Resumo Inteligência Computacional: P1

Aluno: Marcos Seefelder de Assis Araujo UFRJ

31/10/2016

1 Capítulo 2 - Pré-Processamento

- Análise estatística
- Limpeza dos dados

1.1 Análise monovariável

- **média**
- **variância**: medida de dispersão da amostra
- **desvio padrão**: mais usado para dispersão, tem escala da variável
- **mínimo** e **máximo**
- **mediana**: divide o conjunto de valores da variável em dois subconjuntos iguais
- **percentil**: divide de forma que P% dos valores sejam <= a este valor
- **faixa interquartil**: entre o 25 percentil (1º quartil) e o 75 percentil (3º quartil)
- **moda**

1.2 Análise Multivariada

1.2.1 Matriz de correlações

A correlação é uma medida padronizada da covariância de duas variáveis.

O coeficiente de correlação pode ser entendido como uma medida da redundância da informação representada por um par de variáveis. Variáveis totalmente correlacionadas **(positiva ou negativamente)** representam essencialmente a mesma informação e uma delas poderia ser descartada. Entretanto, exceto no caso de correlação total, não é possível decidir sobre qual variável remover apenas a partir da informação de correlação.

Coeficiente de correlação mede **apenas a dependência linear**, portanto um resultado = 0 não representa que não há relação entre as variáveis, apenas que a relação não é linear.

1.2.2 Matriz de distâncias ou similaridades

Utiliza outra métrica que não é a correlação.

1.3 Limpeza de dados

1.3.1 Padronização de variáveis

- Por **mínimo e máximo**:
  - reduz a escala da variável ao intervalo [0,1]
  - não deve ser utilizado na presença de outliers, pois o valor máximo ou mínimo de alguma variável pode gerar distorções
  - Na presença de outliers, os valores de máximo e mínimo podem ser substituídos pelos percentis 95 e 5
- Por **z-score**:
  - robusta a outliers
  - variável padronizada tem média nula e desvio padrão um
  - em distribuições altamente assimétricas pode gerar uma distorção de escala, pois utiliza as estatísticas de média e desvio padrão, mais apropriadas para distribuições simétricas
- **Transformação do logaritmo**:
  - Para tornar simétricas as distribuições de variáveis

1.3.2 Detecção de outliers

Analisar os *boxplots* para ter intuição da situação de *outliers*.

Utilizar ordenação da distância média entre registros (retirada da *matriz de distâncias*) para encontrar os registros mais distantes de todos.

1.3.3 Valores ausentes

Basicamente existem duas estratégias principais para o tratamento de valores ausentes: a *eliminação do registro* contendo valores ausentes ou o *preenchimento* (imputação) dos dados ausentes.

O *preenchimento* pode ser feito por média, k-vizinhos ou algum outro método.

1.3.4 Transformações Lineares

Em mineração de dados uma transformação linear é utilizada para realizar uma mudança de coordenadas do conjunto de dados para um novo conjunto de coordenadas, em geral de menor dimensionalidade, e com características mais apropriadas ao problema. As variáveis transformadas são combinações lineares das anteriores e, em princípio, não podem ser interpretadas no contexto da aplicação.

1.3.4.1 Análise de Componentes Principais (PCA)

A ACP permite a redução da dimensionalidade do conjunto de dados pela eliminação da redundância provocada pela correlação entre as variáveis. A ACP é realizada por um procedimento de cálculo da matriz de transformação linear ortogonal que, aplicada ao conjunto de dados, gera um novo conjunto de dados de variáveis transformadas.

Considera que todas as variáveis de entrada são numéricas, com distribuição normal.

O número n de componentes principais é escolhido de forma que sua variância total represente uma parcela importante (em geral maior que 85%) da variância total dos dados originais.

1.3.4.2 Decomposição em Valores Singulares (SVD)

O cálculo da SVD é equivalente ao cálculo de decomposição em autovalores de matrizes simétricas. Pode ser utilizada para a remoção de vetores linearmente dependentes e realizar a diminuição de dimensionalidade de maneira similar à ACP.

