APRENDIZADO DE MÁQUINA - AM

1-Rule, Naive Bayes e KNN

Tópicos

- Classificação
- Algoritmo 1R
- Algoritmos Bayesianos
  - Naive Bayes
  - Seleção de Atributos (Wrapper Naive Bayes)
  - Noções de Redes Bayesianas
- Algoritmo KNN

Classificação

- Técnica classifica novas entradas (padrões) em uma ou mais dentre diferentes classes discretas
  - Número definido de classes
    - frequentemente apenas duas: classificação binária
- Exemplos
  - Diagnóstico, Análise de crédito, ...
- Existem vária técnicas, para diferentes contextos
  - Sucesso de cada método depende do domínio de aplicação e do problema particular em mãos
  - Técnicas simples frequentemente funcionam muito bem !

Exemplo: Problema Weather

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

1 Rule – 1R

- 1R: Aprende uma árvore de decisão de um nível
  - Todas as regras usam somente um atributo
    - Atributo deve ser (ou ser transformado em) categórico
      - Paradigma simbólico

- Versão Básica:
  - Um ramo para cada valor do atributo
  - Para cada ramo, atribuir a classe mais frequente
  - Para cada ramo, calcular a taxa de erro de classificação:
    - proporção de exemplos que não pertencem à classe mais frequente
  - Escolher o atributo com a menor taxa de erro de classificação

Pseudo-Código para o 1R:

- Para cada atributo:
  - Para cada valor do atributo gerar uma regra como segue:
    - Contar a frequência de cada classe;
    - Encontrar a classe mais frequente*;
    - Formar uma regra que atribui a classe mais frequente a este atributo-valor;
  - Calcular a taxa de erro de classificação das regras;
- Escolher as regras com a menor taxa de erro de classificação.

*** Empates na classe mais frequente podem ser decididos aleatoriamente**

Exemplo: Problema Weather

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Attribute	Rules	Errors	Total Errors
Outlook	Sunny → No	2/5	4/14
	Overcast → Yes	0/4	
	Rainy → Yes	2/5	
Temp	Hot → No*	2/4	5/14
	Mild → Yes	2/6	
	Cool → Yes	1/4	
Humidity	High → No	3/7	4/14
	Normal → Yes	1/7	
Windy	False → Yes	2/8	5/14
	True → No*	3/6	

- 1R seria composto ou das 3 regras para Outlook ou das 2 Regras para Humidity: decisão poderia ser feita, por ex., de acordo com o desempenho em um outro conjunto de dados (dados de teste)

Discussão para o 1R:

- 1R foi descrito por Holte (1993)
  - Contém uma avaliação experimental em 16 bases de dados;
  - Regras simples do 1R não são muito piores do que árvores de decisão mais complexas !
- Interessado em resolver um problema de classificação ou em propor um novo classificador?
  - Experimente o 1R primeiro!
- Implementado no software Weka

Holte, Robert C., Very Simple Classification Rules Perform Well on Most Commonly Used Datasets, Machine Learning 11 (1), pp. 63-90, 1993.

Nota

- 1R pode ser visto como um método de seleção de atributos do tipo embarcado
- Mas pode também ser utilizado como filtro para outros classificadores que não dispõem de seleção de atributos embarcada
  - p. ex. K-NN

Exercício

- Obter um classificador 1R para os dados:

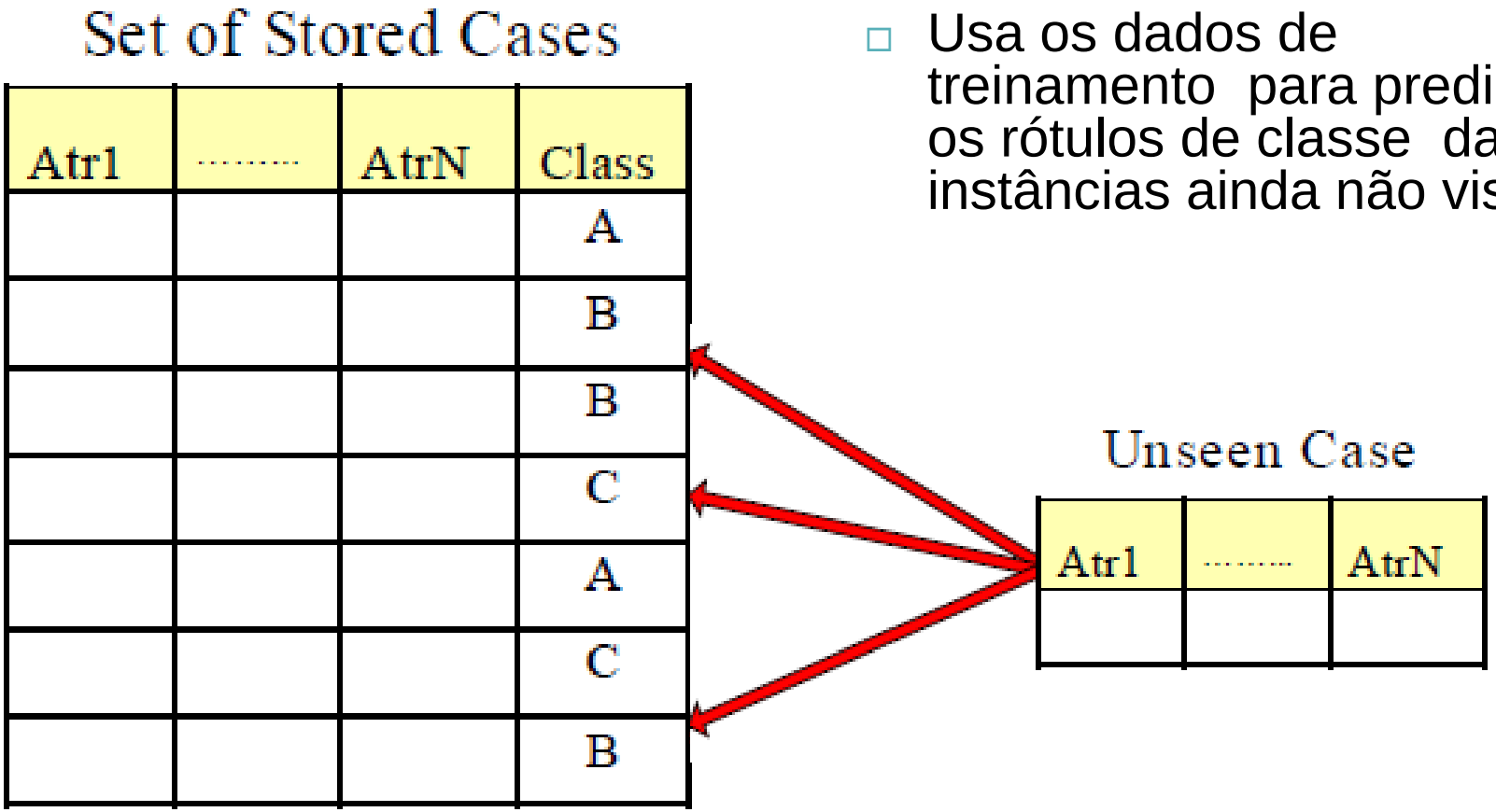
Febre	Enjôo	Mancha	Dor	Diagnóstico
Sim	Sim	Não	Sim	Não
Não	Sim	Não	Não	Sim
Sim	Sim	Sim	Não	Sim
Sim	Não	Não	Sim	Não
Sim	Não	Sim	Sim	Sim
Não	Não	Sim	Sim	Não

K-NN

- O Algoritmo K-NN (K-Vizinhos-Mais-Próximos ou K-Nearest-Neighbors do inglês) é um dos mais simples e bem difundidos algoritmos do paradigma baseado em instâncias

Classificadores Baseados em Instância

- Armazena dados de treinamento
- Usa os dados de treinamento para predizer os rótulos de classe das instâncias ainda não vistas

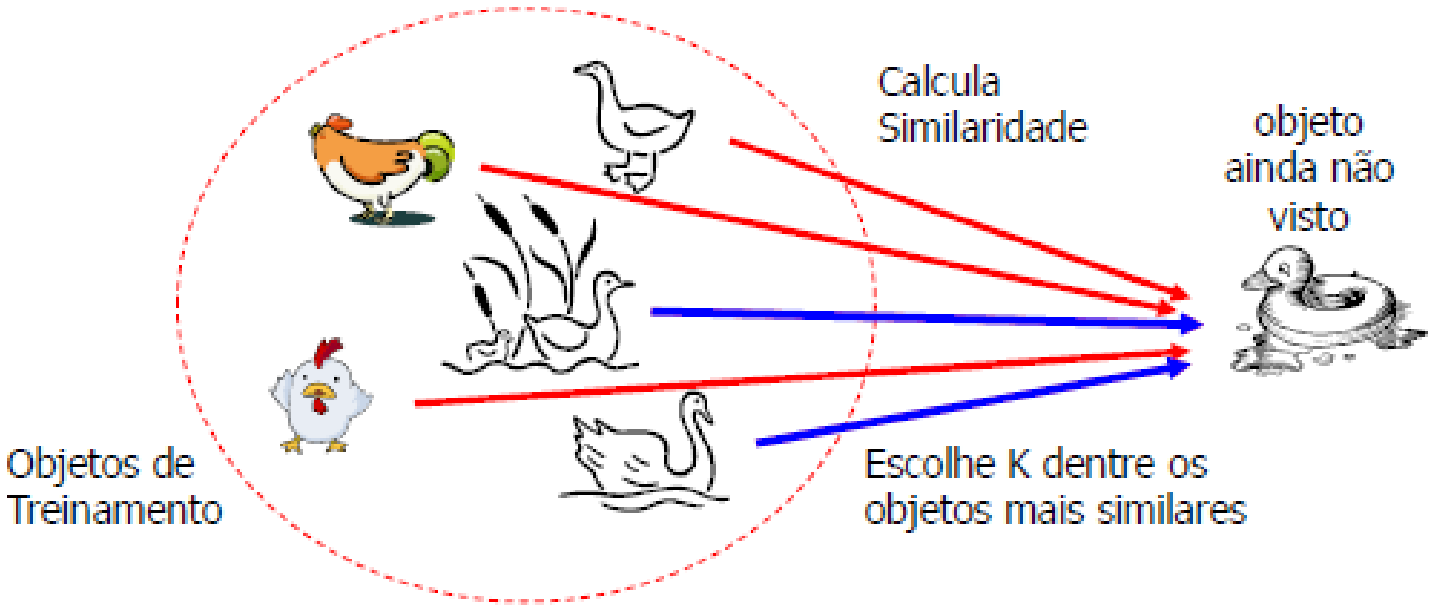


Classificadores Baseados em Instâncias

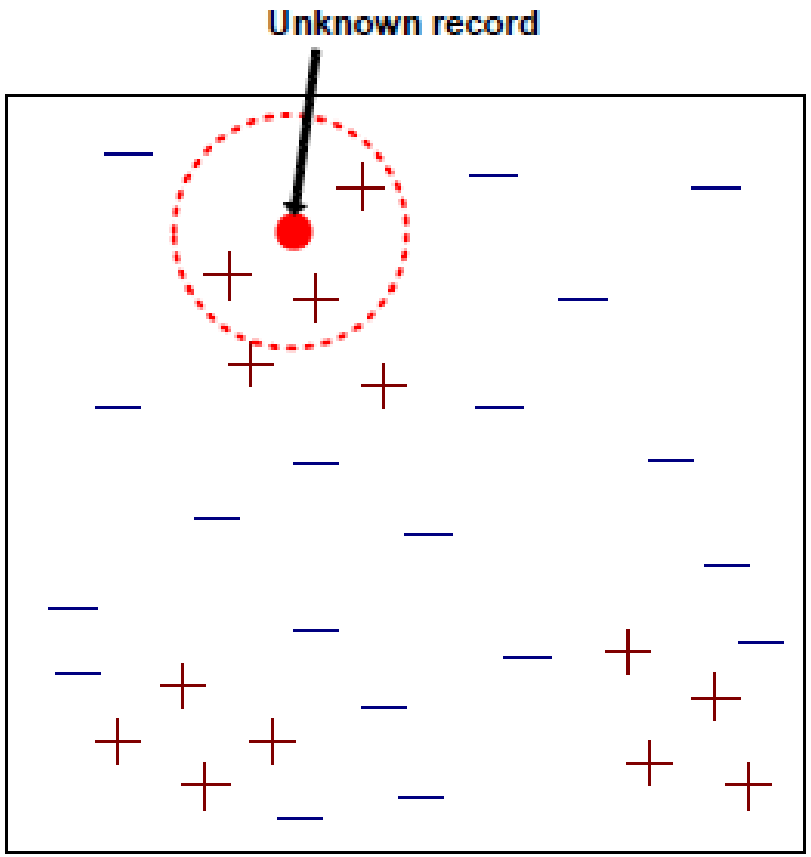
- Exemplos:
  - Rote-learner
    - memoriza o conjunto de treinamento completo e realiza a classificação apenas se os valores dos atributos da instância em questão se casam perfeitamente com um dos exemplos de treinamento
      - Extremamente restritivo...
  - K-NN
    - Usa as K instâncias mais similares (vizinhos mais próximos) para realizar a classificação por votação

