Analise e Seleção de Variáveis

Tópicos

- Por que atributos irrelevantes são um problema
- Quais tipos de algoritmos de aprendizado são afetados
- Abordagens automáticas
  - *Wrapper*
  - Filtros

Introdução

- Muitos algoritmos de AM 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
  - Mais atributos implica em maior poder discriminatório?

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
    - Em geral, de 5% a 10% nos conjuntos de testes

Explicação

- 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 *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

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 processo de aprendizado
  - Ou seja, 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

Analise e Seleção de Variáveis

- Parte de uma área chamada de Redução de Dados
- Obtenção de uma representação reduzida em volume mas que produz resultados de análise idênticos ou similares
- Melhora o desempenho dos modelos de aprendizado
- Objetivo: Eliminar atributos redundantes ou irrelevantes

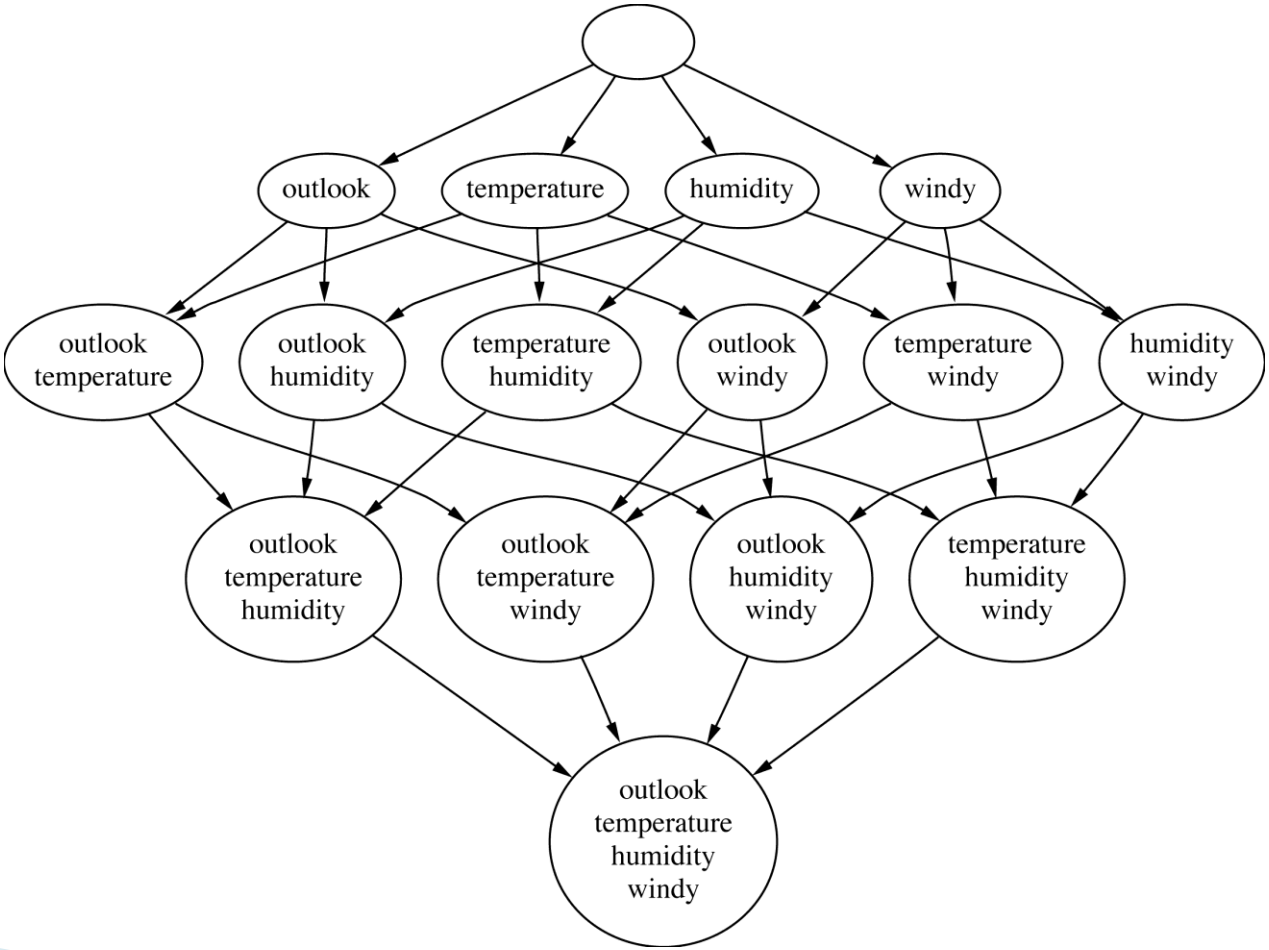
Métodos de Seleção de Atributos

- **Manual**
  - Melhor método se for baseado em um entendimento profundo sobre ambos:
    - O problema de aprendizado
    - O significado de cada atributo
- **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

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: Espaço de Atributos



Busca Heurística no Espaço de Atributos

- **Busca para Frente (Seleção *Forward*)**
  - A busca é iniciada sem atributos e os mesmos são adicionados um a um
  - Cada atributo é adicionado isoladamente e o conjunto resultante é avaliado segundo um critério
  - O atributo que produz o melhor critério é incorporado

Busca Heurística no Espaço de Atributos

- **Busca para trás (Eliminaçao *Backward*)**
  - Similar a Seleção *Forward*
  - Começa com todo o conjunto de atributos, eliminando um atributo a cada passo