1.3.5 Análise Discriminante Linear (LDA)

A Análise Discriminante Linear (ADL), assim como a ACP e a SVD, é uma técnica para o cálculo de uma transformação linear. A ADL, entretanto, leva em consideração a informação da classe para o cálculo da transformação de tal forma que, no novo espaço de coordenadas, a separação entre as classes seja máxima.

A transformação da ADL é calculada de forma que o novo sistema de coordenadas produza dados com **máxima variância** *entre classes* e **mínima** *variância intraclasses*.

1.3.6 Análise de Correlação Canônica

Em problemas de classificação, ADL calcula uma transformação linear que leva em consideração a informação das classes para definir um espaço de coordenadas onde a separação das classes é ótima. Em problemas de regressão é possível obter um resultado semelhante com a Análise de Correlação Canônica (ACC).

2 Capítulo 3 - Regressão Linear

analista: primeiro modelo linear, depois compara com não-linear. Limite inferior

2.1 Regressão

Como ajustar um modelo a partir dos dados observados que seja o mais próximo possível da relação real entre as variáveis que geraram a amostra

2.1.1 Modelo linear

Modelo é linear nos **parâmetros** → vetor de saíde é uma combinação linear dos **regressores**. No modelo linear puro, o **vetor de regressores** é igual às **variáveis de entrada**.

2.2 Mínimos quadrados

Algoritmo mais simples para realizar a **regressão linear**: consistem em ajustas os **parâmetros** utilizando o **Erro Médio Quadrático (EMQ)** como critério.

Estatísticamente, *o EMQ pressupões que o vetor de resíduos tem* ***distribuição normal***, ***média nula*** *e* ***variância constante***.

O *MMQ (Método dos Mínimos Quadrados)* produz os **parâmetros de mínima variância** e *se os* ***resíduos*** *do modelo tem* ***distribuição normal*** a solução é a de **máxima verossimilhança**.

Caso hajam vetores **linearmente dependentes** na **matriz de regressores** a solução pode ser **matematicamente instável**.

Parâmetros com **valores muito altos** são indicativo de que o **modelo tem complexidade inadequada**.

2.3 *Overfitting*

Ocorre quando:

- A **complexidade do modelo** é **maior do que a complexidade da relação**
- &
- O **tmanho da amostra** é **pequeno em relação à complexidade do modelo**

2.4 Etapas da criação de um modelo

1. Escolha do **tipo** do modelo (pré-seleção de um conjunto defunções promissor);
2. Seleção da **estrutura** do modelo (limitar o espaço de soluções a um subconjunto através de hipóteses e conhecimento *a priori*);
3. Definição de um princípio indutivo para gerar o modelo, geralmente gerando um **problema de otimização**;
4. Solução do **problema de otimização** obitdo em 3 gerando os **parâmetros ajustados** do modelo

2.5 Bias x variância

2.5.1 *Bias*:

Diferença entre o valor real da função em $x(t)$ e o valor esperado da aproximação naquele ponto.

Acontece quando a complexidade do modelo é inferior à complexidade da função real.

2.5.2 Variância:

Valor esperado da diferença entre uma estimativa do modelo e o valor esperado de todas as estimativas do modelo

É o efeito da complexidade do modelo ser maior que a complexidade real em relação ao tamanho da amostra.

2.6 Regularização

Diminuir o espaço de solução para encontrar uma solução com complexidade menor e melhor relação de *bias* e variância

2.6.1 Tikhonov

Minimização do EMQR (Erro Médio Quadrático Regularizado).

Adição de uma constante na diagonal principal da Matriz de Coeficientes para aumentar sua estabilidade.

2.6.2 Por Componentes Principais

Utiliza SVD para a seleção apenas das colunas linearmente independentes da Matriz de Coeficientes e monta o modelo com um número r de colunas principais dessas, definido pelo usuário.

2.7 Validação

O método de separar o conjunto de treinameto em uma parte para treinamento e outra para validação é muito simples, masfalha no caso de conjuntos pequenos de dados pois o jeito com o qual a partição é feita pode causar variações importantes nas estatísticas de validação. Por isso o recomendado é a técnica de **validação cruzada**.

2.7.1 Validação cruzada

O conjunto de treinamento é dividido em K subconjuntos e a validação é realizada em K ciclos: em cada ciclo, o modelo é ajustado utilizando $K1$ subconjuntos e avaliado no subconjunto correspondente ao ciclo.