K-NN

- Idéia Básica:
  - Se anda como um pato, “quacks” como um pato, então provavelmente é um pato



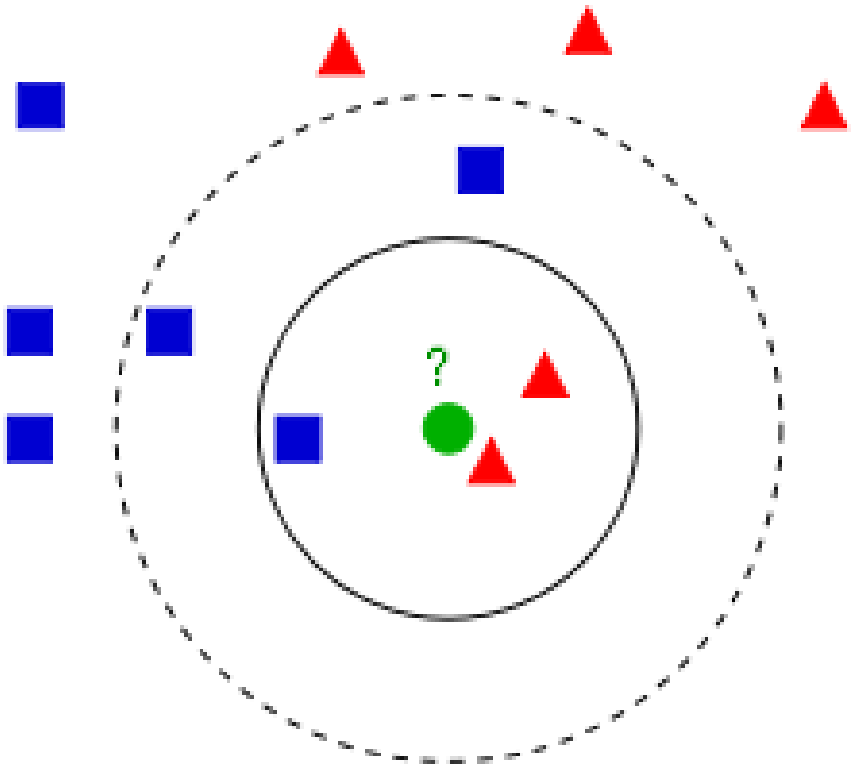
K-NN



- Requer 3 coisas
  - A base de dados de treinamento
  - Uma medida de (dis)similaridade entre os objetos da base
  - O valor de K: no. de vizinhos mais próximos a recuperar
- Para classificar um objeto não visto:
  - Calcule a (dis)similaridade para todos os objetos de treinamento
  - Obtenha os K objetos da base mais similares (mais próximos)
  - Classifique o objeto não visto na classe da maioria dos K vizinhos

Algoritmo K-NN

- Qual a classe do círculo verde?
  - k = 3: triângulo vermelho
  - k = 5: quadrado azul

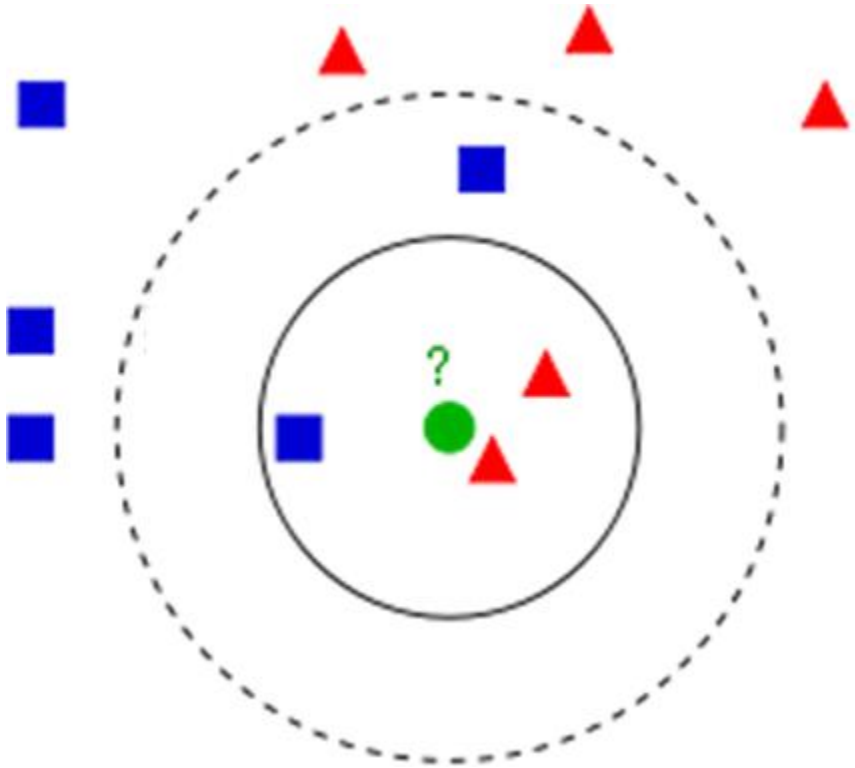


Algoritmo K-NN

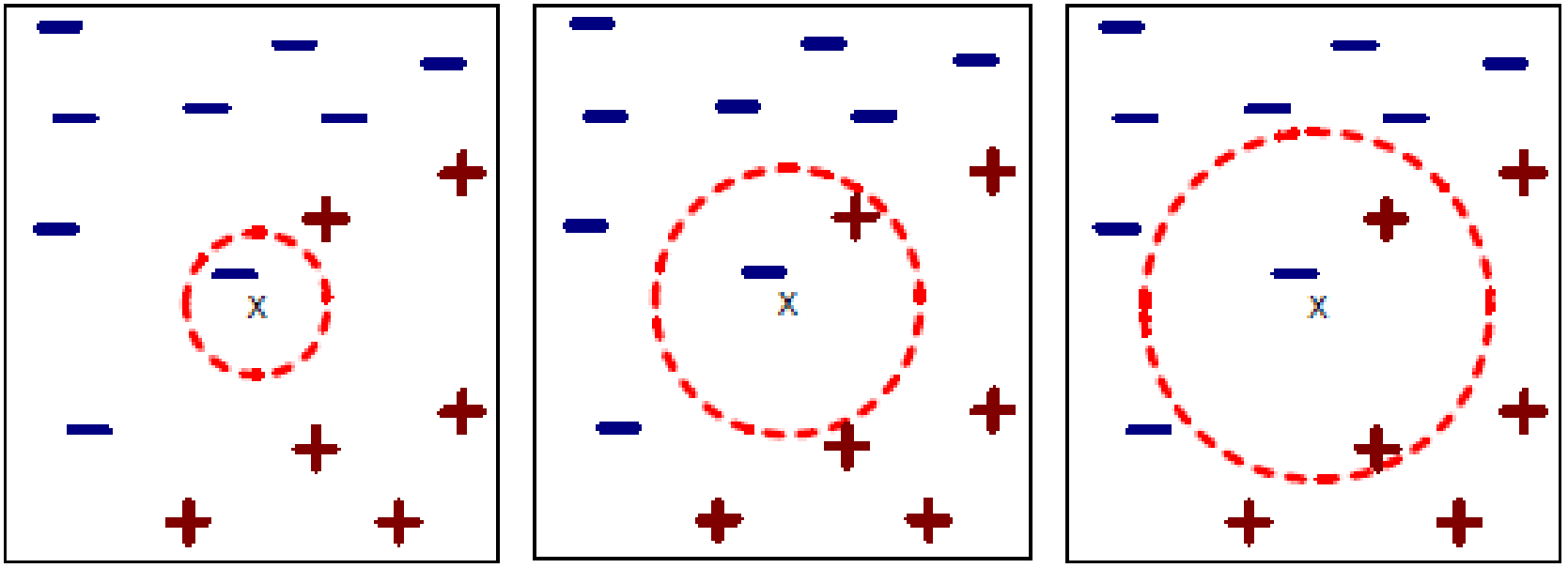
- O que fazer em caso de empate entre duas ou mais classes?
  - Considerar apenas os k-1 vizinhos mais próximos.
  - Em caso de novo empate, repetir esse processo
  - Esse processo para quando uma classe for unânime

Algoritmo K-NN

- Qual a classe do círculo verde?
  - k = 4: empate
  - k = 3: triângulo vermelho



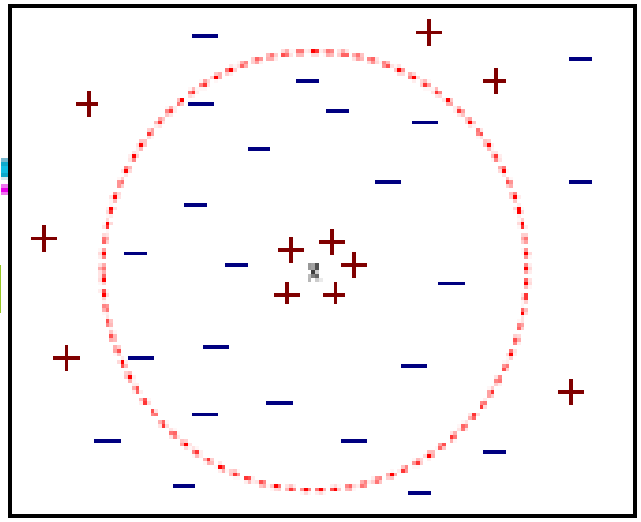
K-NN



(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

- K-NN: Visão geométrica para 2 atributos contínuos e dissimilaridade por distância Euclidiana. K = 1, 2 e 3

