe (antes que possamos de fato observar as amostras).

- No caso do peixe, é a probabilidade de observarmos um salmão ou um robalo.
- As *priors* podem variar dependendo da situação.
  - Se a quantidade de salmões e robalos é a mesma, as *priors* são iguais (ou uniformes).
  - Contudo, dependendo da estação, um dos peixes pode ocorrer com mais frequência.
- A notação é dada por $P(\omega = \omega_1)$ ou $P(\omega_1)$ (para o robalo).
  - As *priors* devem devem ter as propriedades de exclusividade e exaustividade:

$$\sum_{i=1}^{c} P(\omega_i) = 1$$

em que c denota a quantidade de classes.

Regra de decisão

Uma **regra de decisão** determina qual ação tomar com base em uma dada entrada.

- Sugira uma regra de decisão para o seguinte contexto:
  - a única informação disponível é a *prior*.
  - o custo de qualquer classificação incorreta é igual.
- Sugestão: decida ω_1 se $P(\omega_1) > P(\omega_2)$. Caso contrário, decida ω_2.
  - Parece razoável, mas escolhe sempre o mesmo peixe.
  - Se as *priors* são uniformes, o desempenho é fraco.
  - Contudo, para este contexto, é a melhor regra (!). O erro é dado por $P(error) = \min\{P(\omega_1), P(\omega_2)\}$.

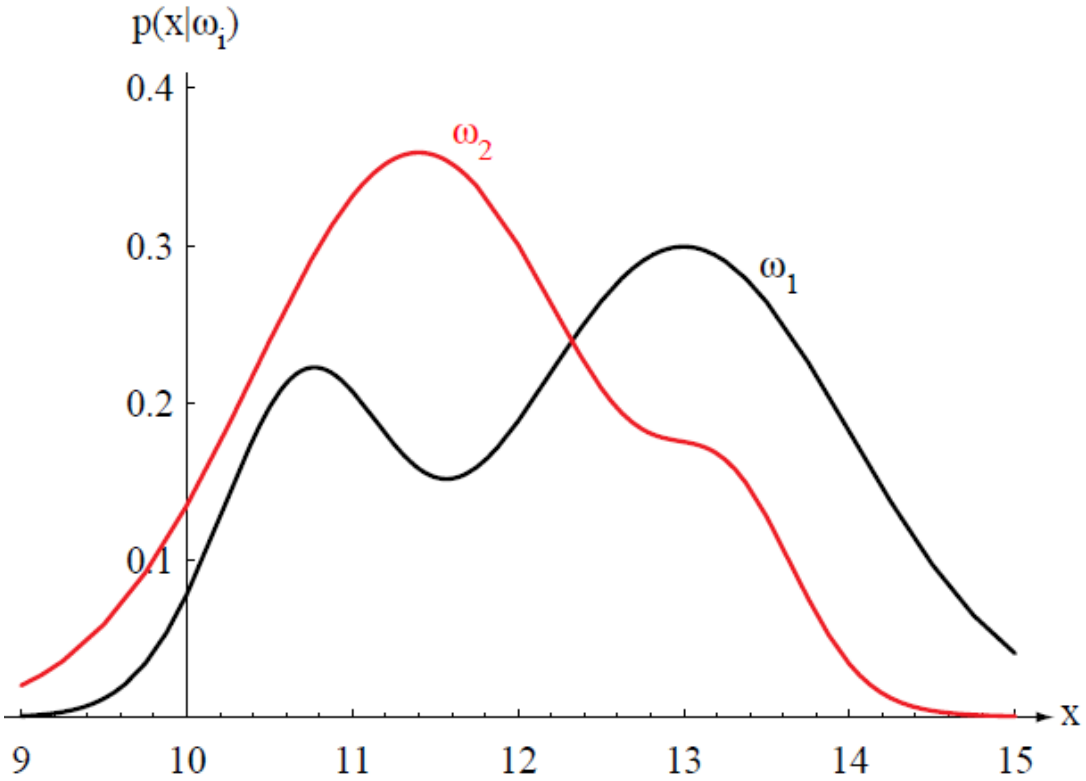
Características e espaços de características

- Tipicamente, mais informações são utilizadas para tomada de decisões. Para tal, são extraídas características (variável que pode ser observada). Exemplos: comprimento, largura, brilho, etc.
- Um espaço de características consiste em um conjunto a partir do qual podemos obter uma amostra.
- Seja x uma característica única e $\mathbf{x}$ um vetor de características, com $\mathbf{x} \in \mathbb{R}^d$, em que d é a dimensão do espaço de características. Por simplicidade, vamos assumir que as características são variáveis contínuas.

Densidade condicional à classe (*likelihood*)

A função de probabilidade de densidade condicional à classe é a função de densidade probabilidade para a característica **x** dado que o estado observado foi ω: $p(\mathbf{x}|\omega)$.

Exemplo: densidade de probabilidade ao medir uma determinada característica x (tamanho) dado que o padrão pertence à classe ω_i (para o problema salmão/robalo).



Evidência

- A função densidade de probabilidade para todas as amostras (também denominada "evidência") é dada por:

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)$$

em que c é a quantidade de classes.