- Tanto na Seleção *Forward* quanto na Eliminação *Backward*, pode-se adicionar um viés por subconjuntos pequenos
  - *Por exemplo, pode-se requerer não apenas que a medida de avaliação crescer a cada passo, mas que ela cresça mais que uma determinada constante*

Busca Heurística no Espaço de Atributos

- Outros métodos de busca
  - Busca bidirecional
  - *Best-first search*
  - *Beam search*
  - Algoritmos genéticos
  - .....

Abordagens para Seleção de Atributos

- Filtros
  - O processo de escolha do subconjunto acontece antes do processo de aprendizado

- *Wrapper*
  - O processo de escolha do subconjunto de atributos está "empacotado" junto com o algoritmo de aprendizado sendo utilizado

Analise e Seleção de Variáveis

- Métodos Dependentes do Modelo (Wrapper)
- Métodos Independentes do Modelo (Filter)

Exemplo: Filtros

- Uso de uma indutor de árvores de decisão (AD) como filtro para o k-NN
  - 1) Aplique um indutor de AD para todo o conjunto de treinamento
  - 2) Selecione o subconjunto de atributos que aparece na AD
  - 3) Aplique o k-NN a apenas este subconjunto
- A combinação pode apresentar melhores resultados do que cada método usando individualmente

Filtros

- Abordagens
  - baseada nas características gerais dos dados
  - Encontrar o menor subconjunto que separe os dados
  - Utilizar diferentes esquemas de aprendizado.
    - Usar os atributos que aparecem no c4.5, 1R

Wrapper

- **Busca para Frente (Seleção *Forward*) + Naive Bayes**
  - (1) Inicialize com o conjunto vazio S={}
  - (2) Resultado_S=0
  - (2) Para cada atributo s_i que não esteja em S
    - Avalie o resultado de (S U s_i): Resultado_ s_i
  - (3) Considere o atributo com maior Resultado_ s_i
    - SE (Resultado_ s_i > Resultado_S)
      ENTAO
      (S=S U s_i) & (Resultado_S= Resultado_ s_i)
      Volte para o Passo (2)
      SENAO
      Pare

