

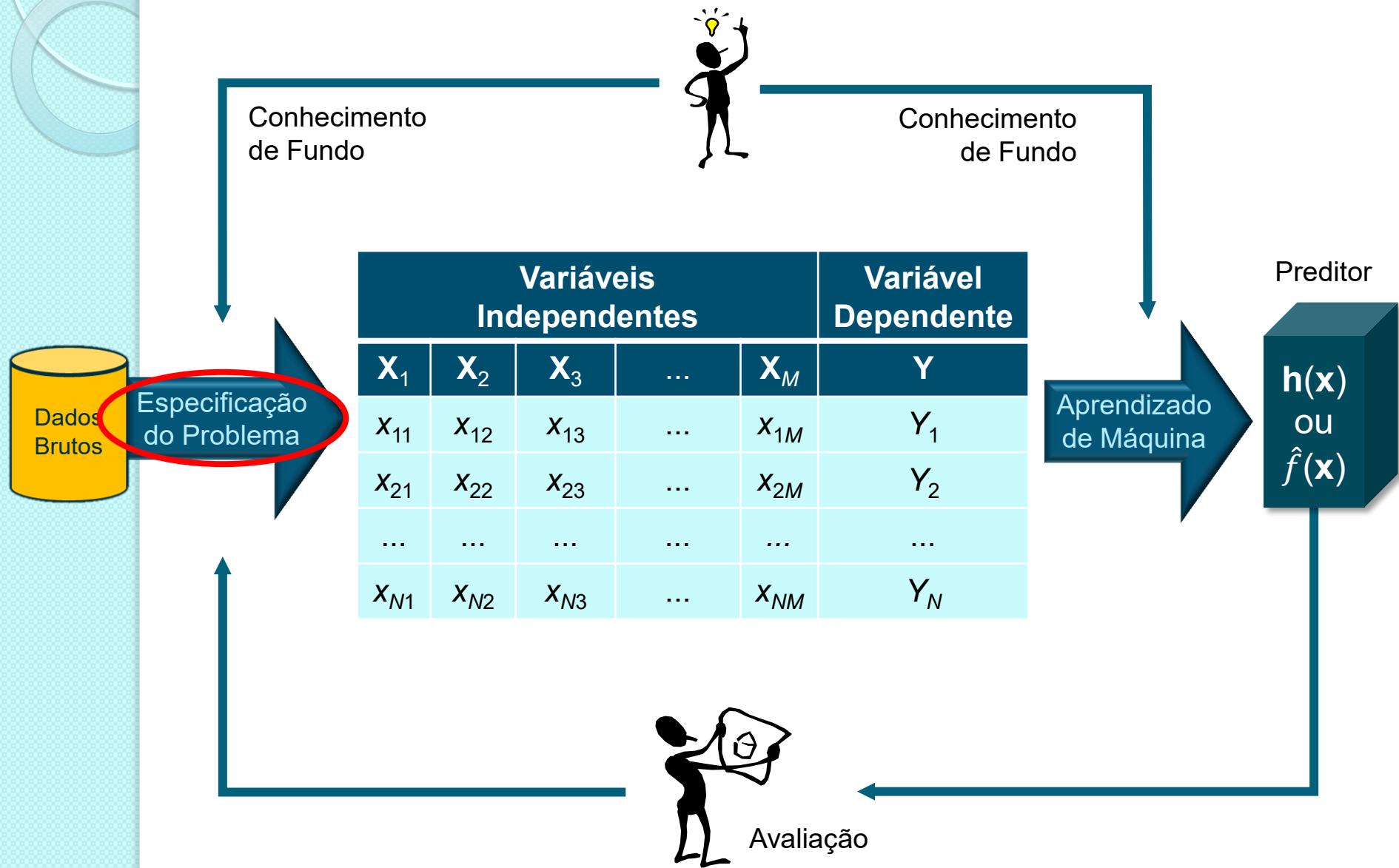
Pré-Processamento de Dados



O que veremos?

- Integração de dados
- Amostragem de dados
- Dados desbalanceados
- Limpeza de dados
- Transformação de dados
- Redução de dimensionalidade

Modelos Preditivos



Integração de Dados

- Dados podem estar distribuídos
 - Necessidade de integração entre os dados
- Necessidade de identificação de entidade
 - Quais os objetos que estão nos diferentes conjuntos a serem combinados?
 - Busca por atributos comuns nos conjuntos
- Dificuldades na integração
 - Atributos correspondentes com nomes diferentes
 - Domínios distintos de atributos equivalentes
 - Necessidade de conversão de dados
 - ...