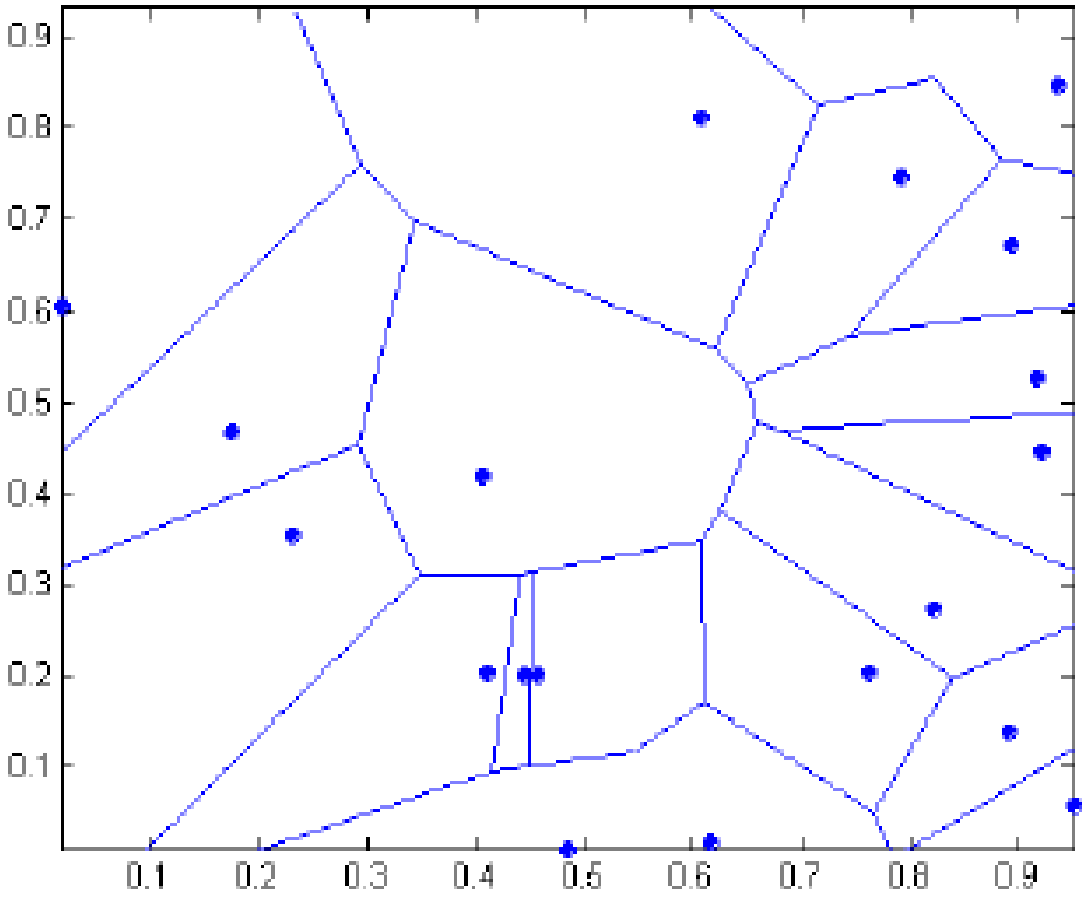
K-NN



- Escolha do Valor de K:
  - Muito pequeno:
    - função de discriminação entre classes muito flexível
    - porém, sensível a ruído
      - classificação pode ser instável
      - menor gasto computacional?
  - Muito grande:
    - mais robusto a ruído
    - aumenta a informação sobre a probabilidade de pertencer à classe
    - porém, vizinhança tende a incluir objetos de outras classes
      - privilegia classe majoritária
      - reduz flexibilidade da função de discriminação
      - maior gasto computacional?

1-NN

- Diagrama de Voronoi

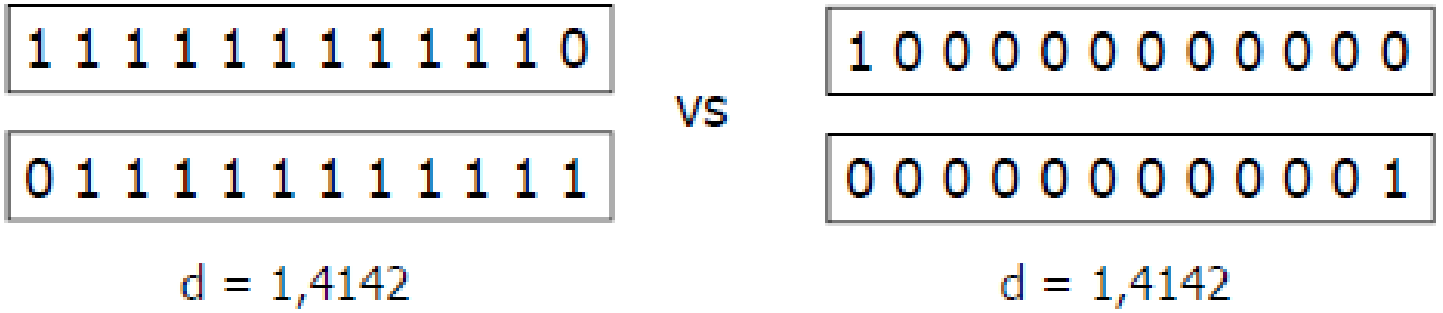


K-NN

- Como calcular as (dis)similaridades... ?
  - Já vimos anteriormente no curso que a medida mais apropriada depende:
  - do(s) tipo(s) do(s) atributos !
  - do domínio de aplicação !
- Por exemplo:
  - Euclidiana
  - Casamento Simples (Simple Matching)
  - Cosseno
  - Pearson...

K-NN

- Exemplo de Escolha Inapropriada:
  - Euclidiana para atributos binários assimétricos
  - Pode produzir resultados contra-intuitivos...



K-NN

- Além da escolha de uma medida apropriada, é preciso condicionar os dados de forma apropriada
  - Por exemplo, atributos podem precisar ser normalizados para evitar que alguns dominem completamente a medida de (dis)similaridade
  - Exemplo:
    - Altura de uma pessoa adulta normal: 1.4m a 2.2m
    - Peso de uma pessoa adulta sadia: 50Kg a 150Kg
    - Salário de uma pessoa adulta: $400 a $30.000

K-NN

- Na versão básica do algoritmo, a indicação da classe de cada vizinho possui o mesmo peso para o classificador
  - 1 voto por vizinho mais próximo
- Isso torna o algoritmo sensível à escolha de K
- Uma forma de reduzir esta sensibilidade e permitir assim o aumento de K (para aumentar a robustez a ruído) é ponderar cada voto pela respectiva distância
  - Heurística Usual: Peso referente ao voto de um vizinho decai de forma inversamente proporcional à distância entre esse vizinho e o objeto em questão

K-NN: Características

- Classificadores K-NN são do tipo **lazy**
  - Ao contrário de classificadores do tipo **eager**, não constroem um modelo explicitamente
    - Atrasam a discriminação até a chegada dos dados
  - Isso torna a classificação de novos objetos relativamente custosa computacionalmente
    - É necessário calcular as distâncias de cada um dos objetos a serem classificados a todos os objetos da base de treinamento!
    - Custo pode ser reduzido com uso de KD-Tree

K-NN: Características

- Sensíveis ao projeto:
  - Escolha de K...
  - Escolha da medida de (dis)similaridade...
- Podem ter poder de classificação elevado:
  - Função de discriminação muito flexível para K pequeno
- Podem ser sensíveis a ruído:
  - Pouco robustos para K pequeno

K-NN: Características

- Ao contrário do Naive Bayes, é sensível a atributos irrelevantes:
  - distorcem o cálculo das distâncias
  - maldição da dimensionalidade...
    - demanda seleção e/ou ponderação de atributos
- Por outro lado, permitem atribuir importâncias distintas para diferentes atributos
  - ponderação de atributos
  - geralmente leva a classificadores mais precisos

Algoritmo K-NN

- Vantagens
  - Simples de implementar
  - Não requer uma etapa de treinamento
  - Ideal para conjuntos de dados pequenos ou médios
  - Usa informação local, podendo ser implementado comportamentos adaptativos
  - Pode ser paralelizado

Algoritmo K-NN

- Desvantagens
  - Custo computacional e de armazenamento alto para conjuntos de dados grandes ou com muitos atributos
  - A constante k usada para definir o número de vizinhos é obtida por tentativa e erro
  - Sua precisão pode ser severamente degradada pela presença de ruído ou atributos irrelevantes

Algoritmo K-NN

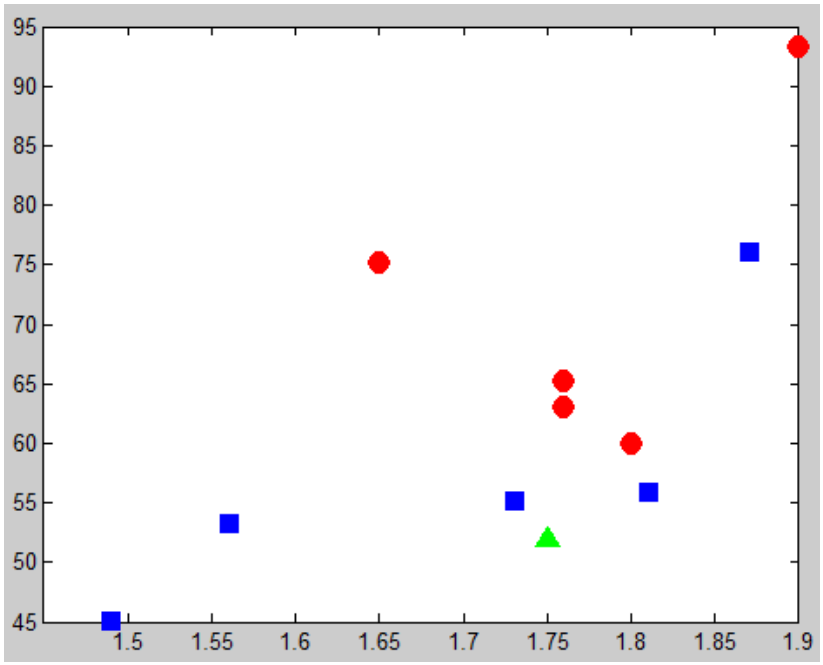
- Exemplo
  - Qual a classe da amostra abaixo, dado o conjunto de treinamento ao lado?

Altura	Peso	Sexo
1,75	52,0	?

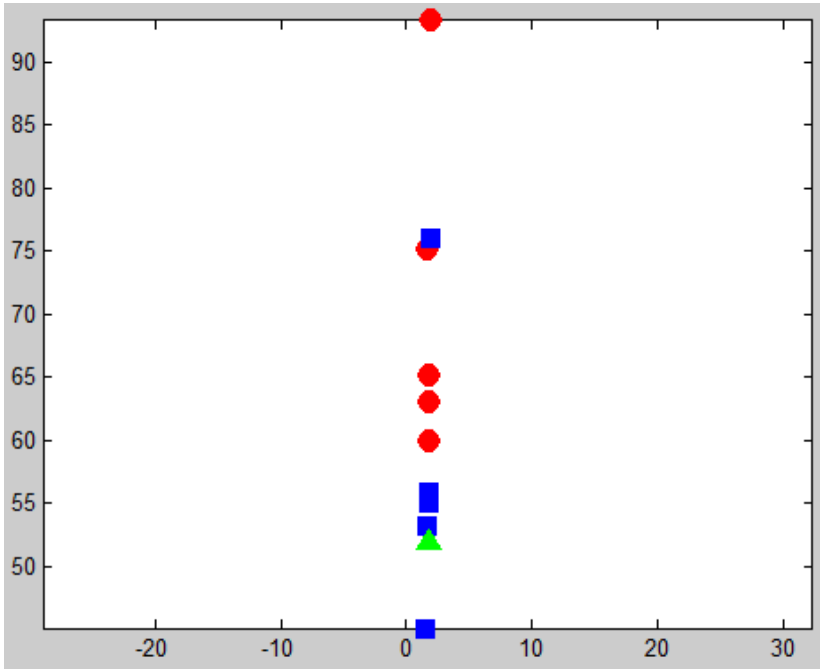
Altura	Peso	Sexo
1,87	76,1	0
1,65	75,2	1
1,80	60,0	1
1,81	55,9	0
1,90	93,3	1
1,74	65,2	1
1,49	45,1	0
1,56	53,2	0
1,73	55,1	0
1,76	63,1	1