Transformação de Dados

Discretização de Variáveis Contínuas

- Transforma atributos contínuos em atributos categóricos
- Absolutamente essencial se o método inteligente só manuseia atributos categóricos
- Em alguns casos, mesmo métodos que manuseiam atributos contínuos têm melhor desempenho com atributos categóricos

Discretização de Variáveis Contínuas

- Diversos métodos de discretização
- Discretização pelo Método 1R (1-rule)
- Discretização Não-supervisionada

Discretização de Variáveis Contínuas

- Discretização pelo Método 1R (1-rule)
- Sub-produto de uma técnica de extração automática de regras
- Utiliza as classes de saída para discretizar cada atributo de entrada separadamente
- Ex: Base de dados hipotética de meteorologia x decisão de realizar ou não um certo jogo

Discretização pelo Método 1R (1-rule)

- Base de Dados Meteorológicos
- **Tempo Temperatura Umidade Vento Jogar? (CLASSE)**

Sol 85 85 Não **Não**
Sol 80 90 Sim **Não**
Nublado 83 86 Não **Sim**
Chuva 70 96 Não **Sim**
Chuva 68 80 Não **Sim**
Chuva 65 70 Sim **Não**
Nublado 64 65 Sim **Sim**
Sol 72 95 Não **Não**
Sol 69 70 Não **Sim**
Chuva 75 80 Não **Sim**
Sol 75 70 Sim **Sim**
Nublado 72 90 Sim **Sim**
Nublado 81 75 Não **Sim**
Chuva 71 91 Sim **Não**

Discretização pelo Método 1R (1-rule)

Primeiro passo: ordenar pela coluna Temperatura

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Segundo passo: discretizar pela Classe de saída

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Terceiro passo: ajustar divisões

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72			Não
Nublado	72			Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

1
2
3
4
5
6
7
8

MUITAS DIVISÕES!

Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Quarto passo: mínimo de valores da maior classe (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72 (2)	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (3)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72 (2)	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (3)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71	91	Sim	Não
Sol	72	95	Não	Não
Nublado	72	90	Sim	Sim
Chuva	75	80	Não	Sim
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Quarto passo: mínimo de valores (ex: 3)

Tempo	Temperatura	Umidade	Vento	Jogar? (CLASSE)
Nublado	64	65	Sim	Sim
Chuva	65	70	Sim	Não
Chuva	68 (1)	80	Não	Sim
Sol	69	70	Não	Sim
Chuva	70	96	Não	Sim
Chuva	71			
Sol	72			
Nublado	72			
Chuva	75			
Sol	75	70	Sim	Sim
Sol	80	90	Sim	Não
Nublado	81 (2)	75	Não	Sim
Nublado	83	86	Não	Sim
Sol	85	85	Não	Não

Categoria 1: Temperatura ≤ 77.5

Categoria 2: Temperatura > 77.5

Discretização de Variáveis Contínuas

- Discretização Não-Supervisionada
  - O método 1R é supervisionado. Considera a variável de saída (classe) na discretização
- Métodos Não Supervisionados consideram somente o atributo a ser discretizado
  - São a única opção no caso de problemas de agrupamento (clustering), onde não se conhecem as classes de saída

- Três abordagens básicas:
  - Número pré-determinado de intervalos
    - uniformes (equal-interval binning)
  - Número uniforme de amostras por intervalo
    - (equal-frequency binning)
  - Agrupamento (clustering): intervalos arbitrários

Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes
  - (equal-interval binning)
- No exemplo (temperatura):
  64 65 68 69 70 71 72 72 75 75 80 81 83 85
- Bins com largura 6: $x \leq 60$
  $60 < x \leq 66$
  $66 < x \leq 72$
  $72 < x \leq 78$
  $78 < x \leq 84$
  $84 < x \leq 90$

Métodos de Discretização Não Supervisionada

- Número pré-determinado de intervalos uniformes
  - (equal-interval binning)
- No exemplo (temperatura):
  64 65 68 69 70 71 72 72 75 75 80 81 83 85
- Bins com largura 6: $x \leq 60$: n.a.