Grande Quantidade de Dados

- Algoritmos de AM necessitam em geral que todos os dados estejam em memória
- Necessidade de balanço entre eficiência computacional e acurácia
 - Quanto mais dados utilizados maior tende a ser a acurácia e menor a eficiência computacional
- Possibilidade de uso de amostra ou subconjunto de dados

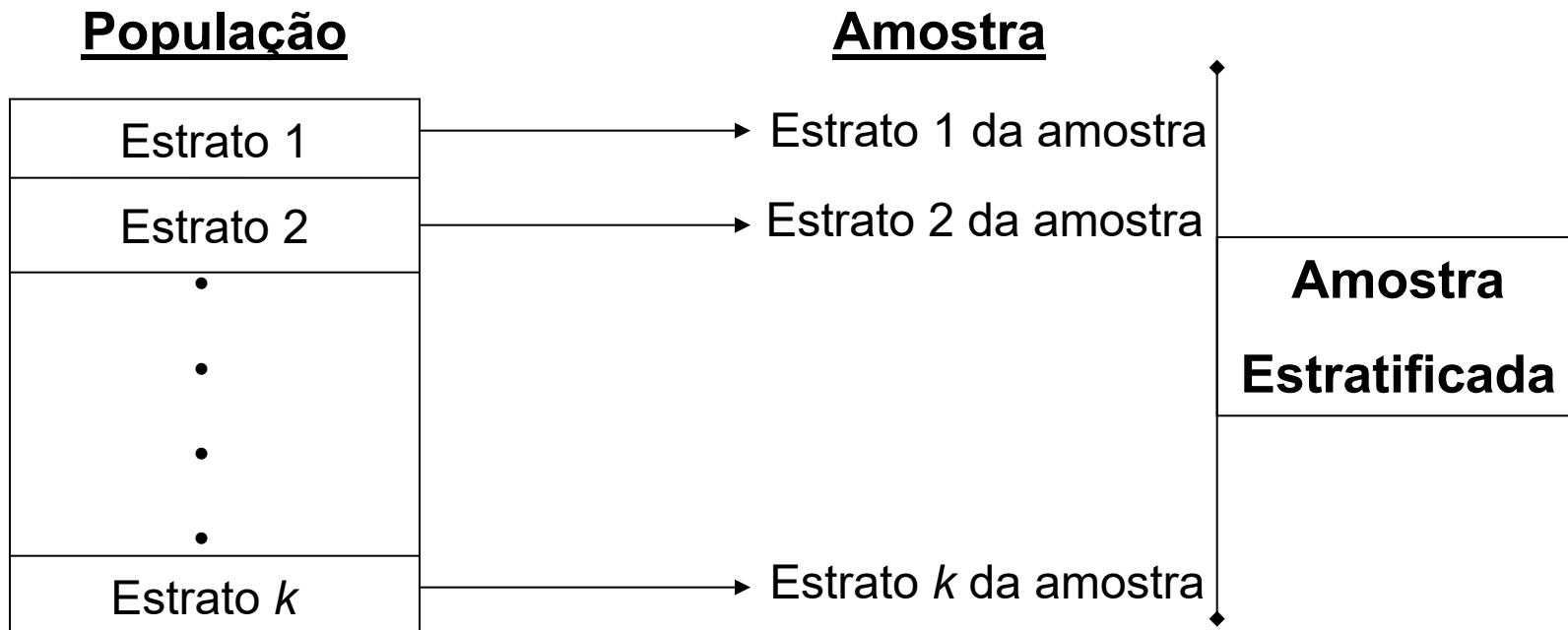
Base de Dados Completa x Amostra Representativa

- Vantagens do uso de amostras representativas – selecionadas aleatoriamente:
 - Minerar uma amostra aleatória representativa é mais fácil e mais eficiente e pode produzir resultados precisos, similares aos produzidos usando toda a base de dados
 - Quando amostras são usadas, a exploração e visualização dos dados ajudam a ganhar conhecimento
 - Conduzem a modelos mais rapidamente e com maior precisão
 - Amostras representativas necessitam relativamente de menor tempo para limpeza, exploração, desenvolvimento e validação modelos – menor custo no processo de MD

Técnicas de Amostragem

- Amostragem aleatória simples
 - Com reposição dos exemplos
 - Probabilidade de seleção dos exemplos se mantém constante
 - Sem reposição dos exemplos
 - Exemplos não se repetem
 - Resultados semelhantes quando amostra é bem menor que o conjunto de dados original
- Amostragem aleatória estratificada
 - É respeitada a distribuição dos exemplos no atributo classe
- Amostragem progressiva
 - Inicia-se o processo com amostra pequena, e vai aumentando conforme a acurácia melhora

Obtenção da amostra



Dados Desbalanceados

- Ex:
 - 80% de pacientes saudáveis
 - 20% de pacientes doentes
- Acurácia:
 - Para ser aceitável, deve ser maior que a atribuição de todos os exemplos à classe majoritária
 - No ex, uma acurácia aceitável deve ser maior que 80%

Dados Desbalanceados

- Balanceamento artificial de dados:
 - Redefinição do tamanho do conjunto de dados
 - Replicação de exemplos da classe minoritária
 - Remoção de exemplos da classe majoritária
 - Utilização de diferentes custos de classificação
 - Dependente do algoritmo de aprendizado aceitar matriz de custo
 - Indução de um modelo para cada classe
 - Aprendizado de uma única classe para a classe minoritária

Limpeza dos Dados

- Dados Incompletos
 - Motivos:
 - Atributo não foi considerado importante na coleta dos primeiros dados
 - E-mail não era comum na década de 90, por exemplo
 - Desconhecimento do valor no preenchimento dos valores do objeto
 - Falta de necessidade de apresentar um valor
 - Inexistência do valor
 - Número de partos para pacientes do sexo masculino

Limpeza dos Dados

- Dados Incompletos
 - Pode gerar erro no processo de indução
 - O que fazer?
 - Eliminar objetos com valores ausentes
 - Definir e preencher manualmente os valores faltantes
 - Utilizar heurística para preencher os valores automaticamente
 - Média, moda ou mediana
 - Utilizar algoritmos de aprendizado para predizer os valores do atributo