Algoritmo K-NN

Com ajuste de escala



Sem ajuste de escala



Algoritmo K-NN

- Primeiro passo
  - Normalizar os valores
    - z-score

Média Altura	Média Peso
1,73	64,22

Desvio Altura	Desvio Peso
0,13	14,01

Altura	Peso	Sexo
1,87	76,1	0
1,65	75,2	1
1,80	60,0	1
1,81	55,9	0
1,90	93,3	1
1,74	65,2	1
1,49	45,1	0
1,56	53,2	0
1,73	55,1	0
1,76	63,1	1

Algoritmo K-NN

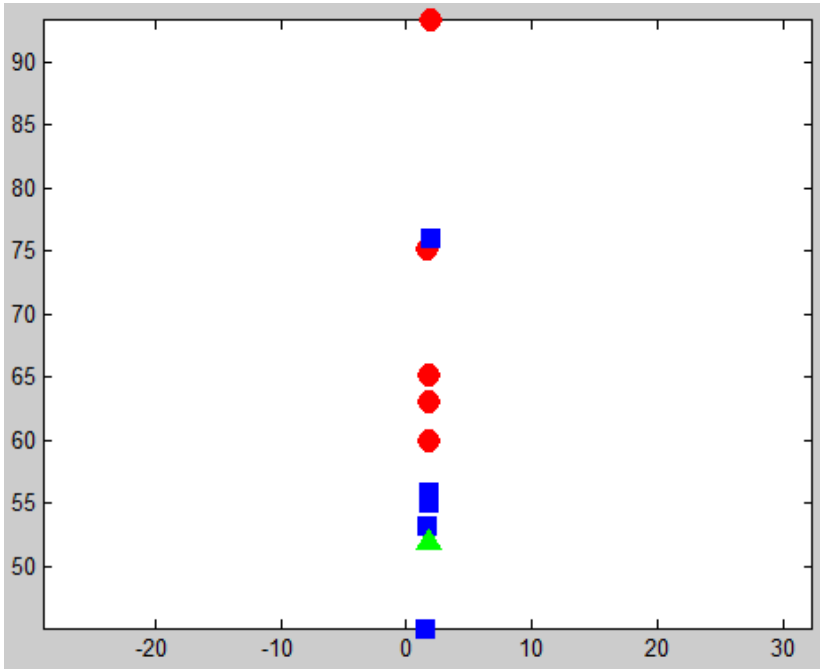
- Primeiro passo
  - Normalizar os valores
    - z-score

Altura	Peso	Sexo
0,12	-0,87	?

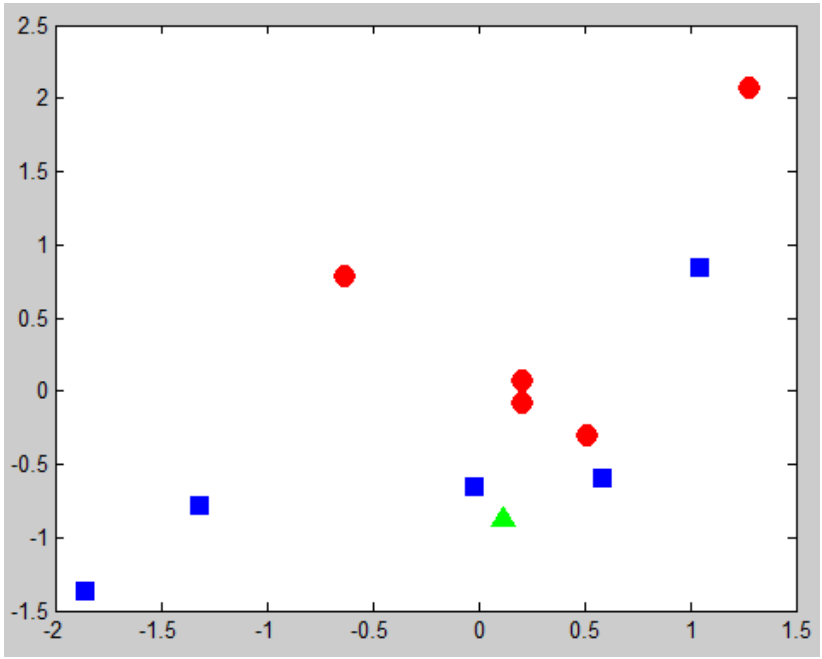
Altura	Peso	Sexo
1,04	0,84	0
-0,63	0,78	1
0,51	-0,30	1
0,58	-0,59	0
1,27	2,07	1
0,20	0,06	1
-1,85	-1,36	0
-1,32	-0,78	0
-0,02	-0,65	0
0,20	-0,07	1

Algoritmo K-NN

Sem ajuste de escala



zscore



Algoritmo K-NN

- Segundo passo
  - Calcular as distâncias da amostra desconhecida para as conhecidas

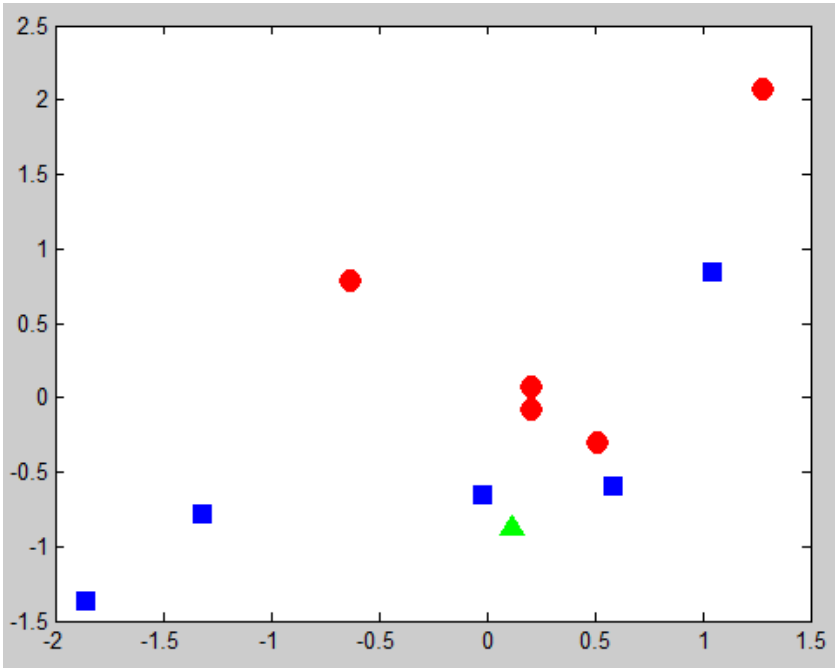
Altura	Peso	Sexo
0,12	-0,87	?

Altura	Peso	Sexo	D
1,04	0,84	0	2,64
-0,63	0,78	1	0,90
0,51	-0,30	1	0,96
0,58	-0,59	0	0,74
1,27	2,07	1	4,10
0,20	0,06	1	1,03
-1,85	-1,36	0	2,47
-1,32	-0,78	0	1,36
-0,02	-0,65	0	0,08
0,20	-0,07	1	0,88

Algoritmo K-NN

- Terceiro passo
  - Classificação: k = 3

Altura	Peso	Sexo
0,12	-0,87	? = 0

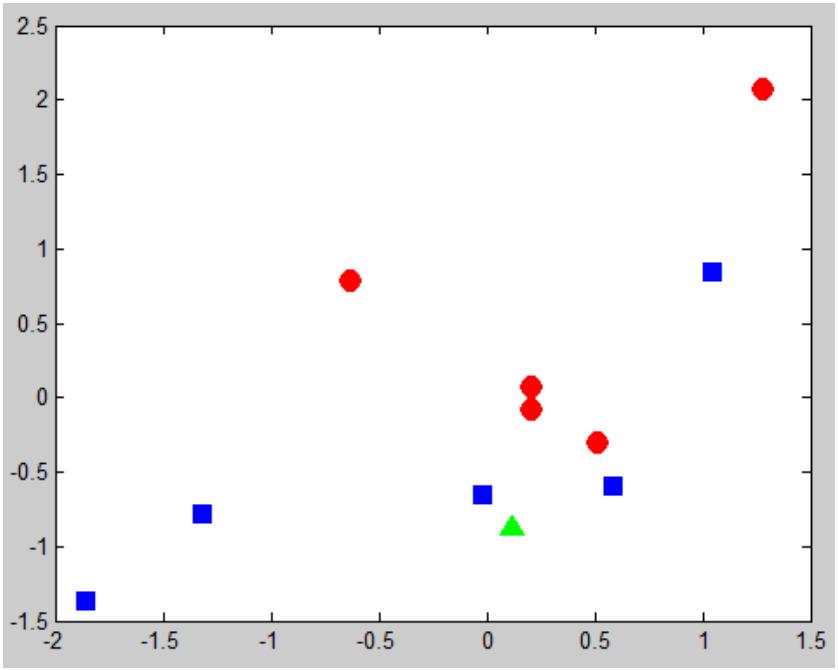


Altura	Peso	Sexo	D
1,04	0,84	0	2,64
-0,63	0,78	1	0,90
0,51	-0,30	1	0,96
0,58	**-0,59**	**0**	**0,74**
1,27	2,07	1	4,10
0,20	0,06	1	1,03
-1,85	-1,36	0	2,47
-1,32	-0,78	0	1,36
-0,02	**-0,65**	**0**	**0,08**
0,20	**-0,07**	**1**	**0,88**

Algoritmo K-NN

- Terceiro passo
  - Classificação: k = 5

Altura	Peso	Sexo
0,12	-0,87	? = 1



Altura	Peso	Sexo	D
1,04	0,84	0	2,64
-0,63	**0,78**	**1**	**0,90**
0,51	**-0,30**	**1**	**0,96**
0,58	**-0,59**	**0**	**0,74**
1,27	2,07	1	4,10
0,20	0,06	1	1,03
-1,85	-1,36	0	2,47
-1,32	-0,78	0	1,36
-0,02	**-0,65**	**0**	**0,08**
0,20	**-0,07**	**1**	**0,88**

Algoritmo 1-NN

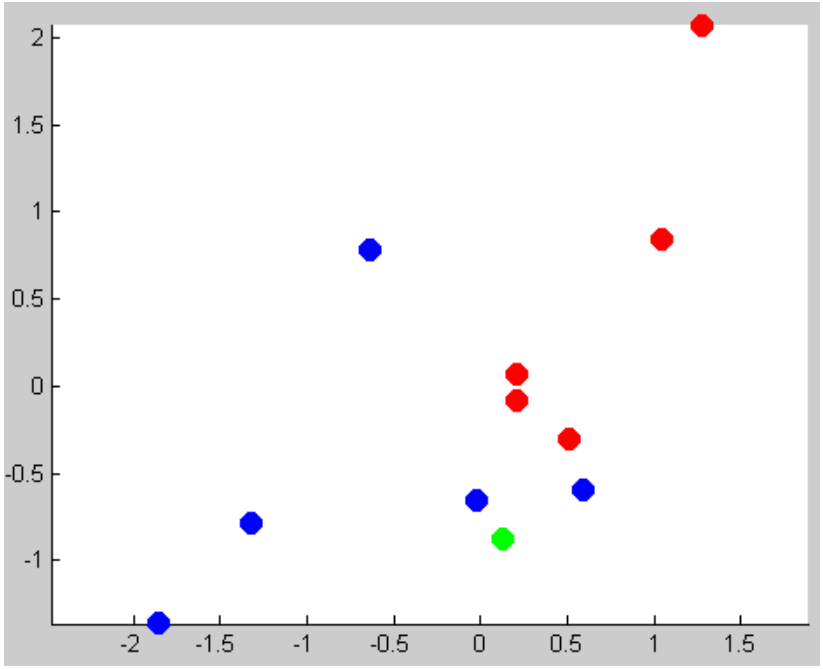
- Uma amostra desconhecida é considerada como pertencente a mesma classe da amostra conhecida que apresentar a menor distância até ela
  - É um caso particular do algoritmo K-NN
    - Nesse caso, k = 1
  - Qual a classe do círculo verde?