- Pode ser interpretada como a frequência que uma determinada característica é encontrada.

Probabilidade *a posteriori*

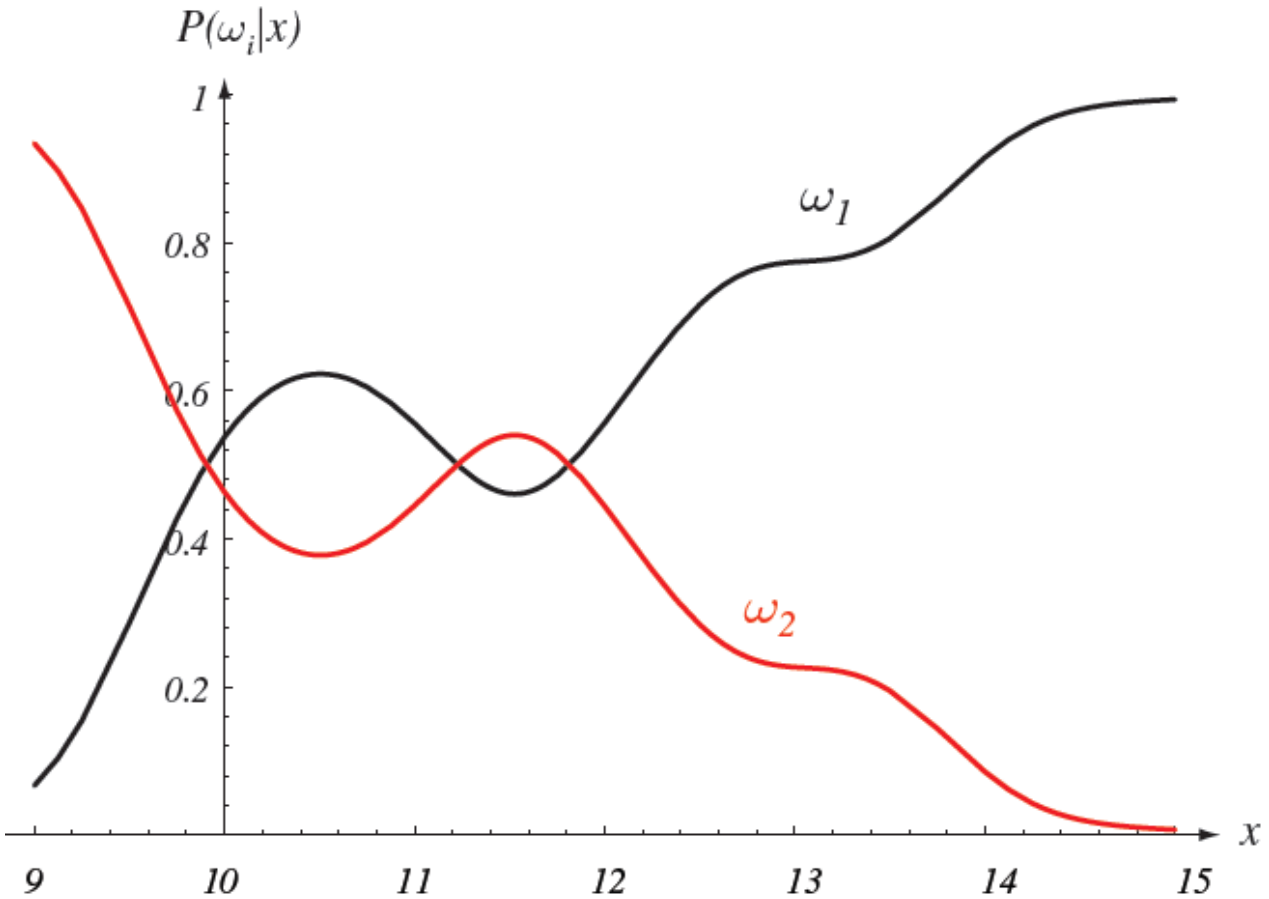
- Utilizando o Teorema de Bayes, é possível estimar as probabilidades *a posteriori* com base nas *priors* e nas respectivas funções de densidade de probabilidade condicionais (*likelihoods*).
- A fórmula é dada por:

$$P(\omega|\mathbf{x}) = \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})}$$

O termo $p(\mathbf{x})$ atua como um fator de escala para normalizar a densidade.

Exemplo: probabilidades *a posteriori*

Para $P(\omega_1) = 2/3$ e $P(\omega_2) = 1/3$, a posterior é dada por:



Por exemplo, para $x = 14$, a probabilidade de ω_1 é 0.92 e de ω_2 é 0.08 (observe que, para cada x, as probabilidades somam 1).

Estrutura de um classificador Bayesiano

1. Treinamento: determinar $p(\mathbf{x}|\omega_i)$ para cada classe. Estimar o conhecimento *a priori*: medir (ou estimar) $P(\omega_i)$ na população geral.