Limpeza dos Dados

- Dados Inconsistentes
 - Exemplos replicados com valores na classe diferentes
 - Verificação manual da inconsistência
 - Remoção dos exemplos
- Dados Redundantes
 - Replicação de exemplos
 - Remoção das redundâncias

Limpeza dos Dados

- Dados com Ruídos
 - Dados que aparentemente não pertencem à distribuição dos dados
 - Variância ou erro aleatório no valor gerado ou medido
 - Ex: Peso de uma pessoa = 300
 - É ruído ou é real?
 - Verificação com especialista pode ser indicado
 - Diferentes técnicas para lidar com ruídos

Transformação dos Dados

- Após integrar os atributos e selecionar do domínio, deve-se fazer análise:
 - Tipo dos dados aceitos pelo algoritmo de AM

Tipos de Atributos

- Classificados em dois tipos
 - 1. Contínuos
 - Variáveis numéricas que descrevem quantidades e tem uma escala contínua
 - Média e desvio padrão são medidas para quantificar uma medida de tendência central e dispersão, respectivamente
 - Exs: Total de vendas por consumidor, custo por produto, o total de vendas por produto, o número de unidades adquiridas por cada consumidor, a renda anual por consumidor...
 - Uma variável contínua é necessária para modelagem preditiva em regressão linear múltipla e redes neurais artificiais

Tipos de Atributos

2. Categóricos

- Podem ser classificados como:
 - Ordinal
 - Variável com rank (ordenação) categorizada ou discreta com mais de dois níveis
 - Exemplo: grupo de idades
 - Nominal
 - Variável categorizada com mais de dois níveis e não ordenada
 - A moda é a estatística mais utilizada para tendência central, e o estudo da distribuição de freqüência é a técnica mais utilizada para descrição
 - Ex: diferentes tipos de serviços bancários, raça
 - Binárias
 - Uma variável binária com apenas dois níveis
 - Ex: bom e ruim, vendeu e não vendeu

Transformando os Dados

Valores Contínuos – Números

- Normalização
 - Os valores resultantes são dados dentro de uma certa faixa
 - Ex: [0,1]
 - Esta transformação não muda a forma da distribuição dos valores
 - Normalização pode ser útil quando usamos técnicas que realizam operações de multiplicação sobre os dados
 - Exs: Redes Neurais e Cluster Analysis
 - Árvores de Decisão não são afetadas pela normalização, pois não muda a ordem dos valores

$$v' = \frac{Valor - Min}{(Máx - Min)} (1,0 - 0,0) + 0,0$$

Transformando os dados

Valores Contínuos – Números

- Caixas com igual largura (Equal-width binning) – Discretização
 - Transforma as variáveis em faixas de tamanhos fixos
 - A variável resultante tem aproximadamente a mesma distribuição da variável original
 - Ex: se os valores observados estão entre 0-100, podemos criar 5 faixas:
 - Comprimento = $(100 - 0)/5 = 20$
 - Faixas [0-20], (20-40], (40-60], (60-80], (80-100]
- Existem outras técnicas de discretização

Transformando os dados Datas e Tempos

- Um formato típico para datas e tempo é o número de dias ou horas desde alguma data no passado
- Neste caso os algoritmos tratam datas como números

Transformando os Dados Variáveis Categorizadas

- Os algoritmos trabalham melhor com poucas categorias
- Redes Neurais e Algoritmos de Clusterização entendem variáveis quantitativas
 - Na presença de variáveis categorizadas, utilizar variáveis binárias

Atributos Irrelevantes x Algoritmos de AM

- Algoritmos mais afetados
 - Indutores de árvores e regras de decisão
 - Continuamente reduzem a quantidade de dados em que baseiam suas escolhas
 - Indutores baseados em instâncias (e.g., k-NN)
 - Sempre trabalha com vizinhanças locais
 - Leva em consideração apenas algumas poucas instâncias (k)
 - Foi mostrado que, para se alcançar um certo nível de desempenho, a quantidade de instâncias necessária cresce exponencialmente com o número de atributos irrelevantes (Mitchell, 1997)

Seleção de atributos antes do aprendizado

- Melhora o desempenho preditivo
- Acelera o processo de aprendizado
 - O processo de seleção de atributos, às vezes, pode ser muito mais custoso que o próprio processo de aprendizado
 - Quando somarmos os custos das duas etapas, pode não haver vantagem
- Produz uma representação mais compacta do conceito a ser aprendido
 - O foco será nos atributos que realmente são importantes para a definição do conceito

Redução de Dimensionalidade

- Parte de uma área chamada de Redução de Dados
 - Necessário devido à “Maldição da Dimensionalidade” (*Curse of Dimensionality*)
- Objetivo:
 - Obtenção de uma representação reduzida em volume mas que produz resultados de análise idênticos ou similares
 - Melhora o desempenho dos modelos de aprendizado
- Duas abordagens:
 - Agregação
 - Seleção de Atributos

Agregação

- Combinação dos atributos originais por meio de funções lineares ou não lineares
- Análise de Componentes Principais – PCA
 - Correlaciona estatisticamente os exemplos, reduzindo a dimensionalidade do conjunto de dados original pela eliminação de redundâncias
- Agregação leva à perda dos dados originais
 - Em algumas áreas é importante a preservação dos dados originais para interpretar os resultados obtidos
 - Ex: finanças, medicina, biologia, etc

Seleção de Atributos

- Muitos algoritmos de aprendizado são projetados de modo a selecionar os atributos mais apropriados para a tomada de decisão
 - Algoritmos de indução de árvores de decisão são projetados para:
 - Escolher o atributo mais promissor para particionar o conjunto de dados
 - Nunca selecionar atributos irrelevantes
 - **Pergunta:** Mais atributos implica em maior poder discriminatório?