Nearest Prototype Classifier

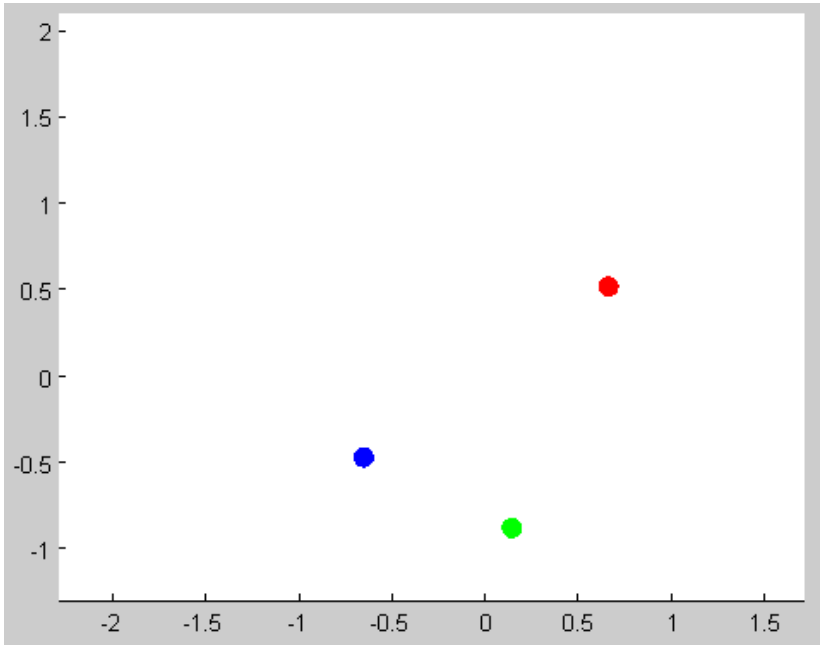
- Também conhecido como algoritmo do *centroide mais próximos*
  - Similar ao algoritmo 1-NN
  - Atribui uma observação à classe de amostras de treinamento cujo centroide (média) se encontra mais próximo

Nearest Prototype Classifier

zscore



Centroides



Limitações da distância Euclidiana

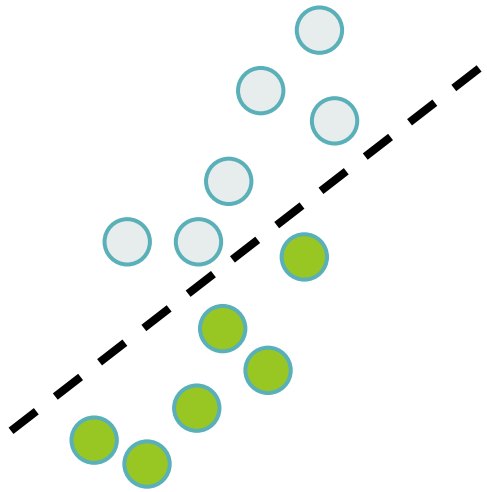
- Se um classificador baseado em distância mínima tem um elevado número de classificações corretas, não há razão para procurar classificadores mais complexos
  - No entanto, é frequente ocorrer um baixo desempenho

Limitações da distância Euclidiana

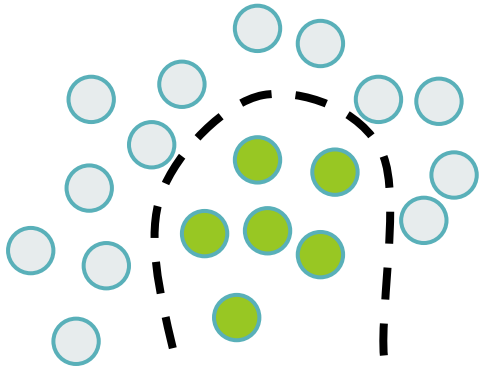
- Por que isso ocorre?
  - Atributos inadequados
  - Atributos correlacionados
  - As superfícies de decisão podem ser não-lineares
  - Existência de subclasses distintas
  - Espaço de padrões muito complexo

Limitações da distância Euclidiana

- Atributos correlacionados

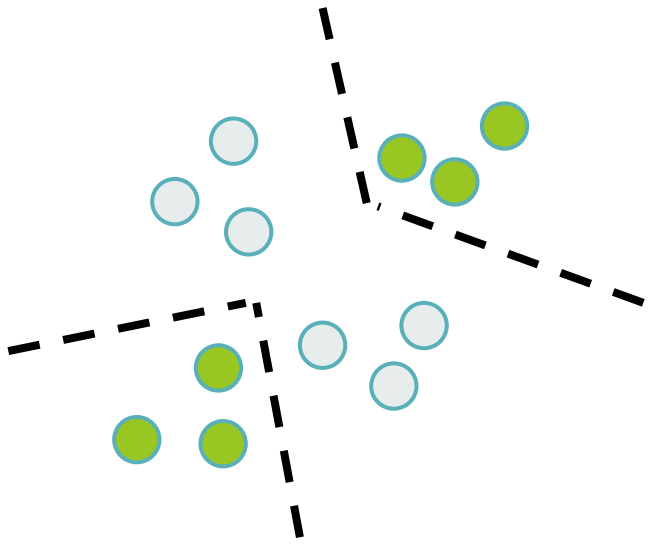


- As superfícies de decisão podem ser não-lineares

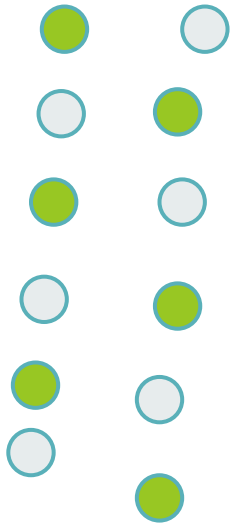


Limitações da distância Euclidiana

- Existência de subclasses distintas



- Espaço de padrões muito complexo



Maldição da dimensionalidade

- Suponha o seguinte problema
  - Um conjunto de dados é descrito por 20 atributos
  - Destes, apenas 2 são relevantes
    - Os demais são atributos ruins ou correlacionados
  - O resultado será um mau desempenho na classificação
- O algoritmo K-NN é normalmente enganado quando o número de atributos é grande

Maldição da dimensionalidade

- Maldição da dimensionalidade (ou *Curse of dimensionality*)
  - Termo que se refere a vários fenômenos que surgem na análise de dados em espaços com muitas dimensões (atributos)
    - Muitas vezes com centenas ou milhares de dimensões
  - Basicamente, adicionar características não significa sempre melhora no desempenho de um classificador

Maldição da dimensionalidade

- Teorema do patinho feio (de Watanabe)
  - Caso haja um conjunto suficientemente grande de características em comum, sem uma outra referência previamente estabelecida, é possível fazer com que dois padrões arbitrários sejam considerados similares.
  - Um cisne e um pato e um par de cisnes podem ficar igualmente similares

Maldição da dimensionalidade

- Um grande número de atributos tende a gerar informação redundante
  - Isso prejudica o desempenho do sistema

- Classificar dados significa encontrar grupos com propriedades similares
  - Em espaços com muitas dimensões as amostras se tornan esparsas e pouco similares
  - Isso impossibilita estratégias comuns de organização dos dados de serem eficiente.

Maldição da dimensionalidade

- É possível evitar isso?
  - O número de amostras de treinamento deve aumentar exponencialmente com o aumento do número de atributos
  - Isso nem sempre é possível na prática
    - De onde tirar novos dados?

Maldição da dimensionalidade

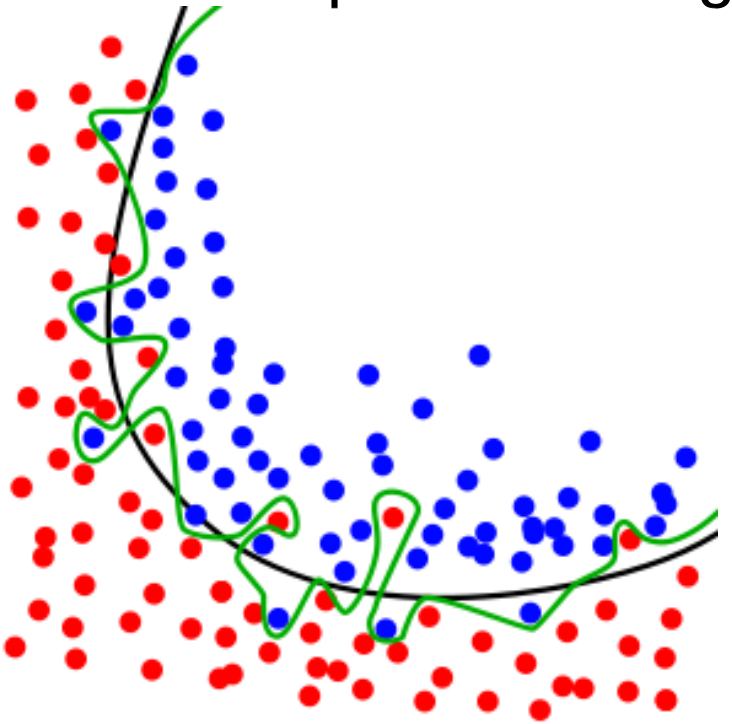
- É possível evitar isso?
  - Podemos reduzir o número de dimensões do espaço de características
    - Etapa importante no projeto de um sistema de classificação
    - Pode ser feita selecionando e/ou compondo as características mais adequadas
    - Uso de técnicas de seleção e combinação de características

Overfitting

- Fenômeno que ocorre quando o modelo estatístico se ajusta em demasiado ao conjunto de dados/amostra
  - Também conhecido com **sobrejuste** ou **over-training**
  - Ao invés de aprender o padrão, o modelo estatístico aprende suas “esquisitices”

Overfitting

- É comum haver desvios causados por erros de medição ou fatores aleatórios nas amostras
  - No overfitting, o modelo se ajusta a estes desvios e se esquece do comportamento geral dos dados

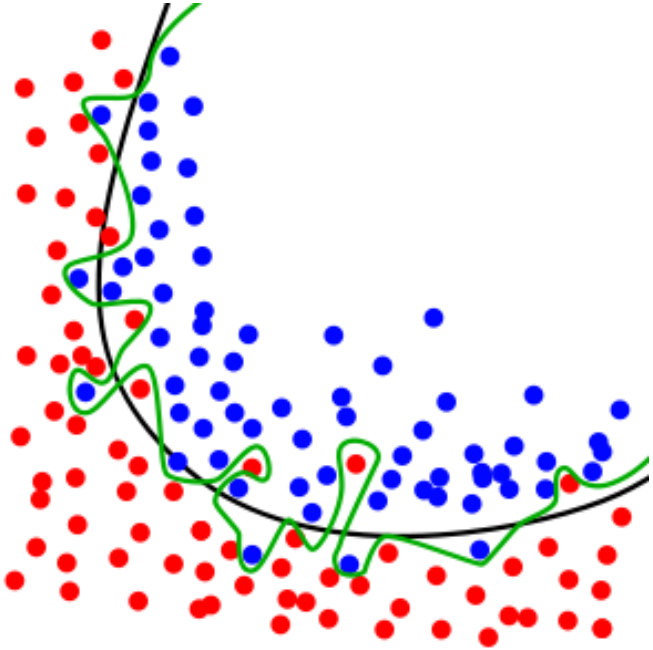


Overfitting

- Modelo com overfitting
  - Se degenera e se especializa no conjunto de treinamento
    - Alta precisão quando testado com seu conjunto de treinamento
  - Esse modelo não representa a realidade e deve ser evitado
    - Não tem capacidade de generalização.