2. Classificação:
   - Obter o vetor de características $\mathbf{x}$ para o caso de teste.
   - Calcular a probabilidade *a posteriori* $P(\omega_i|\mathbf{x})$ para cada classe.
   - Atribuir ao caso de teste a classe com maior $P(\omega_i|\mathbf{x})$.

Para uma dada observação $\mathbf{x}$, a decisão é governada pela *posterior*:

$$\omega^* = \arg\max_i P(\omega_i|\mathbf{x})$$

Probabilidade de erro

- A probabilidade de erro é dada por:

$$P(error|\mathbf{x}) = \begin{cases} P(\omega_1|\mathbf{x}), & \text{se a decisão for } \omega_2 \\ P(\omega_2|\mathbf{x}), & \text{se a decisão for } \omega_1 \end{cases}$$

- Portanto, podemos minimizar a probabilidade de erro escolhendo a seguinte posterior:

$$\text{Decida } \omega_1 \text{ se } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}), \text{ou } \omega_2 \text{ caso contrário.}$$

$$\text{Decida } \omega_1 \text{ se } p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2), \text{ou } \omega_2 \text{ caso contrário.}$$

$$\text{Decida } \omega_1 \text{ se } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \quad (\textit{likelihood ratio})$$

- Observe que, se as *likelihoods* são iguais, a decisão depende das *priors* (e vice-versa). A pdf $p(\mathbf{x})$ não é levada em consideração (fator de normalização).

Importante

A tomada de decisão depende tanto da função de densidade de probabilidade condicional (*likelihoods*) quanto das *priors*. A regra de decisão de Bayes combina essas duas informações para obter a mínima probabilidade de erro.

Cabe ressaltar que essa abordagem também minimiza a probabilidade média de erro:

$$P(error) = \int_{-\infty}^{\infty} P(error|\mathbf{x})p(\mathbf{x})d\mathbf{x}$$

dado que a integral será minimizada ao garantir que cada $P(error|\mathbf{x})$ é o menor possível.

Exemplo

- Considere que nesta época do ano a probabilidade de pescar salmão é maior que a de pescar robalo, sendo $P(salmao) = 0.82$ e $P(robalo) = 0.18$.
- A característica adicional considerada é o brilho, sendo que 49.5% dos salmões e 85% dos robalos tem intensidade clara.
- A probabilidade de um dado caso de teste ser um salmão dado que tem intensidade clara é dada por:

$$P(S|C) = \frac{P(S)P(C|S)}{P(C)} = \frac{0.82 * 0.495}{0.82 * 0.495 + 0.18 * 0.85} = 0.726.$$

Função de perda (*Loss function*)

Dados:

- Um conjunto de c classes, $\{\omega_1, \ldots, \omega_c\}$
- Um conjunto de a possíveis ações, $\{\alpha_1, \ldots, \alpha_a\}$

A **função de perda** $\lambda(\alpha_i|\omega_j)$ representa a perda resultante da escolha da ação α_i quando a classe é ω_j.

Exemplo: a função de perda 0-1 é dada por

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases} \qquad i, j = 1, 2, \ldots, c$$

ou seja, a perda é 0 para uma decisão correta e 1 para cada decisão incorreta.

Função de perda: exemplo

Com a função de perda, é possível mensurar de forma diferente a perda para cada tipo de classificação errada. Por exemplo:

$$\lambda(decision = healthy|patient = sick) \gg \lambda(sick|healthy)$$

Isso pode ser formalizado utilizando-se uma matriz de perda: L_{kj} representa a perda para a decisão C_j quando a resposta correta seria C_k. Exemplo:

$$L_{diagnostico} = \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix}$$

A *likelihood ratio* seria:

$$\text{Decida } \omega_1 \text{ se } \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12}\lambda_{22}P(\omega_2)}{\lambda_{21}\lambda_{11}P(\omega_1)}$$

Risco condicional (*expected loss*)

O risco condicional é definido como:

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

Com isso, dada uma observação **x**, podemos minimizar a perda selecionando a ação que minimiza o risco condicional. **Isso é o que a Regra de Decisão de Bayes faz**.

Classificador ótimo

A regra de decisão Bayesiana nos fornece um método para minimizar o risco total, escolhendo-se a ação que minimiza o risco condicional:

$$\begin{aligned} \alpha* &= \arg\min_{\alpha_i} R(\alpha_i|\mathbf{x}) \\ &= \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x}) \end{aligned} \tag{1}$$

Com isso, temos o classificador ótimo.