3 Capítulo 4 - Modelos Dinâmicos

3.1 Autocovariância e Autocorrelação:

As funções de autocorrelação e de correlação cruzada permitem identificar as características dinâmicas do processo em estudo, conforme modelos de séries temporais e sistemas dinâmicos.

Vamos considerar **processos estacionários**:

3.1.1 Processo Estacionário

- Também chamados de **Sistemas Dinâmicos Inverantes no Tempo**;
- Características estatísticas não se alteram em amostras distintas;
- Distribuição de probabilidade de uma sequência é idêntica à distribuição da sequência defasada de k registros.

3.1.2 Autocovariância:

- É a covariância entre um registro $y(t)$ e outro registro $y(t+k)$;
- Em um **processo estacionário** autocovariância entre dois registros separados de k instantes de tempo é a mesma para qualquer instante t e depende apenas do valor de k, chamado de atraso.

3.1.3 Autocorrelação:

- Valor adimensional;
- Um valor para cada atraso k;
- O conjunto dos valores de autocorrelação para todos os valores de atraso é chamado **função de autocorrelação** (ACF) ou **correlograma**;
- Os valores de autocorrelação são simétricos em relação à $k = 0$.

3.1.4 Covariância e correlação cruzadas:

- Em sistemas dinâmicos em que as entradas e saídas do sistema são observadas, é possível calcular as funções de covariância e de correlação cruzadas;
- A função de correlação cruzada mostra a correlação entre as variáveis de entrada e saída do sistema em relação a diversos valores de atraso.

4 Capítulo 5 - Classificação

Classificador: um modelo capaz de predizer a classe correta de um registro a partir dos valores das variáveis de entrada.

Superfície de decisão: Fronteira entre as regiões.

4.1 Classificação Bayesiana

4.1.1 Tipos de probabilidade

- **Probabilidade *a priori***: Probabilidade de ocorrência da classe C_i na ausência de qualquer observação;
- **Probabilidade condicional:** Distribuição de probabilidades das variáveis x quando a classe observada é C_i;
- **Probabilidade *a posteriori***: Probabilidade de observar a classe C_i conhecendo a observação (evidência) dos valores das variáveis x.

4.1.2 Métodos de decisão

- Pelo mínimo erro de classificação: Minimiza o erro de classificação
- Pelo mínimo risco codicional: Erro global é maior, mas erro é menor na classe de maior custo de classificação incorreta.

4.1.3 Classificadores

4.1.3.1 Classificador Bayesiano Simples

Realiza a decisão por uma das formas apresentada acima, geralmente pelo mínimo erro de classificação.

A distribuição de probabilidade condicional dos registros do conjunto de treinamento é calculada, para cada classe, por uma estimativa que considera as variáveis de entrada independentes. Desta forma, o classificador Bayesiano Simples não leva em consideração a covariância entre as variáveis.

Em geral, a probabilidade condicional é estimada como uma distribuição normal monovariável cuja média e variância são estimados no conjunto de treinamento para a variável $x_i(t)$ nos registros da classe C_j.

A utilização da ACP para eliminar a correlação dos dados nem sempre produz bons resultados, mas é um recurso que pode ser empregado na busca pelo melhor modelo de classificação.

4.1.3.2 Classificador Bayesiano Quadrático

É possível generalizar regras de decisão como **funções discriminantes** em problemas de múltiplas classes.

O classificador Bayesiano com distribuições normais multivariadas (aproximando a distribuição condicional) é chamado de classificador Bayesiano Quadrático.

A estimativa dos parâmetros da equação normal para o cálculo da função discriminante pode se tornar inviável em problemas de alta dimensionalidade, uma vez que são necessários $ p + p(p + 1)/2$ parâmetros para cada classe. Para isso pode ser usada uma diminuição de dimensionalidade através de ACP ou ADL, por exemplo.

4.2 Modelos Lineares de Classificação

Alguns métodos podem ser utilizados diretamente no ajuste de parâmetros do modelo linear, sem a necessidade de hipóteses ou estimativa da distribuição de probabilidades condicional.