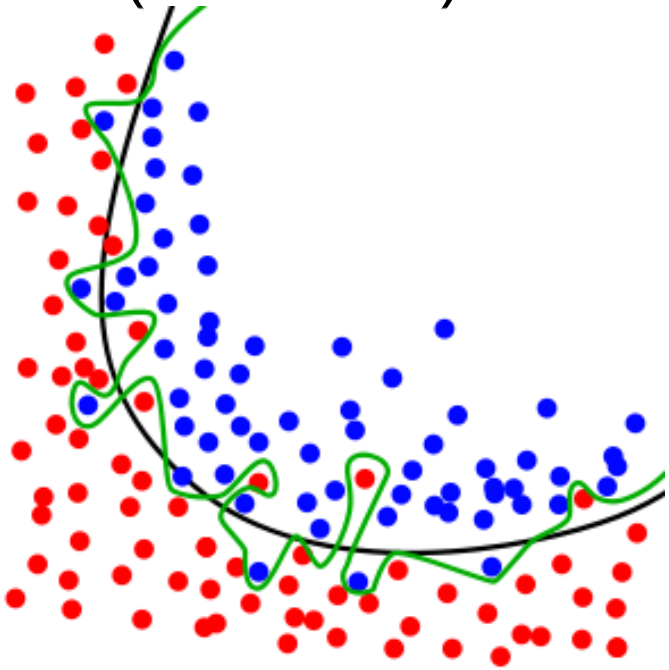
Overfitting

- Simplicidade é a resposta
  - Ajuda a proteger contra overfitting
    - É preferível um classificador que erre os dados estranhos no pressuposto de que eles são, de fato, estranhos e sem valor preditivo.



Overfitting

- Simplicidade é a resposta
  - Navalha de Occam
    - "*Se em tudo o mais forem idênticas as várias explicações de um fenômeno, a mais simples é a melhor*“ - William de Ockham (século XIV)



Overfitting

- Como contornar esse problema?
  - Regularização
    - Manter todos os atributos, mas reduzir a magnitude/valor dos deles
    - Penalizar atributos pela imposição de restrições de suavidade ao modelo de aproximação

Overfitting

- Como contornar esse problema?
  - Métodos de poda (*pruning*)
    - Voltados para árvores de decisão
    - Muitas das arestas ou sub-árvores podem refletir ruídos ou erros.
    - Necessidade de detectar e excluir essas arestas e sub-árvores
    - Simplifica a árvore e facilita sua interpretabilidade por parte do usuário

Overfitting

- Como contornar esse problema?
  - Cross-validation
    - Consiste em separar os dados em Treinamento e Teste
    - Essa divisão dos dados em subconjuntos ajuda a evitar que o modelo aprenda as particularidades dos dados

Teorema de Bayes

- Frequentemente, uma informação é apresentada na forma de probabilidade condicional
  - Probabilidade de um evento ocorrer dada uma condição
    - Probabilidade de um evento *B*, sabendo qual será o resultado de um evento *A*
- Esse tipo de problema é tratado usando o Teorema de Bayes

Naïve Bayes

- Naive Bayes é um dos mais simples e bem difundidos classificadores baseados no Teorema de Bayes
  - Paradigma probabilístico
- Para compreender esse classificador, devemos relembrar alguns conceitos elementares da teoria de probabilidade:
  - Probabilidade Conjunta
  - Probabilidade Condicional
  - Independência Condicional

Eventos independentes

- Dizemos que dois eventos são independentes quando a realização ou a não realização de um dos eventos não afeta a probabilidade da realização do outro e vice-versa.
  - Por exemplo, quando lançamos dois dados, o resultado obtido em um deles independe do resultado obtido no outro.
  - Nesse caso, a probabilidade de que ambos aconteçam ao mesmo tempo é igual ao produto de suas probabilidades
    - $P(A \text{ e } B) = P(A \cap B) = P(A) * P(B)$

Evento independente

- Exemplo de evento independente
  - A probabilidade de em uma família nascer um menino e ele ter olhos azuis
  - Nesse caso, a probabilidade do sexo da criança em nada interfere na probabilidade dela vir a ter olhos azuis

Eventos mutualmente exclusivos

- Dizemos que dois ou mais eventos são mutuamente exclusivos quando a realização de um exclui a realização dos outro.
  - Assim, no lançamento de uma moeda, o evento “tirar cara” e o evento “tirar coroa” são mutuamente exclusivos, já que, ao se realizar um deles, o outro não se realiza.
  - Se dois eventos são mutuamente exclusivos, a probabilidade de que um ou outro se realize é igual à soma das probabilidades de que cada um deles se realize:

Eventos mutualmente exclusivos

- Dizemos que dois ou mais eventos são mutuamente exclusivos quando a realização de um exclui a realização dos outro.
  - Assim, no lançamento de uma moeda, o evento “tirar cara” e o evento “tirar coroa” são mutuamente exclusivos, já que, ao se realizar um deles, o outro não se realiza.
  - Se dois eventos são mutuamente exclusivos, a probabilidade de que um ou outro se realize é igual à soma das probabilidades de que cada um deles se realize:
  - Exemplo: estimar a probabilidade de nascer uma criança com olhos castanhos ou azuis
    - P(A) = P(menino de olhos castanhos) = 3/8
    - P(B) = P(meninas de olhos azuis) = 1/8
    - P(A ou B) = P(A) + P(B)= 3/8 + 1/8 = 1/4

Eventos mutualmente exclusivos

- Lei da soma para eventos mutuamente exclusivos
  - Neste caso podemos definir a seguinte expressão de probabilidade
    - P(A ou B) = P(A) + P(B) – P(A e B)

- Exemplo: Estimar a probabilidade de nascer um menino ou uma criança de olhos azuis. Assim, tem-se:
  - P(A) = P(menino) = 1/2
  - P(B) = P(olhos azuis) = 1/4
  - P(A e B) = P(meninos de olhos azuis) = 1/8
  - P(A ou B) = P(A) + P(B) – P(A e B) = 1/2 + 1/4 – 1/8

- A necessidade de subtrair a probabilidade de meninos de olhos azuis na P(A ou B) pode ser constatada pois tanto a valor P(menino) quanto P(olhos azuis) inclui a possibilidade de sair menino de olhos azuis, consequentemente esta probabilidade estaria sendo somada duas vezes caso não houvesse aquela subtração.

Eventos dependentes

- Dados dois eventos A e B, temos que a ocorrência do evento A exerce influência na probabilidade de ocorrência do outro evento, B

Combinação de Eventos

- Exemplo de evento dependente
  - A probabilidade de em uma família nascer um menino e ele ser daltônico
    - O gene do daltonismo na espécie humana está ligado ao sexo. Ele é provocado por genes recessivos localizados no cromossomo X (sem alelos no Y). Assim, o problema ocorre muito mais frequentemente nos homens que nas mulheres

Combinação de Eventos

- Eventos dependentes
  - Nesse caso, a probabilidade de ambos ocorrerem **ao mesmo tempo** assume um valor diferente dependendo da natureza da relação
  - Dados dois eventos A e B, a **probabilidade condicional de A dado B** é definida como o quociente entre a probabilidade conjunta de A e B, e a probabilidade de B:
    - $P(A|B) = \frac{P(A \cap B)}{P(B)}$
    - $P(B) > 0$

Combinação de Eventos

- Exemplo: cálculo da probabilidade condicional de um evento dependente
  - 250 alunos estão matriculados numa universidade
    - 100 homens e 150 mulheres
    - 110 no BCC e 140 no BSI

Sexo\Curso	BCC	BSI	Total
H	40	60	100
M	70	80	150
Total	110	140	250

Combinação de Eventos

- Exemplo: cálculo da probabilidade condicional de um evento dependente
  - Num sorteio, qual a probabilidade de sair alguém do BSI dado que o sorteada uma mulher?
- $P(BSI|M) = \frac{P(BSI \cap M)}{P(M)} = \frac{\frac{80}{250}}{\frac{150}{250}} = \frac{80}{150}\ 0{,}53 = 53\%$

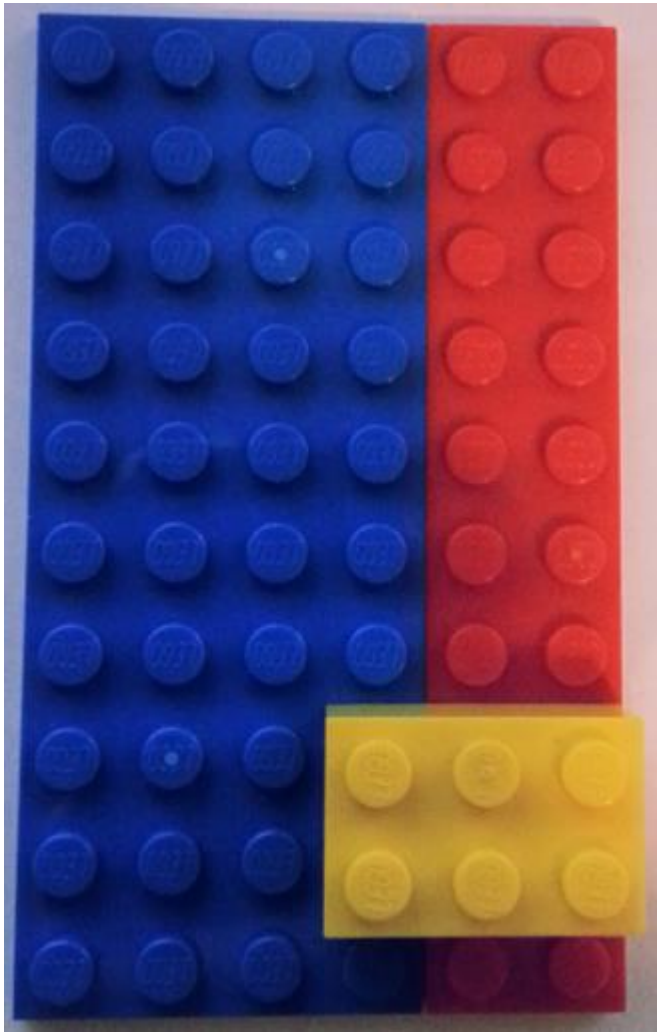
Complemento de probabilidade

- Por fim, temos também o complemento de uma probabilidade
  - $P(A^C) = 1 - P(A)$
- A probabilidade complementar de um evento A é a probabilidade de A não ocorrer
  - Ao lançarmos um dado, a probabilidade de sair um 6 será: P(6) = 1/6
  - A probabilidade de sair qualquer outro número será: $P(6^C) = 1 – 1/6 = 5/6$

Teorema de Bayes

- Exemplo
  - Considere o conjunto de peças de Lego ao lado
  - Perceba que o Lego Amarelo sempre esconde uma das cores: vermelho ou azul
  - Qual a probabilidade de sair a cor vermelha dado selecionamos um ponto amarelo?