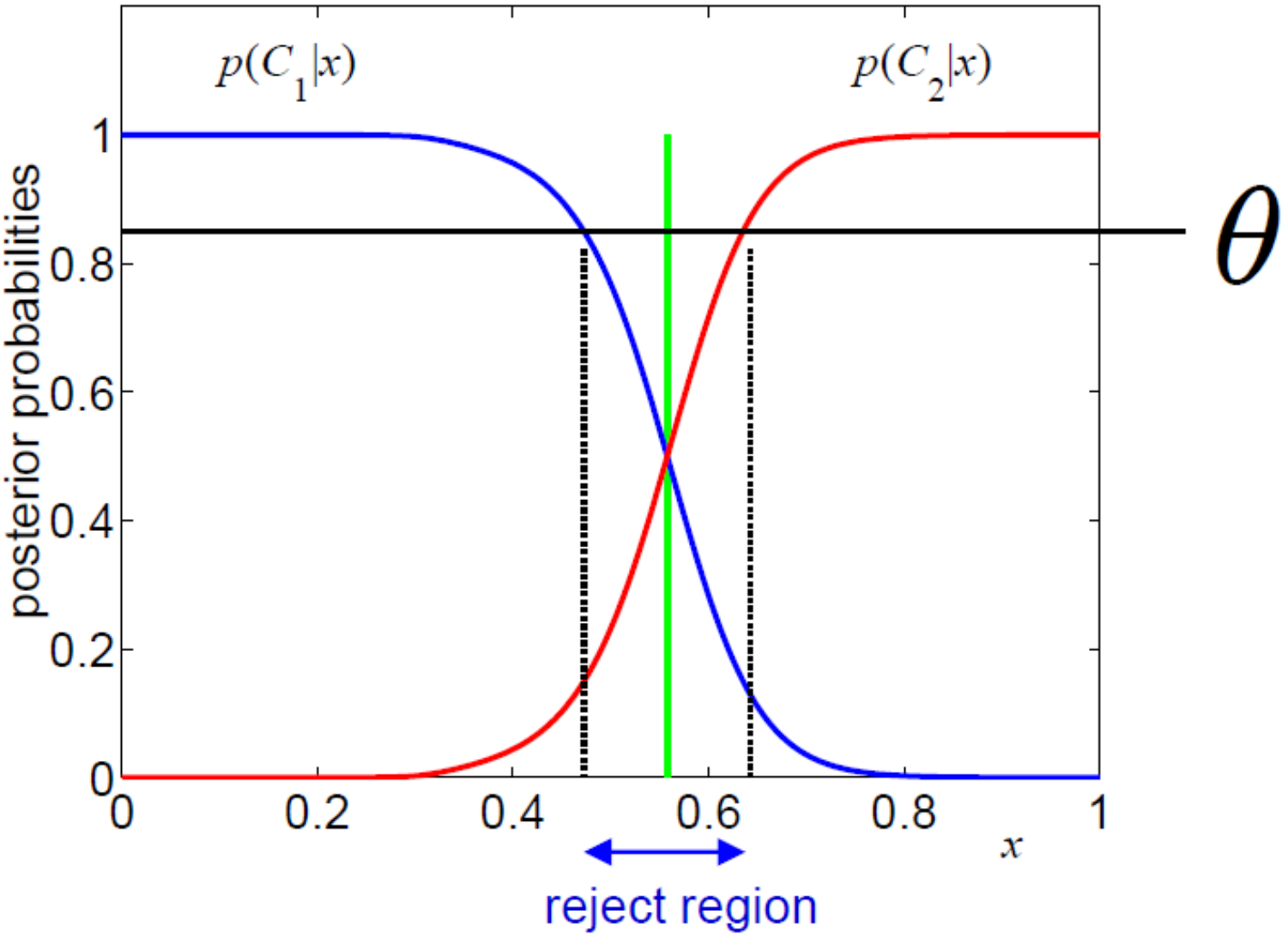
Classificador ótimo

Se o classificador Bayesiano é um classificador ótimo, então por que outros classificadores são utilizados?

- O motivo é que o classificador de Bayes só pode ser executado se a probabilidade *a priori* e a *likelihood* forem conhecidas, o que geralmente não ocorre.
- Em problemas práticos, na fase de treinamento são utilizados métodos de estimação dessas probabilidades. Entretanto, quando a distribuição das classes possui formas "complicadas" e descontínuas, o preço computacional desses métodos torna-se muito alto quando se deseja obter uma representação precisa dessas probabilidades.

Reject option

Em algumas ocasiões, pode ser mais adequado rejeitar a decisão automática, deixando a resposta final a cargo de um especialista.



Naive Bayes

Classificador Naive Bayes

- Classificadores baseados em árvores de decisão
  - Dados geram regras de decisão.
  - Os novos exemplos são classificados com base nestas regras.
- Classificadores Naive Bayes
  - Dados geram modelos probabilísticos sobre relacionamentos entre classes e atributos.
  - Os novos exemplos são classificados via inferência baseada nestes modelos.

Distribuição conjunta

A forma mais geral para descrever relacionamentos entre variáveis aleatórias. Considere o exemplo: $P(Temperature, Wind, Play)$

Temperature	Wind	PlayTennis	Probability
Hot	Weak	No	0.1
Hot	Weak	Yes	0
Hot	Strong	No	0.2
Hot	Strong	Yes	0.3
Cool	Weak	No	0.1
Cool	Weak	Yes	0.2
Cool	Strong	No	0.1
Cool	Strong	Yes	0

Cada combinação de valores possui uma probabilidade. Além disso, a soma de todas as probabilidades é 1.