4.2.1 Mínimos Quadrados

Equivalente à decisão pelo menor erro ajustando um classificador.

4.2.2 Mínimos Quadrados Ponderado

Efeitos equivalentes a decisão pelo risco condicional.

4.2.3 Regressão Logística

Usa função sigmoide para ajuste de parâmetros.

4.3 Avaliação de Classificadores

As estatísticas de avaliação de classificadores são calculadas a partir da matriz de confusão.

Muito utilizada em estatística para representar as possíveis soluções de um teste de hipóteses e utiliza a mesma terminologia. Em problemas de classificação, as linhas representam as classes observadas e as colunas representam as classes preditas pelo modelo.

Acurácia do modelo: Taxa de classificação correta do modelo

Erro global do modelo: Complemento da acurácia.

O erro global destaca melhor a diferença de desempenho de dois classificadores. Por exemplo, um modelo com acurácia de 96% tem o dobro do erro de outro modelo com 98% de acurácia.

5 Capítulo 6 - Análise de Agrupamentos

O objetivo da análise de agrupamentos é a divisão do conjunto de dados em subconjuntos chamados de grupos. Frequentemente, a análise de agrupamentos também é chamada de **classificação não supervisionada**.

5.1 Métodos Hierárquicos

Foco em **métodos aglomerativos**: Partem de um nível mais baixo da hierarquia com $K = N$ conjuntos e vai fundindo os conjuntos conforme sobe, utilizando uma **função de ligação** para tal.

Essa **função de ligação** pode ser:

- Vizinho mais próximo: Utiliza os elementos mais próximos dos conjuntos
- Vizinho mais distante
- Distâncias médias: Utiliza a distância média entre os elementos dos grupos
- Centroides: Utiliza a distância entre centroides

Os algoritmos hierárquicos aglomerativos em geral trabalham com a matriz de distâncias na memória.

5.2 Métodos de partição

5.2.1 K-medias

Minimiza o critério do **custo quadrático**.

Resultado muito dependente da inicialização, uma vez que usa estratégia gulosa e pode cair em máximo local.

- Enquanto os centros ainda mudam bastante:
  - Para cada registro r
    - calcular a distância para cada centro i e atribuir registro r ao conjunto C_i com centro mais próximo
  - Recalcular os centros dos conjuntos

O resultado do algoritmo k-médias é muito sensível à presença de outliers e ruído no conjunto de dados. Um outliers será sempre associado a algum grupo e será contabilizado no cálculo da média, o que pode afastar os centros calculados dos centros do conjunto mais denso de registros.

5.3 Validação de grupos

- **Medidas externas**: que avaliam a qualidade do agrupamento em relação a uma informação externa sobre a estrutura de classes do problema:
  - *pureza* de um grupo é a porção dos registros do grupo que pertence à classe que contém;
  - *entropia* pode ser utilizada para avaliar a heterogeneidade do agrupamento, quanto menor a entropia, melhor o agrupamento. a maioria dos registros daquele grupo.
- **Medidas internas**: que avaliam a qualidade do agrupamento a partir dos registros do conjunto de dados, em geral utilizando critérios geométricos, sabendo que um bom agrupamento é o que gera grupos coesos e separados:
  - PI (**índice de partição**) quando calculado com funções de pertinência binárias, é a razão entre o critério quadrático do k-médias e a soma ponderada das distâncias entre centro. Para avaliação de agrupamentos obtidos pelo algoritmo k-médias e suas variantes, o PI é um bom índice para a escolha do melhor agrupamento com o mesmo número de grupos K, considerando um conjunto de soluções obtidas com diversas inicializações aleatórias;
  - VAT (**visualização qualitativa da tendência de agrupamento**): Um bom agrupamento mostra um contraste mais intenso que um agrupamento de pior qualidade. A observação do data image do conjunto de dados com os registros simplesmente ordenados segundo o resultado do agrupamento permite uma avaliação qualitativa do resultado.

- **Medidas relativas**: que produzem um índice calculado pela combinação de critérios (internos) que avaliam a qualidade de um agrupamento em relação a outros.

5.4 Particionamento de grafos

Solução do **problema do corte mínimo**.

Grafo pode ser obtido através da **similaridade entre os registros**