Resposta...

- Em algum momento durante a geração das árvores:
 - O atributo irrelevante é escolhido
 - Isto causa erros aleatórios durante o teste
- Por que o atributo irrelevante é escolhido?
 - Na medida em que a árvore é construída, menos e menos dados estão disponíveis para auxiliar a escolha do atributo
 - Chega a um ponto em que atributos aleatórios parecem bons apenas por acaso
 - A chance disto acontece aumenta com a profundidade da árvore

Atributos Irrelevantes

- Adição de atributos irrelevantes às instâncias de uma base de dados, geralmente, “confunde” o algoritmo de aprendizado
- Experimento (exemplo)
 - Indutor de árvores de decisão (C4.5)
 - Base de dados D
 - Adicione às instâncias em D um atributo binário cujos valores sejam gerados aleatoriamente
- Resultado
 - A acurácia da classificação cai entre 5% e 10% nos conjuntos de testes

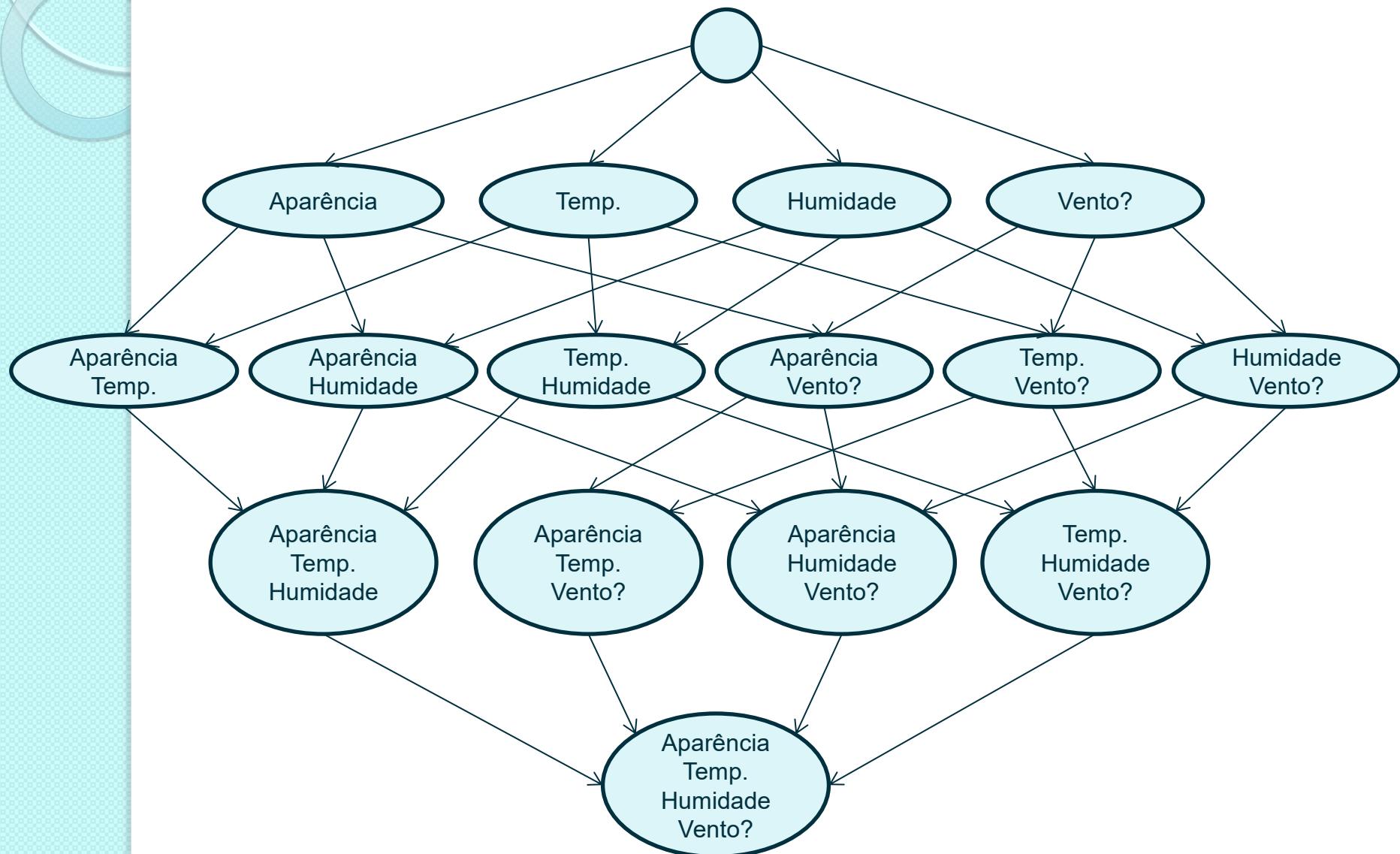
Seleção Automática

- Implica em uma busca no “espaço” de atributos
 - Quantos subconjuntos há?
 - 2^N , em que N é o número total de atributos
 - Portanto, na maioria dos casos práticos, uma busca exaustiva não é viável
 - Solução: busca heurística