Distribuição conjunta e classificação

Para classificar com base em uma distribuição conjunta, calcula-se a probabilidade. Por exemplo, deve-se jogar tênis em um dia quente com vento forte?

$$
\begin{aligned}
P(Play = y | T = h, W = s) &= \frac{P(Play = y, T = h, W = s)}{P(T = h, W = s)} \\
&= \frac{0.3}{0.2 + 0.3} = 0.6
\end{aligned}
$$

Resposta: jogar tênis.

Distribuição conjunta e classificação

É possível trabalhar com um subconjunto de atributos. Por exemplo, deve-se jogar tênis quando o vento está fraco?

$$
\begin{aligned}
P(Play = y | W = w) &= \frac{P(Play = y, W = w)}{P(W = w)} \\
&= \frac{0 + 0.2}{0.1 + 0 + 0.1 + 0.2} = 0.67
\end{aligned}
$$

Resposta: jogar tênis.

O caso geral

- Suponha que tenhamos uma distribuição conjunta:

$$P(A_1, A_2, \ldots, A_n, C)$$

  com atributos A_i e a variável de classe C.
- Para um novo exemplo com valores de atributos $a_1, a_2, \ldots, a_n$ e para cada possível valor v_j, calcula-se a probabilidade:

$$P(C = v_j | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$$

- E atribui-se o exemplo à classe v_j^* com a maior probabilidade *a posteriori* (MAP - *Maximum a Posteriori Estimation*).

$$c_j^* = \underset{v_j \in classlabels}{\arg\max} \; P(C = v_j | A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n)$$

Limitações

- Supondo 100 atributos binários e uma classe com saída binária, seriam 2^{101} entradas para a distribuição de probabilidade conjunta.
- Pode ser necessário lidar com muitos dados, havendo o risco de *overfitting*.

A modelagem baseada em Naive Bayes reduz a quantidade de parâmetros do modelo assumindo **independência**.

Independência

Duas variáveis aleatórias X e Y são marginalmente independentes se, para qualquer estado x de X e qualquer estado y de Y:

$$P(X = x|Y = y) = P(X = x)$$

ou seja, aprender o valor de Y não dá nenhuma informação sobre X e vice-versa.

Exemplos:

- X: resultado de jogar uma moeda justa pela primeira vez. Y: resultado de jogar uma moeda justa pela segunda vez.
- X: resultado das eleições dos EUA. Y: sua nota final nesta disciplina.

Você consegue propor um contra-exemplo?

Independência condicional

Duas variáveis aleatórias X e Y são condicionalmente independentes dada uma terceira variável Z se:

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z)$$

ou seja, aprender o valor de Y não dá nenhuma informação sobre X e vice-versa.

Isso significa que, se soubermos o estado de Z, aprender o estado de Y não dá informação adicional sobre X. Mesmo que Y tenha alguma informação sobre X, ela já está em Z.

Exemplo de independência condicional

Suponha uma sacola com 100 moedas. Delas, 10 são viciadas, resultando em cara 80% das jogadas. As demais moedas são justas.

Considere o seguinte experimento: tire aleatoriamente uma moeda da sacola e jogue-a algumas vezes. X_i denota o resultado da i-ésima jogada e Y representa se a moeda é viciada ou não.

Exemplo: X_i não são independentes entre si

Neste caso:

- Se saírem 9 caras em 10 jogadas, a moeda é provavelmente viciada. Portanto, a próxima jogada tem mais probabilidade de ser cara do que coroa.
- Aprender o valor de X_i dá alguma informação sobre o fato de a moeda ser viciada, o que dá (em termos) alguma informação sobre X_j.

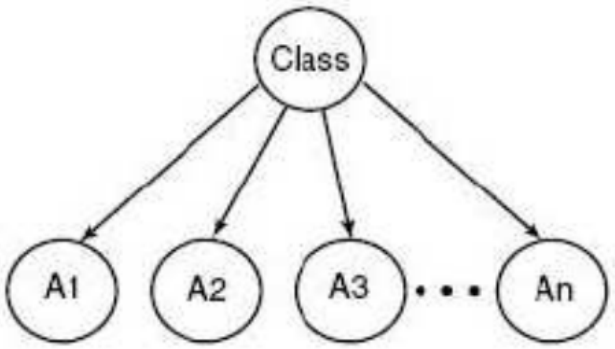
Exemplo: X_i são condicionalmente independentes entre si dado Y

- Se a moeda não é viciada, a probabilidade de obter uma cara em uma jogada é $1/2$, independentemente dos outros resultados.
- Se a moeda é viciada, a probabilidade de obter uma cara é 80%, independentemente das outras jogadas.

Em suma: se já sabemos se a moeda é ou não viciada, aprender o valor de X_i não dá nenhuma informação adicional sobre X_j.

O modelo Naive Bayes

Esse modelo assume que os atributos são mutuamente independentes entre si, dada a variável que representa a classe. Graficamente,



Em outras palavras: o classificador é baseado na suposição de que os valores dos atributos são condicionalmente independentes dado o valor alvo.