$$P(vermelho|amarelo) = ?$$



Teorema de Bayes

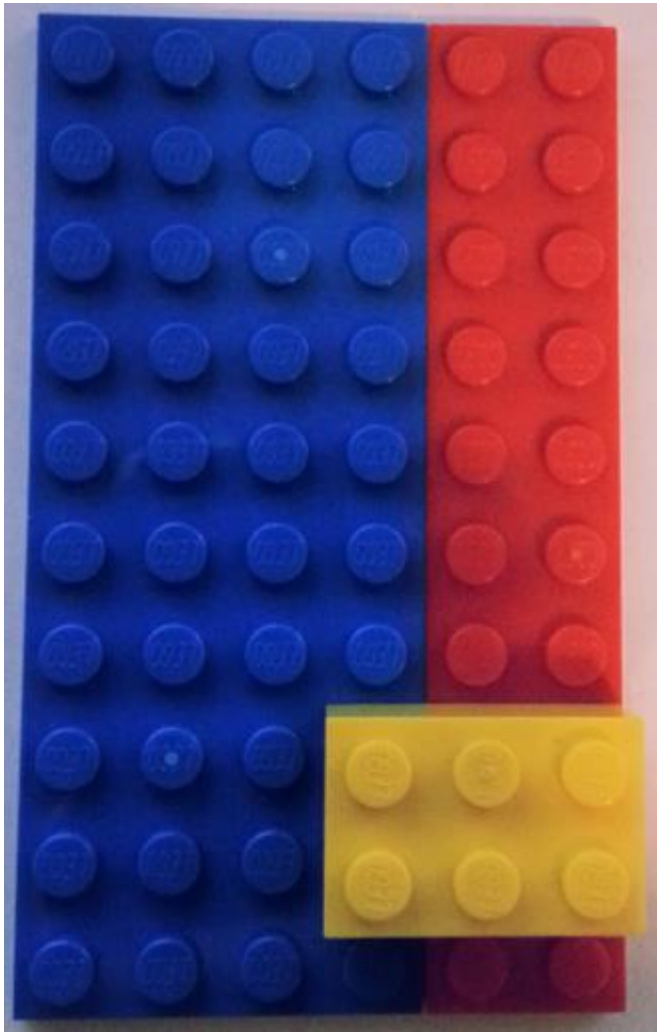
- Temos 60 pontos: calcular probabilidades
  - Probabilidade vermelho

$$P(vermelho) = \frac{20}{60} = \frac{1}{3}$$

  - Probabilidade azul

$$P(azul) = \frac{40}{60} = \frac{2}{3}$$

  - Soma das probabilidades dá 1

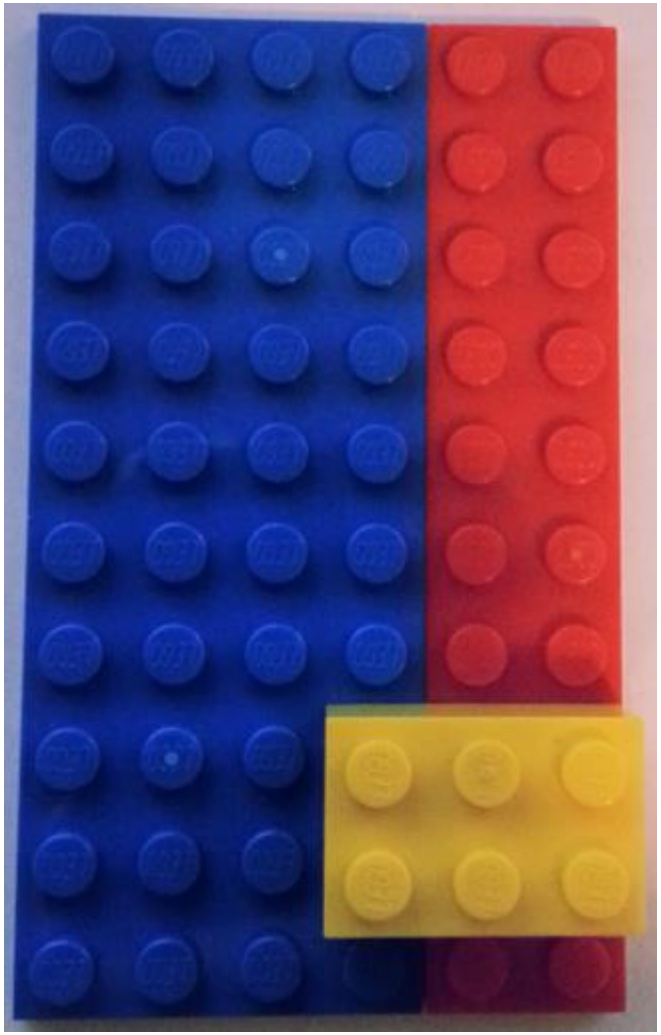


Teorema de Bayes

- Faltou calcular a probabilidade do amarelo
  - Probabilidade amarelo

$$P(amarelo) = \frac{6}{60} = \frac{1}{10}$$

  - Se somarmos as 3 probabilidades, o resultado é maior do que 1!
    - A peça amarela sempre vem com um outra cor
    - Probabilidade condicional!



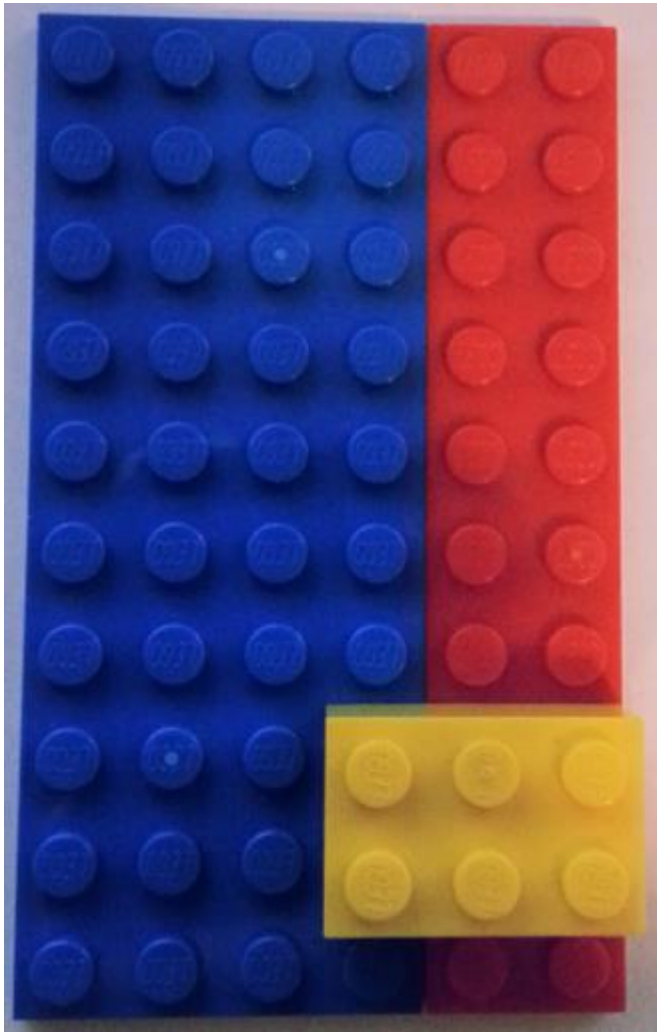
Teorema de Bayes

- Probabilidades do amarelo
  - Em relação ao vermelho

$$P(amarelo|vermelho) = \frac{4}{20} = \frac{1}{5}$$

  - Em relação ao azul

$$P(amarelo|azul) = \frac{2}{40} = \frac{1}{20}$$



Teorema de Bayes

- Voltando ao problema
  - Qual a probabilidade de sair a cor vermelha dado selecionamos um ponto amarelo?

$$P(vermelho|amarelo)$$

  - Isso equivale a calcular

$$\frac{P(vermelho)P(amarelo|vermelho)}{P(amarelo)}$$

Teorema de Bayes

- Voltando ao problema
  - Qual a probabilidade de sair a cor vermelha dado selecionamos um ponto amarelo?

$$\frac{P(vermelho)P(amarelo|vermelho)}{P(amarelo)}$$

  - Substituindo as probabilidades

$$P(vermelho|amarelo) = \frac{{}^{1}/_{3} * {}^{1}/_{5}}{{}^{1}/_{10}} = \frac{2}{3}$$

Probabilidade Conjunta

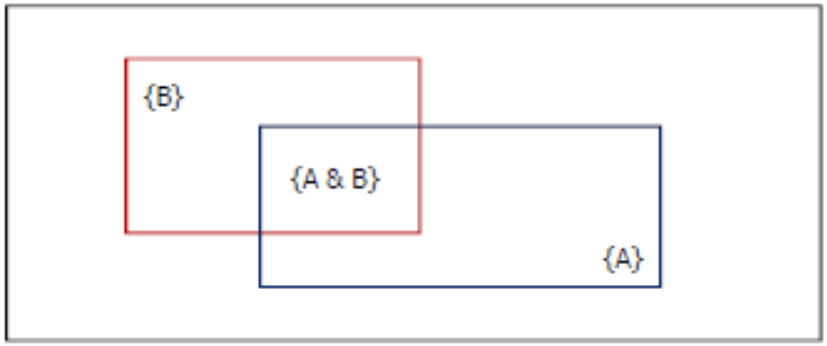
- É simplesmente a probabilidade de 2 eventos A e B (2 valores associados a 2 variáveis aleatórias) ocorrerem
  - P(A): Probabilidade do evento A ocorrer
  - P(B): Probabilidade do evento B ocorrer
  - P(A & B) ou P(A,B): Probabilidade de A e B ocorrerem
  - P(A & B) = P(A)*P(B) se A e B forem **eventos independentes**
    - A ocorrência de um não afeta a probabilidade de ocorrência do outro
    - Por exemplo, dois dados jogados independentemente

Probabilidade Condicional

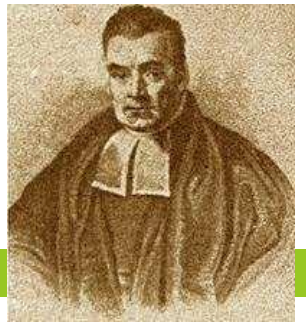
- Se A e B não forem eventos independentes, tem-se:
  - P(A & B) = P(A)*P(B|A)

    onde P(B|A) = P(A & B) / P(A) é a probabilidade que B ocorra dado que A ocorreu (**probabilidade condicional** de B dado A)

- Exemplo: várias bolas de 2 cores diversas em uma caixa
  - A = bola com 1 cor azul
  - B = bola com 1 cor vermelha
  - A & B = bola azul e vermelha



Teorema de Bayes

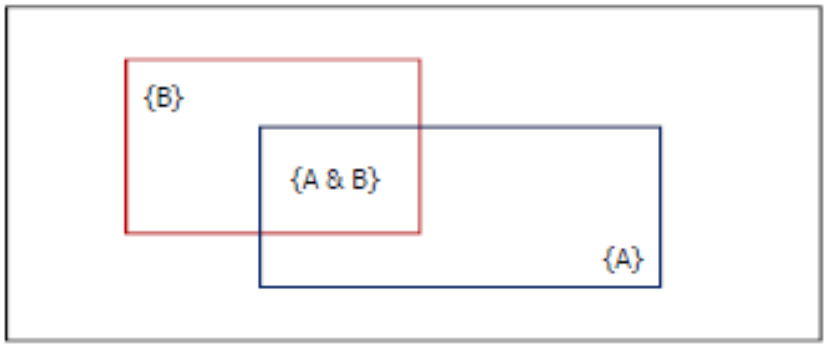


Thomas Bayes
(1702-1761)

- Note que P(A & B) = P(B & A) e portanto
  - P(B|A)*P(A) = P(A|B)*P(B)

- Teorema de Bayes:
  - P(B|A) = P(A|B)*P(B) / P(A)

  - {A} = conj. bolas azuis
  - {B} = conj. bolas vermelhas
  - {A & B} = conj. bolas azuis e vermelhas



Teorema de Bayes

- O Teorema de Bayes relaciona as probabilidades de *A* e *B* com suas respectivas probabilidades condicionadas

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, para\ P(B) > 0$$

  - Onde
    - *P(A)* e *P(B)*: probabilidades a **priori** de *A* e *B*
    - *P(B|A)* e *P(A|B)*: probabilidades a **posteriori** de *B* condicional a *A* e de *A* condicional a *B* respectivamente.