Exemplo: O Conjunto Voyage

Ex. #	Aparência	Temp.	Humi.	Vento?	Viajar?
T ₁	sol	25	72	sim	sim
T ₂	sol	28	91	sim	não
T ₃	sol	22	70	não	sim
T ₄	sol	23	95	não	não
T ₅	sol	30	85	não	não
T ₆	nuvens	23	90	sim	sim
T ₇	nuvens	29	78	não	sim
T ₈	nuvens	19	65	sim	não
T ₉	nuvens	26	75	não	sim
T ₁₀	nuvens	20	87	sim	sim
T ₁₁	chuva	22	95	não	sim
T ₁₂	chuva	19	70	sim	não
T ₁₃	chuva	23	80	sim	não
T ₁₄	chuva	25	81	não	sim
T ₁₅	chuva	21	80	não	sim

Exemplo: Espaço de Atributos



Seleção de Atributos

- Manual
 - Melhor método se for baseado em um entendimento profundo sobre:
 - O problema de aprendizado
 - O significado de cada atributo
 - Ex: Diagnóstico clínico de paciente
 - Id e Nome são **irrelevantes**
 - Idade, Sexo, Peso, Manchas, Temp. Corporal, Núm. Internações e Diagnóstico são **relevantes**

Avaliação dos Atributos

- Atributos com apenas um valor
 - Ex: Todos os clientes são do sexo feminino
 - Atributo sexo possui somente o valor F
 - Não contém informação que possa fazer a distinção entre linhas da base de dados
 - Como ela não representa informação dever ser desprezada para MD
- Atributo com grande predominância de apenas um único valor
 - Questão: quando este(s) atributo(s) pode(m) ser desprezado(s)?
 - Praticamente todos os registros devem ter o mesmo valor
 - Poucos registros com valores diferentes e que representam uma porção desprezível dos dados – muito pequena para ter importância

Avaliação dos Atributos

- Atributo com valores únicos
 - Variáveis categóricas que, para cada linha, assumem um valor diferente
 - Ex: nome do cliente, endereço, número do telefone, etc
 - Tais atributos não têm valor preditivo para MD
 - Objetivo é generalizar!
- Atributos sinônimos com a variável alvo
 - Quando uma coluna é altamente correlacionada com a coluna alvo isto pode significar que ela é sinônimo
 - Ex: se um cliente está com o seu cartão de crédito em inatividade, pode indicar que ele não vai responder a uma campanha de marketing
 - Variáveis sinônimas com a variável alvo devem ser ignoradas da análise

Avaliação dos Atributos

- Não é recomendável incluir variáveis descritoras contínuas altamente correlacionadas
 - Coeficiente de correlação $\geq 0,95$
 - Podem produzir modelos instáveis que trabalham bem somente com a particular amostra usada

Seleção de Atributos

- Automático
 - Filtros: método usado antes do processo de aprendizado para selecionar o subconjunto de atributos
 - Wrappers: o processo de escolha do subconjunto de atributos está “empacotado” junto com o algoritmo de aprendizado sendo utilizado

Abordagens para Seleção de Atributos

- Embutida
- Filtros
- Wrapper

Abordagens para Seleção de Atributos – Embutida

- O próprio algoritmo de aprendizado faz a seleção dos atributos considerados relevantes
 - Ex: Árvores de Decisão

Abordagens para Seleção de Atributos – Filtros

- O processo de escolha do subconjunto acontece antes do processo de aprendizado
- Não leva em consideração qual o algoritmo de aprendizado será utilizado
 - Exemplos de filtros:
 - Verificação de correlação entre atributos
- Mais rápidas que Wrapper

Abordagens para Seleção de Atributos – Wrapper

- O processo de escolha do subconjunto de atributos está “empacotado” junto com o algoritmo de aprendizado sendo utilizado
 - Algoritmo de aprendizado utilizado como caixa-preta para seleção
 - Para cada subconjunto de atributos:
 - O algoritmo é consultado e o subconjunto que apresentar melhor combinação entre redução de taxa de erro e redução de atributos é selecionado
- Abordagens mais lentas que os filtros