  $60 < x \leq 66$: 64, 65
  $66 < x \leq 72$: 68, 69, 70, 71, 72, 72
  $72 < x \leq 78$: 75, 75
  $78 < x \leq 84$: 80, 81, 83
  $84 < x \leq 90$: 85

Métodos de Discretização Não Supervisionada

- Equal-interval binning: Problemas
- Como qualquer método não supervisionado, arrisca destruir distinções úteis, devido a divisões muito grandes ou fronteiras inadequadas
- Distribuição das amostras muito irregular, com algumas bins com muitas amostras e outras com poucas amostras

Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- Também chamado de equalização do histograma
- Cada bin tem o mesmo número aproximado de amostras
- Histograma é plano
- Heurística para o número de bins: $\sqrt{N}$
- N = número de amostras

Métodos de Discretização Não Supervisionada

- Número uniforme de amostras por intervalo
  - (equal-frequency binning)
- No exemplo (temperatura):
- 64 65 68 69 |70 71 72 72 |75 75 80| 81 83 85
- 14 amostras: 4 Bins
  - $x \leq 69{,}5$: 64, 65, 68, 69
  - $69{,}5 < x \leq 73{,}5$: 70, 71, 72, 72
  - $73{,}5 < x \leq 80{,}5$: 75, 75, 80
  - $x > 80{,}5$: 81, 83, 85

Métodos de Discretização Não Supervisionada

- Agrupamento (Clustering)
- Pode-se aplicar um algoritmo de agrupamento
- no caso unidimensional
- Para cada grupo (cluster), atribuir um valor discreto

Transformar

Análise de Componentes Principais (PCA)

- Dado um conjunto D com *n* instâncias e *p* atributos (x_1, x_2,..., x_p), uma transformação linear para um novo conjunto de atributos z_1, z_2,..., z_p pode ser calculada como:

$$z_1 = a_{11}\, x_1 + a_{21}\, x_2 + ... + a_{p1}\, x_p$$
$$z_2 = a_{12}\, x_1 + a_{22}\, x_2 + ... + a_{p2}\, x_p$$
$$...$$
$$z_p = a_{1p}\, x_1 + a_{2p}\, x_2 + ... + a_{pp}\, x_p$$

- Componentes Principais (PCs) são tipos específicos de combinações lineares que são escolhidas de tal modo que z_p (PCs) tenham as seguintes características

PCA: Características

- As *p* componentes principais (PC) são não-correlacionadas (independentes)
- As PCs são ordenadas de acordo com quantidade da variância dos dados originais que elas contêm (ordem decrescente)
  - A primeira PC “explica” (contém) a maior porcentagem da variabilidade do conjunto de dados original
  - A segunda PC define a próxima maior parte, e assim por diante
  - Em geral, apenas algumas das primeiras PCs são responsáveis pela maior parte da variabilidade do conjunto de dados
  - O restante das PCs tem uma contribuição insignificante
- PCA é usada em Aprendizado de Máquina principalmente para a redução de dimensionalidade

PCA: Cálculo

- PCA pode reduzida ao problema de encontrar os auto-valores e auto-vetores da matriz de covariância (ou correlação) do conjunto de dados
- A proporção da variância do conjunto de dados originais explicada pela *i*-ésima PC é igual ao *i*-ésimo auto-valor divido pela soma de todos os *p* auto-valores
- Ou seja, as PCs são ordenadas - decrescente - de acordo com os valores dos auto-valores
- Quando os valores dos diferentes atributos estão em diferentes escalas, é preferível usar a matriz de correlação em lugar da matriz de covariância

Análise de Componentes Principais

- Principais Limitações
  - Assume apenas relações lineares entre os atributos
  - A interpretação dos resultados (e.g., classificador gerado) em termos dos atributos originais pode ficar mais difícil