A distribuição conjunta é dada por:

$$P(C, A_1, A_2, \ldots, A_n) = P(C) \prod_{i=1}^{n} P(A_i|C)$$

Aprendizagem com Naive Bayes

- A partir dos dados, é preciso estimar: $P(C)$, $P(A_1|C), \ldots, P(A_n|C)$.
- São simples de calcular:

$$\hat{P}(C = v_j) = \frac{\#\text{ de exemplos com } C = v_j}{\#\text{ total de exemplos}}$$

$$\hat{P}(A_i = a_i|C = v_j) = \frac{\#\text{ de exemplos com } C = v_j \text{ e } A_i = a_i}{\#\text{ de exemplos com } C = v_j}$$

Embora simples, estes são os parâmetros MLE (*Maximum Likelihood Estimates*).

Classificação com Naive Bayes

Para um novo exemplo valores de atributos $a_1, a_2, \ldots, a_n$, a seguinte classe será atribuída:

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(C = v_j) \prod_{i=1}^{n} \hat{P}(A_i = a_i | C = v_j)$$

Exemplo: jogar tênis

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Exemplo: jogar tênis

O primeiro passo consiste em construir o modelo de probabilidades condicionais Naive Bayes. A seguinte tabela mostra a frequência de diferentes evidências.

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								

Por exemplo, existem duas instâncias mostrando $(outlook = sunny)$ para $(play = y)$

Exemplo: jogar tênis

Após definir todas as frequências é necessário calcular todas as probabilidades condicionais e as probabilidades *a priori*.

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Por exemplo:

- $P(outlook = sunny|play = y) = 2/9$
- $P(play = y) = 9/14$

Exemplo: jogar tênis

$$P(PlayTennis = y) = 9/14$$
$$P(Outlook = sunny|y) = 2/9$$
$$P(Outlook = overcast|y) = 4/9$$
$$P(Outlook = rain|y) = 3/9$$
$$P(Temp = hot|y) = 2/9$$
$$P(Temp = mild|y) = 4/9$$
$$P(Temp = cool|y) = 3/9$$
$$P(Humidity = high|y) = 3/9$$
$$P(Humidity = normal|y) = 6/9$$
$$P(Wind = strong|y) = 3/9$$
$$P(Wind = weak|y) = 6/9$$

$$P(PlayTennis = n) = 5/14$$
$$P(Outlook = sunny|n) = 3/5$$
$$P(Outlook = overcast|n) = 0/5$$
$$P(Outlook = rain|n) = 2/5$$
$$P(Temp = hot|PlayTennis = n) = 2/5$$
$$P(Temp = mild|n) = 2/5$$
$$P(Temp = cool|n) = 1/5$$
$$P(Humidity = normal|n) = 1/5$$
$$P(Humidity = high|n) = 4/5$$
$$P(Wind = strong|n) = 3/5$$
$$P(Wind = weak|n) = 2/5$$

Exemplo de classificação

Para o caso ($Sunny$, $Cool$, $High$, $Strong$), deve jogar tênis? A inferência é:

$$P(y)P(sunny|y)P(cool|y)P(high|y)P(strong|y) = 0.005$$

$$P(n)P(sunny|n)P(cool|n)P(high|n)P(strong|n) = 0.021$$

Conclusão: não jogar!

```
P(Yes|E) = P(Outlook = sunny | Yes) X
P(Temp = cool|Yes) X
P(Humidity = high|Yes) X
P(Wind = strong|Yes) X
P(Yes) =
= (2/9 * 3/9 * 3/9 * 3/9 * 9/14) = 0.005
```

Técnica de Suavização

- Em alguns casos, nenhuma das instâncias de treinamento com alvo v_j tem o atributo a_i (por exemplo, $P(outlook = overcast|play = No) = 0/5$). Com isso, temos que $\hat{P}(a_i|v_j) = 0$ e, portanto: $\hat{P}(v_j)\prod_i \hat{P}(a_i|v_j) = 0$.
- Com a suavização, adiciona-se 1 a cada caso (*Laplace Smoothing/Correction*)

$$\hat{P}(A_i = a_i|C = v_j) = \frac{(\#\text{ de exemplos com } C = v_j \text{ e } A_i = a_i) + 1}{(\#\text{ de exemplos com } C = v_j) + |C|}$$

Relembrando...

Relembrando: probabilidade

Em um modelo onde os resultados são equiprováveis, o espaço amostral é um conjunto finito Ω e a medida de probabilidade é proporcional à quantidade de resultados que fazem parte de um dado evento:

$$P(B) = \frac{\#B}{\#\Omega}$$

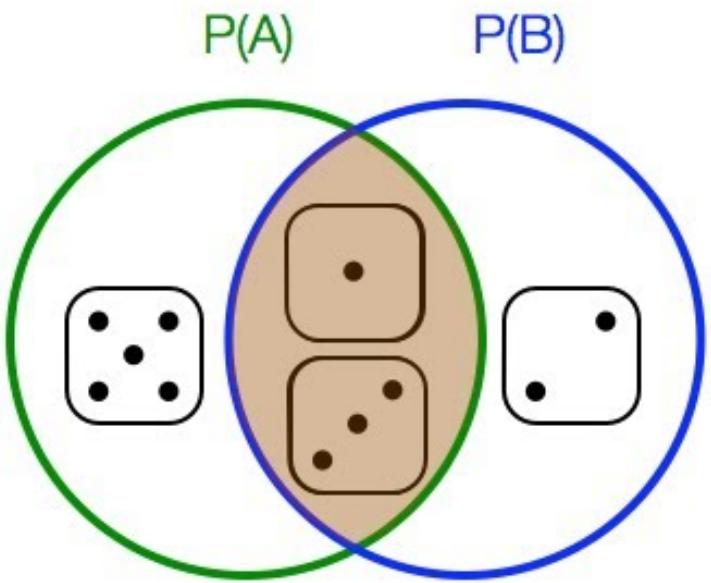
Exemplo: Imagine o sorteio de uma carta em um baralho. Queremos saber a probabilidade de um jogador tirar $4\clubsuit$, $7\heartsuit$, $A\spadesuit$ ou $7\diamondsuit$, evento que será denotado por B. Temos:

$$P(B) == \frac{\#B}{\#\Omega} = \frac{4}{52} \approx 8\%$$

Relembrando: probabilidade condicional

Considere os eventos:

- B = jogar um dado e o valor resultante ser menor que 4...
- A = ... sabendo que o valor é ímpar.



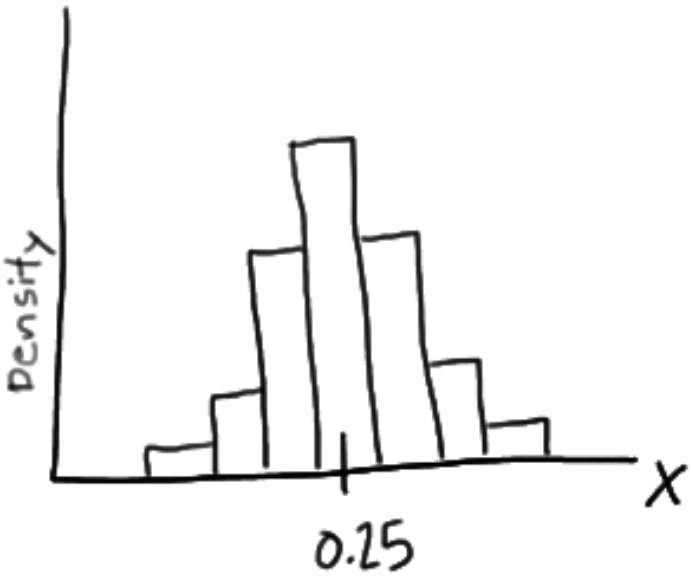
A probabilidade condicional $P(B|A)$ é dada por:

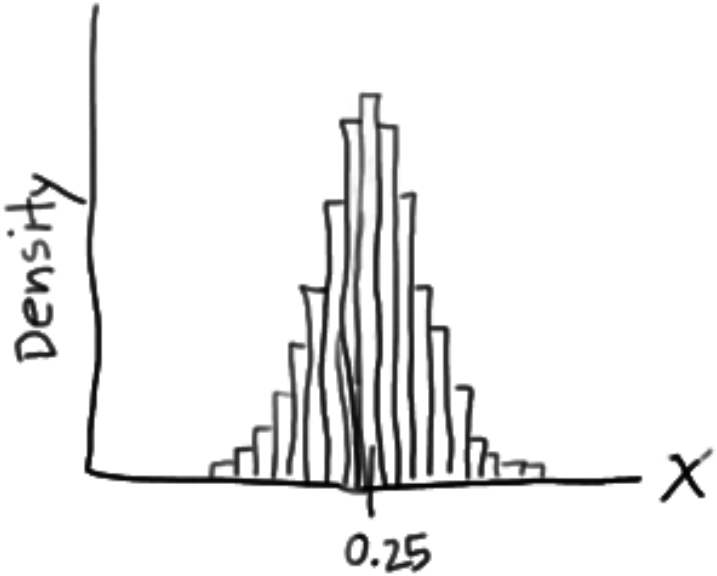
$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

Relembrando: função densidade de probabilidade

Suponha o seguinte exemplo: um rede de sanduíches diz que o hambúrguer de um lanche pesa 0.25 libras. Contudo, esse valor não é exato, ou seja, uma amostra pode pesar 0.23 e a outra 0.27 libras. Qual a probabilidade de que uma amostra pese entre 0.2 e 0.3 libras?