Filtros x Wrappers

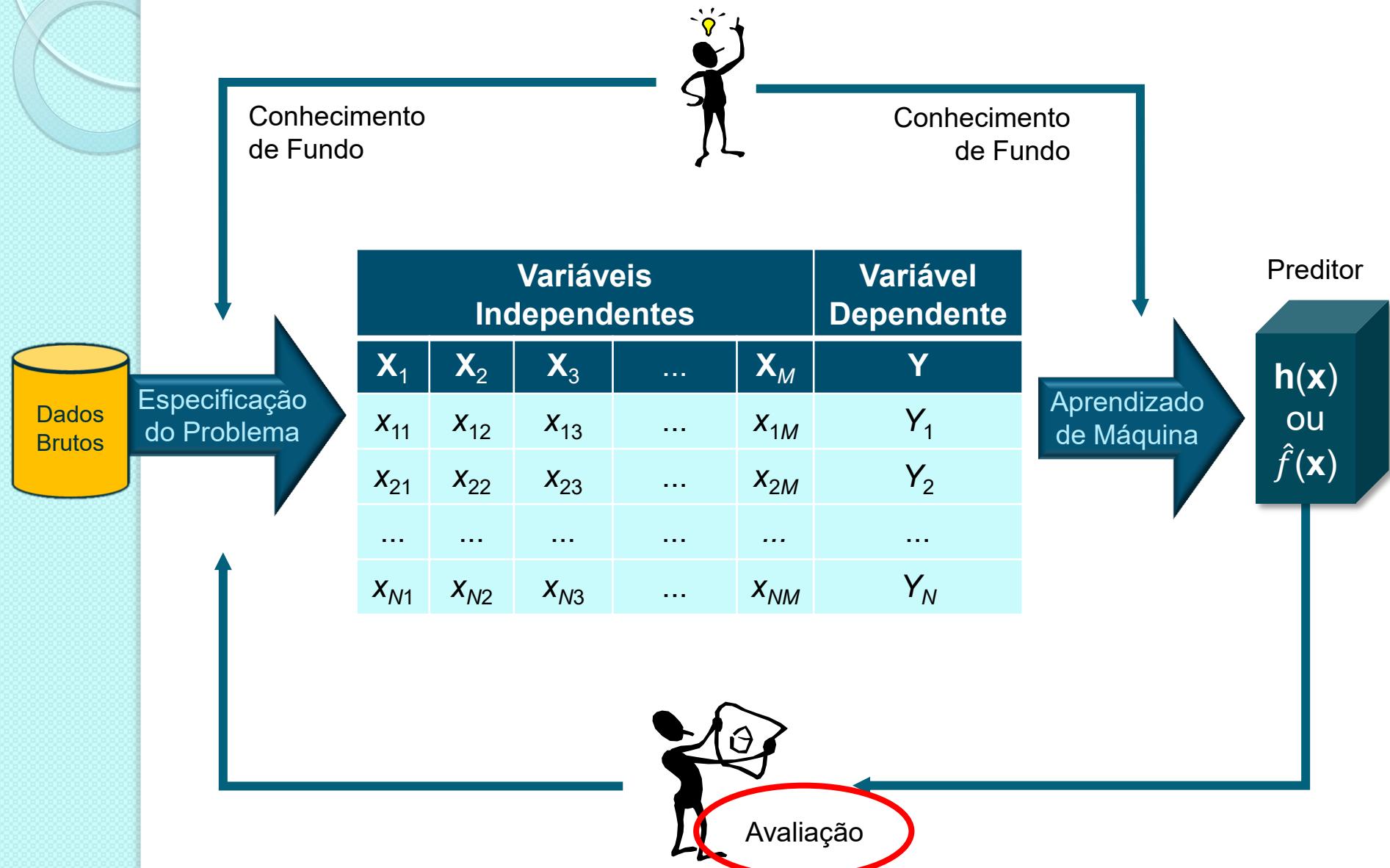
- Wrappers:
 - Alternativa simples e poderosa para selecionar atributos
 - Criticadas por serem técnicas de força bruta
 - Estratégias eficientes de busca têm sido utilizadas

Vantagens e Desvantagens – Filtros

- Vantagens dos filtros:
 - Processo de seleção não depende de indutor – atributos selecionados podem ser utilizados por diferentes indutores
 - Heurísticas utilizadas para avaliação são pouco custosas computacionalmente
 - Conseguem lidar eficientemente com grande quantidade de dados
- Desvantagens dos filtros:
 - Independência em relação ao algoritmo de AM
 - O viés do algoritmo de seleção de atributos não interage com o viés do algoritmo de aprendizado, podendo levar à construção de classificadores com desempenho aquém do desejado
- Wrappers:
 - Alternativa simples e poderosa para selecionar atributos
 - Criticadas por serem técnicas de força bruta
 - Estratégias eficientes de busca têm sido utilizadas

Avaliação de Modelos Preditivos

Modelos Preditivos



Erro e Precisão de um Classificador

Recordando a notação adotada

- Exemplo $(x, y) = (x, f(x))$
- Atributos: x
- Classe (rotulada): $y = f(x)$
- Classe (classificada): $h(x)$
- n é o número de exemplos

Erro e Precisão de um Classificador

- Classificação

$$ce(h) = \frac{1}{n} \sum_{i=1}^n \|y_i \neq h(x_i)\| \quad (\text{erro})$$

$$ca(h) = 1 - ce(h) \quad (\text{precisão})$$

- O operador $\| E \|$ retorna:
 - 1 se E é verdadeiro
 - 0 se E é falso

Avaliação de um Classificador: Matriz de Confusão para 2 Classes

Class label	predicted C_+	predicted C_-	Class error rate	Total error rate
true C_+	T_P	F_N	$\frac{F_N}{T_P + F_N}$	$\frac{F_P + F_N}{n}$
true C_-	F_P	T_N	$\frac{F_P}{F_P + T_N}$	

T_P = True Positive

F_N = False Negative

F_P = False Positive

T_N = True Negative

$n = (T_P + F_N + F_P + T_N)$

Erro e Precisão de um Regressor

- Regressão

$$\text{pe-mad}(h) = \frac{1}{n} \sum_{i=1}^n |y_i - h(x_i)|$$

$$\text{pe-mse}(h) = \frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2$$

$$\text{pe-rmse}(h) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - h(x_i))^2}$$