Teorema de Bayes

- O Teorema de Bayes nos permite calcular a probabilidade a posteriori para um determinado padrão pertencente a uma determinada classe
  - Em resumo

$$Prob\ Posteriori = \frac{Prob\ Priori\ * Distrib\ Prob}{Evidencia}$$

Example of Bayes Theorem

- Given:
  - A doctor knows that meningitis causes stiff neck 50% of the times
  - **Prior probability** of any patient having meningitis is 1/50000
  - **Prior probability** of any patient having stiff neck is 1/20

- If a patient has stiff neck (evidence), what's the posterior probability he/she has meningitis?

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

Teorema de Bayes

- Para várias variáveis aleatórias A_1, A_2, ..., An, e B: raciocínio análogo
- **Teorema de Bayes**:
  - $P(B|A_1,A_2,...,A_n) = P(A_1, A_2,...,A_n|B)*P(B) / P(A_1,A_2,...,A_n)$

- Se as vars aleatórias A_1 ... A_n forem independentes entre si tem-se:
  - $P(A_1,A_2,...,A_n) = P(A_1)*P(A_2)* ... *P(A_n)$ [**independência**]
  - $P(A_1, A_2,...,A_n|B) = P(A_1|B)*P(A_2|B)* ... * P(A_n|B)$ [**independência condicional**]
  - e finalmente...

$$P(B \mid A_1,\ldots,A_n) = \frac{P(B)\cdot\prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

Naïve Bayes

$$P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

- Naive Bayes é um algoritmo que utiliza o Teorema de Bayes com a hipótese de independência entre atributos

- Porque assumir independência entre atributos A_1 ... A_n?
  - Estimar probabilidades conjuntas $P(A_1, A_2,...,A_n)$ e $P(A_1, A_2,...,A_n|B)$ demandaria uma quantidade mínima de exemplos de cada combinação possível de valores de $A_1, A_2,...,A_n$
    - impraticável, especialmente para quantidades elevadas de atributos !
  - Apesar da hipótese ser quase sempre violada, o método (Naive Bayes) se mostra bastante competitivo na prática !

Exemplo:

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	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								
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								

- Idéia é estimar a probabilidade de cada valor B do atributo meta (valorda classe) dados os valores $A_1, A_2, \ldots, A_n$ dos demais atributos

$$P(B \mid A_1, \ldots, A_n) = \frac{P(B) \cdot \prod_{i=1}^{n} P(A_i \mid B)}{\prod_{i=1}^{n} P(A_i)}$$

Outlook	Temp	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No

Continuando…

Outlook (A_1)			Temperature (A_2)			Humidity (A_3)			Windy (A_4)			Play (B)	
	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								
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								

- Para um novo dia:

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	???

P(Yes|Sunny, Cool, High, True) = (2/9 × 3/9 × 3/9 × 3/9 × 9/14) / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = (3/5 × 1/5 × 4/5 × 3/5 × 5/14) / P(Sunny, Cool, High, True)

P(Yes|Sunny, Cool, High, True) = **0.0053** / P(Sunny, Cool, High, True)

P(No|Sunny, Cool, High, True) = **0.0206** / P(Sunny, Cool, High, True)

Play = No

Problema da Frequência Zero

- O que acontece se um determinado valor de atributo não aparece na base de treinamento, mas aparece no exemplo de teste?
  - Por exemplo: “Outlook = Overcast” para classe “No”
  - Probabilidade correspondente será zero
    - P(Overcast | “No”) = 0
  - Probabilidade a posteriori será também zero!
    - P(“No” | Overcast, ...) = 0
    - Não importa as probabilidades referentes aos demais atributos !
    - Muito radical, especialmente considerando que a base de treinamento pode não ser totalmente representativa
      - Por exemplo, classes minoritárias com instâncias raras

Problema da Frequência Zero

- Possível solução (Estimador de Laplace):
  - Adicionar 1 unidade fictícia para cada combinação de valor-classe
    - Como resultado, probabilidades nunca serão zero!
  - Exemplo (atributo Outlook – classe No):

$$\frac{3+1}{5+3} \quad \frac{0+1}{5+3} \quad \frac{2+1}{5+3}$$

Sunny ***Overcast*** ***Rainy***

  - Nota: Deve ser feito para todas as classes, para não inserir viés nas probabilidades de apenas uma classe

Problema da Frequência Zero

- Solução mais geral (Estimativa m):
  - Adicionar múltiplas unidades fictícias para cada combinação de valor-classe
  - Exemplo (atributo Outlook – classe No):

$$\frac{3+\frac{m}{3}}{5+m} \qquad \frac{0+\frac{m}{3}}{5+m} \qquad \frac{2+\frac{m}{3}}{5+m}$$

Sunny ***Overcast*** ***Rainy***

- Solução ainda mais geral:
  - Substituir o termo 1/n no numerador (onde n é o no. de valores do atributo) por uma probabilidade p qualquer

Valores Ausentes

- Treinamento:
  - excluir exemplo do conjunto de treinamento
- Classificação:
  - considerar apenas os demais atributos
- Exemplo:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	???

Verossimilhança para "Yes" = $3/9 \times 3/9 \times 3/9 \times 9/14 = 0.0238$

Verossimilhança para "No" = $1/5 \times 4/5 \times 3/5 \times 5/14 = 0.0343$

Probabilidade Estimada ("Yes") = 0.0238 / (0.0238 + 0.0343) = 41%

Probabilidade Estimada ("No") = 0.0343 / (0.0238 + 0.0343) = 59%

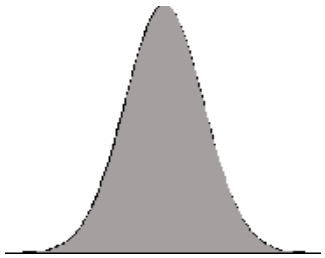
Atributos Numéricos

- **Alternativa 1**: Discretização
- **Alternativa 2**: Assumir ou estimar alguma função de densidade de probabilidade para estimar as probabilidades
  - Usualmente distribuição Gaussiana (Normal)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \mu)^2$$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Karl Gauss
1777-1855

Estatísticas para “weather”

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, …	85, …	80, …	95, …					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

- Valor de **densidade**:

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}\,6.2} e^{-\frac{(66-73)^2}{2\times 6.2^2}} = 0.0340$$

  - Porque o Teorema de Bayes convenientemente permite usar o valor de **densidade** de probabilidade para estimar a **probabilidade** de um valor pontual (teoricamente nula)... ?

Exercício

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

$$P(A_i \mid C_j) = \frac{1}{\sqrt{2\pi\sigma_{ij}^2}} e^{-\frac{(A_i - \mu_{ij})^2}{2\sigma_{ij}^2}}$$

- P(Taxable Income = 120 | No) ????
- P(Taxable Income = 120 | Yes) ????

Exercício

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	???

Naive Bayes: Características

- Robusto a ruídos isolados
  - Por exemplo, outliers ilegítimos
  - Afetam pouco o cálculo das probabilidades
- Robusto a atributos irrelevantes
  - Afetam pouco as probabilidades relativas entre classes
- Capaz de classificar instâncias com valores ausentes
- Assume que atributos são igualmente importantes
- Desempenho pode ser (mas muitas vezes não é) afetado pela presença de atributos correlacionados

Naive Bayes: Características

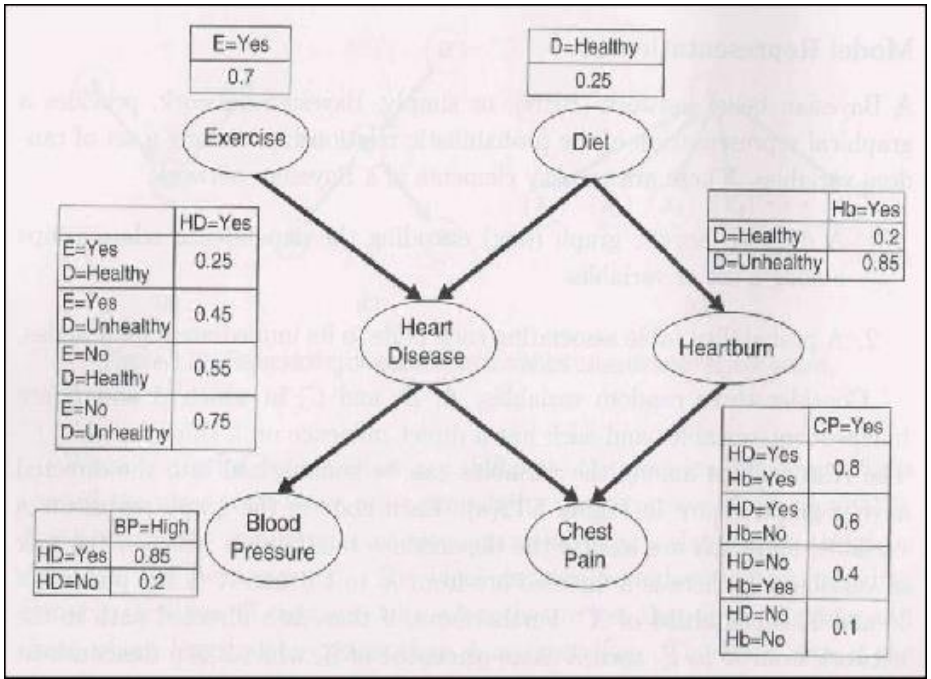
- Contrariamente ao 1R, usa todos os atributos
  - Mas ainda assim possui complexidade computacional linear em todas as variáveis do problema!
  - Tentou o 1R e não ficou satisfeito(a) ?
    - Experimente o Naive Bayes !
  - Ainda não ficou satisfeito(a) ?
    - Razão pode ser uma presença significativa de atributos correlacionados
    - Se forem simplesmente redundantes, seleção de atributos pode resolver
    - Caso contrário, faz-se necessária uma outra abordagem
      - Por exemplo, Redes Bayesianas

Seleção de Atributos: Wrapper Naive Bayes

- Algoritmo Guloso:
  - Selecione o melhor classificador Naive Bayes com um único atributo (avaliando todos em um conjunto de dados de teste)
  - Enquanto houver melhora no desempenho do classificador faça
    - Selecione o melhor classificador Naive Bayes com os atributos já selecionados anteriormente adicionados a um dentre os atributos ainda não selecionados
    - Nota: Apesar de ser um wrapper, o algoritmo acima é relativamente rápido devido à sua simplicidade e à eficiência computacional do Naive Bayes !

Redes Bayesianas

- Modelo probabilístico de Grafo
  - Nós: variáveis não determinísticas (probabilísticas)
  - Arcos: Interdependências entre variáveis
- Exemplo:



Redes Bayesianas

- Teoricamente, são capazes de solucionar o Calcanhar de Aquiles do Naive Bayes
- No entanto, obtenção dos modelos não é trivial
  - Determinação da topologia do grafo não é simples de ser sistematizada, especialmente sem conhecimento de domínio do problema
  - Determinação frequentista das probabilidades (como no Naive Bayes) só pode ser feita se todas as variáveis forem observáveis
    - Caso contrário, métodos de otimização são necessários para estimar essas probabilidades a partir das variáveis observáveis

Redes Bayesianas

- Além da construção do modelo, a inferência também não é trivial...
  - Teoricamente, as probabilidades de qualquer subconjunto das variáveis da rede podem ser obtidas (inferidas) a partir dos valores observados e/ou probabilidades das demais variáveis
  - No entanto, a inferência exata é um problema NP-hard
    - Até mesmo a obtenção de boas aproximações pode ser um problema complexo
  - Apesar disso, seu poder e flexibilidade têm motivado o interesse e pesquisa crescentes nesses modelos