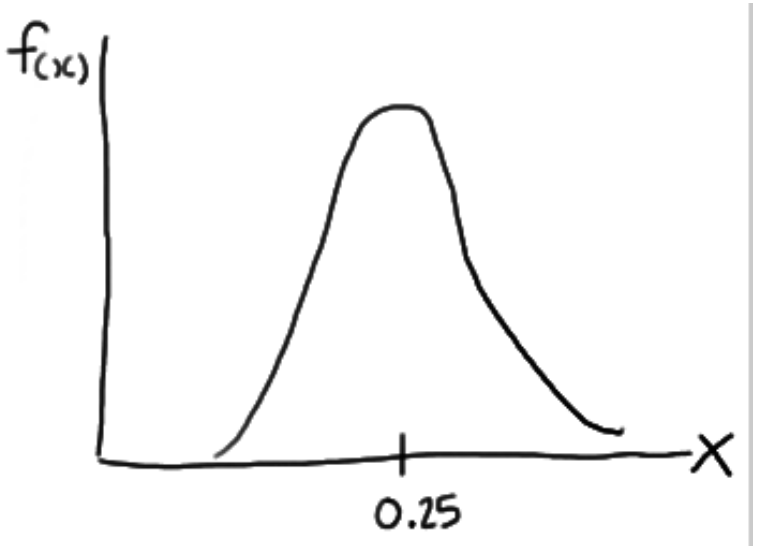
1. Inicialmente, criamos um histograma de densidade a partir do peso de 100 amostras selecionadas aleatoriamente.





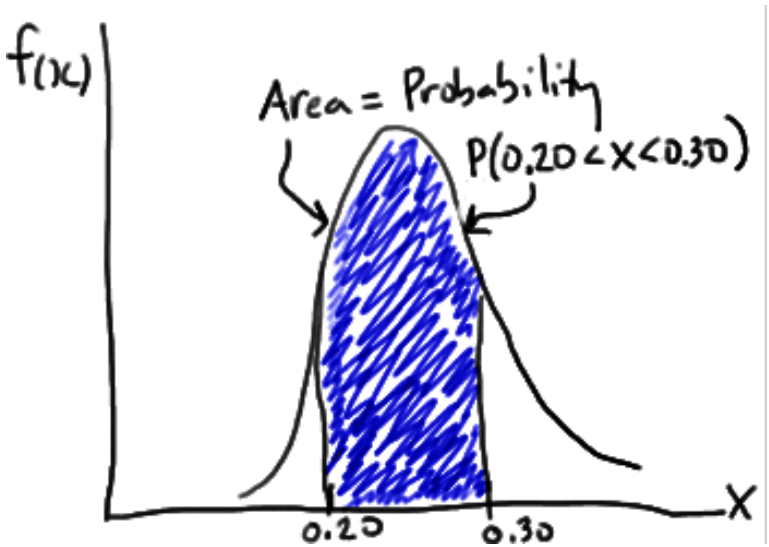
Relembrando: função densidade de probabilidade

2 Suponha que os intervalos fiquem tão pequenos de tal forma que a densidade de probabilidade não seja representada por um histograma de densidades, mas sim por uma "curva": a função densidade de probabilidade.



Relembrando: função densidade de probabilidade

3 Estimar a probabilidade que uma variável aleatória X caia em um determinado intervalo consiste em encontrar a área sob a curva dentro de tal intervalo. Para o exemplo:



Relembrando: função densidade de probabilidade

Seja X uma variável aleatória contínua.

Definition

A **função densidade de probabilidade** (pdf) de X é a função $f(x)$ tal que, para qualquer $a \leq b$

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

As seguintes propriedades também são válidas:

1. $f(x) \geq 0$ para todo x
2. $\int_{-\infty}^{\infty} f(x)dx = 1$

Relembrando: Bayes

Por definição,

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

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

Com isso, temos que:

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

Portanto:

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

Exemplo A

Um médico sabe que a meningite causa torcicolo em 50% dos casos. Porém, o médico sabe que a meningite atinge 1/50000 e também que a probabilidade de se ter torcicolo é de 1/20. Usando Bayes pra saber a probabilidade de uma pessoa ter meningite dado que ela está com torcicolo.

Temos que:

- $P(T|M) = 0.5$
- $P(M) = 1/50000$
- $P(T) = 1/20$

$$P(M|T) = \frac{P(M)P(T|M)}{P(T)} = \frac{1/50000 \times 0.5}{1/20} = 0.0002$$

Exemplo B

Um armário tem duas gavetas, A e B. A gaveta A tem 2 meias azuis e 3 meias pretas, e a gaveta B tem 3 meias azuis e 3 meias vermelhas. Abre-se uma gaveta ao acaso e retira-se uma meia ao acaso da gaveta escolhida. Problema: Qual a probabilidade de escolher-se uma meia azul?

Exemplo B

A solução utiliza a lei da probabilidade total. Começamos pelos valores conhecidos: $P(A) = P(B) = \frac{1}{2}, P(azul|A) = \frac{2}{5}$ e $P(azul|B) = \frac{3}{6}$. Assim,

$$\begin{aligned} P(azul) &= P(A)P(azul|A) + P(B)P(azul|B) \\ &= \frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times \frac{3}{6} \\ &= \frac{9}{20} \end{aligned} \tag{2}$$

Exemplo C

No problema do exemplo B, sabendo-se que uma meia azul foi retirada, qual a probabilidade de ter sido aberta a gaveta A? Pela **Fórmula de Bayes** temos:

$$\begin{aligned} P(A|azul) &= \frac{P(A)P(azul|A)}{P(A)P(azul|A) + P(B)P(azul|B)} \\ &= \frac{\frac{1}{5}}{\frac{9}{20}} \\ &= \frac{4}{9} \end{aligned} \tag{3}$$