Avaliação

- Posso calcular as medidas ditas acima com os dados usados para treinar o modelo?
- A resposta é não...
 - Avaliar o modelo com os dados utilizados para treino é oferecer uma estimativa aparente do comportamento preditivo
- É necessário fazer reamostragem dos dados para avaliação dos modelos

Estimativa de Medida Utilizando Reamostragem

- Para se estimar o erro verdadeiro de um classificador a amostra para teste deve ser aleatoriamente escolhida
- Amostras não devem ser pré-selecionadas de nenhuma maneira
- Para problemas reais, tem-se uma amostra de uma única população, de tamanho n , e a tarefa é estimar o erro verdadeiro para essa população

Métodos para estimar erro de um classificador

- Holdout
- k -fold cross-validation
- k -fold cross-validation estratificado
- Leave-one-out
- Bootstrap

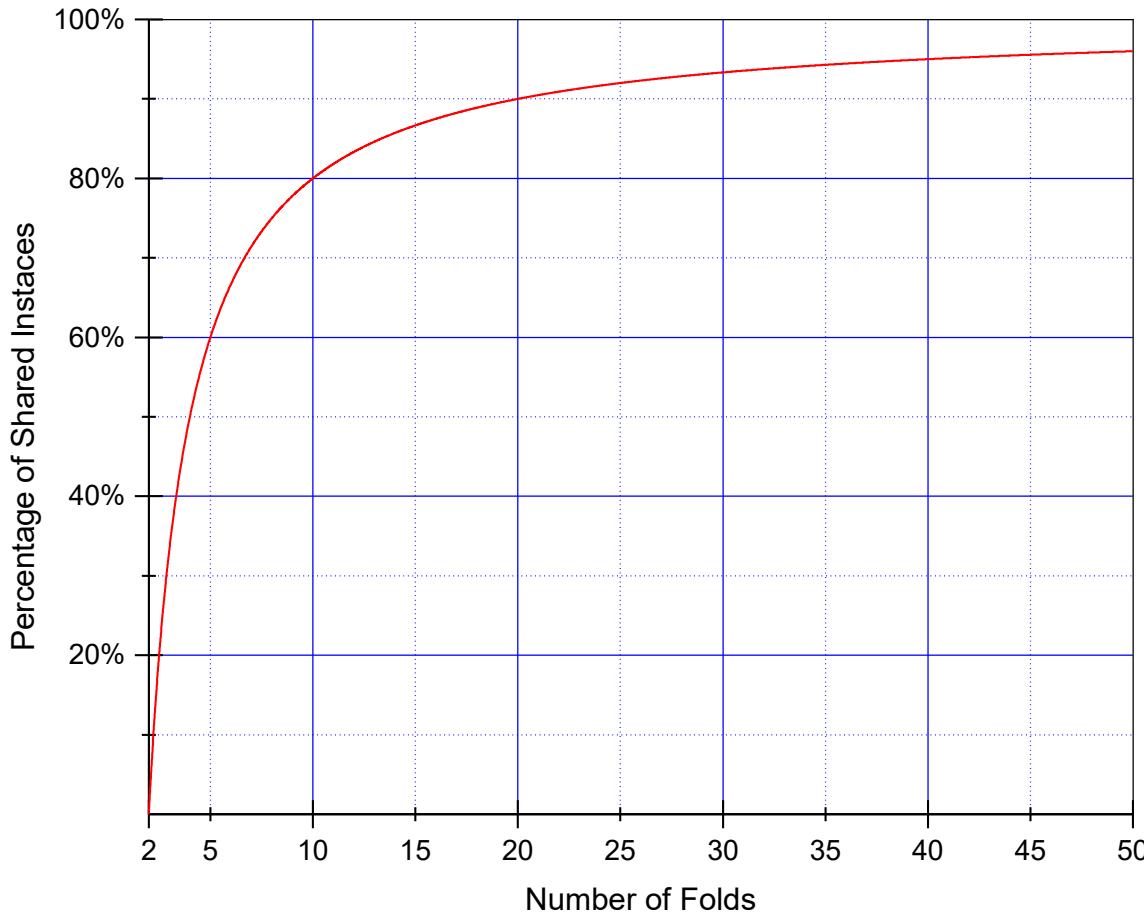
Holdout

- Divide os dados em uma porcentagem fixa p para treinamento e $(1-p)$ para teste
 - geralmente $p=2/3$ e $(1-p)=1/3$
 - para que os resultados não dependam da divisão dos dados (exemplos), pode-se calcular a média de vários resultados de holdout

k-fold cross-validation

- Os exemplos são aleatoriamente divididas em k partições (*folds*) de tamanho aproximadamente igual (n/k)
- Os exemplos de $(k-1)$ folds são independentemente usados no treinamento e os classificadores obtidos são testados com o fold remanescente
- O processo é repetido k vezes, e a cada repetição um *fold* diferente é usado para teste
 - O erro do cross-validation é a média dos erros dos k folds

Relação entre o número de Folds e a porcentagem de exemplos compartilhadas



k-fold cross-validation estratificado

- É similar ao cross-validation mas no processo de geração dos folds a distribuição das classes no conjunto de exemplos é levada em consideração durante à amostragem.
- Ex:
 - Se conjunto de exemplos tiver duas classes com uma distribuição de 80% para uma classe e 20% para outra, cada fold também terá essa proporção

Leave-one-out

- Para um exemplo de tamanho n , um classificador é gerado usando $n-1$ exemplos, e testado no exemplo remanescente.
- O processo é repetido n vezes, utilizando cada um dos n exemplos para teste. O erro é a soma dos erros dos testes para cada exemplo dividido por n
- Caso especial de validação cruzada
- Computacionalmente caro e usado apenas quando o conjunto de exemplos é pequeno

Bootstrap

- Repetir diversas vezes o processo inteiro de classificação e, para estimar quantidades como o bias, p.ex., cada experimento é baseado em um conjunto de treinamento novo, obtido por resampling com reposição do conjunto de dados original
 - Existem muitos estimadores bootstrap

e0 bootstrap

- O conjunto de treinamento é de n exemplos (mesmo tamanho do dataset) extraídos com reposição
- As exemplos que não aparecem no conjunto de treinamento são colocadas no conjunto de teste

Comparando Modelos

- Não há um único bom algoritmo de AM para todas as tarefas
- É importante conhecer o poder e as limitações de indutores diferentes
- Na prática, devemos testar algoritmos diferentes, estimar sua precisão e escolher entre os algoritmos aquele que apresentar maior precisão, por exemplo, para um domínio específico

Metodologia de Avaliação (Russel e Norvig, 2003)

- 1 Coletar um conjunto de exemplos, de preferencia sem “ruido”.
- 2 Dividir randomicamente o conjunto de exemplos em um conjunto de teste e um conjunto de treinamento.
- 3 Aplicar um ou mais indutores ao conjunto de treinamento, obtendo uma hipótese h para cada indutor
- 4 Medir a performance dos classificadores com o conjunto de teste
- 5 Estudar a eficiência e robustez de cada indutor, repetindo os passos 2 a 4 para diferentes conjuntos e tamanhos do conjunto de treinamento
- 6 Se estiver propondo um ajuste ao indutor, voltar ao passo 1

Calculando Média e Desvio Padrão usando Resampling

Usando validação cruzada: Dado um algoritmo A, para cada fold i , calculamos o erro $pe(h_i)$, $i = 1, 2, \dots, r$, temos:

$$\hat{\mu} = \frac{1}{r} \sum_{i=1}^r pe(h_i)$$

$$\hat{\sigma}^2 = \frac{1}{r} \left[\frac{1}{r-1} \sum_{i=1}^r (pe(h_i) - \text{média}(A))^2 \right]$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2(A)}$$

Calculando Média e Desvio Padrão usando Resampling

- Exemplo: Considerando um exemplo de cross-validation 10-fold ($r=10$), para um algoritmo A que apresente os erros 5.5, 11.4, 12.7, 5.2, 5.9, 11.30, 10.9, 11.2, 4.9 e 11.0, então:

$$\text{média}(A) = \frac{90.0}{10} = 9.0$$

$$\text{desvio padrão} = \sqrt{\frac{1}{10(9)} 90.3} = 1.0$$

Comparando dois Algoritmos

$A_S \Rightarrow$ algoritmo standard

$A_P \Rightarrow$ algoritmo proposto

$$\text{média}(A_S - A_P) = \text{média}(A_S) - \text{média}(A_P)$$

$$sd(A_S - A_P) = \sqrt{\frac{sd(A_S)^2 + sd(A_p)^2}{2}}$$

$$ad(A_S - A_P) = \frac{\text{média}(A_S - A_P)}{sd(A_S - A_P)}$$

Comparando dois Algoritmos

- Se $ad(AS-AP) > 0$ AP tem melhor performance que AS
- Se $ad(AS-AP) \geq 2$ AP tem melhor performance que AS com um nível de confiança de 95%.
- Se $ad(AS-AP) \leq 0$ As tem melhor performance que Ap
- Se $ad(AS-AP) \leq -2$ As tem melhor performance que Ap com um nível de confiança de 95%.

Comparando dois Algoritmos

- Exemplo: considerando que $A_S = 9.00 \pm 1.00$ e $A_P = 7.50 \pm 0.80$

$$\text{média}(A_S - A_P) = 9.00 - 7.50 = 1.50$$

$$sd(A_S - A_P) = \sqrt{\frac{1.00^2 + 0.80^2}{2}} = 0.91$$

$$ad(A_S - A_P) = \frac{1.50}{0.91} = 1.65$$

Como $ad(A_S - A_P) < 2$, A_P não tem uma performance significantivamente melhor que A_S , com um nível de confiança de 95%.

Para o trabalho

- Utilizar os seguintes algoritmos:
 - Redes Neurais
 - Árvores de Decisão
 - K-NN
 - NB
- Comparar os modelos dois a dois usando o teste t

Referências Bibliográficas

- FACELLI, K.; LORENA, A.C.; GAMA, J.; CARVALHO, A.C.P.L.F. *Inteligência Artificial -- Uma Abordagem de Aprendizado de Máquina.* Ed. LTC, 2011. – Caps. 4, 6, 7, 10, 11, 12.
- REZENDE, S.O. *Sistemas Inteligentes: Fundamentos e Aplicações.* Ed. Manole, 2003.
- MITCHELL, T. M. *Machine Learning.* McGraw-Hill, 1997
- DUDA, R.O.; HART, P.E.; STORK, D.G. *Pattern Classification, 2nd Edition.* Ed. Wiley, 2001.