O Capítulo 4, intitulado **"Data Pre-processing: Cleaning Process Data"**, aborda técnicas importantes para lidar com dados de processos industriais que inevitavelmente sofrem de diversas imperfeições. O capítulo enfatiza que a qualidade dos dados de treinamento impacta diretamente a utilidade dos modelos de aprendizado de máquina, seguindo o princípio de "lixo entra, lixo sai".

Os principais tópicos cobertos no Capítulo 4 são:

- **Remoção de ruído de medição (Signal De-noising)**: As medições de processos são frequentemente contaminadas por ruído de alta frequência, o que pode levar a erros na estimativa de parâmetros do modelo ou características desfavoráveis nas variáveis previstas. O capítulo menciona dois métodos para de-noising de sinais:
    - **Filtro de média móvel (Moving window average filter)**: Um filtro simples onde cada ponto de dados é substituído pela média dos seus vizinhos dentro de uma janela especificada.
    - **Filtro Savitzky-Golay (SG filter)**: Um método mais sofisticado que ajusta um polinômio local aos dados dentro de uma janela e usa o valor central do polinômio ajustado como o novo valor do ponto de dados, sendo mais eficaz na preservação de características do sinal em comparação com o filtro de média móvel.
- **Seleção de variáveis/características (Variable Selection / Feature Selection)**: Em processos industriais complexos, pode haver um grande número de variáveis, nem todas relevantes para a modelagem. A inclusão de entradas irrelevantes pode prejudicar o desempenho do modelo. O capítulo descreve três categorias de métodos de seleção de variáveis:
    - **Métodos de filtro (Filter methods)**: Usam medidas estatísticas, como coeficientes de correlação e informação mútua, para quantificar a relevância de cada entrada em relação à variável alvo. As variáveis são então classificadas e as de maior relevância são selecionadas.
    - **Métodos de wrapper (Wrapper methods)**: Utilizam o próprio modelo para avaliar o poder preditivo de diferentes subconjuntos de entrada através de métricas como MSE e BIC. O **Recursive Feature Elimination (RFE)** e a **Sequential Feature Selection (SFS)** são mencionados como exemplos. O RFE constrói um modelo com todas as variáveis e as elimina iterativamente com base na sua importância, enquanto o SFS adiciona (forward) ou remove (backward) variáveis sequencialmente para otimizar o desempenho do modelo.

    - **Métodos embarcados (Embedded methods)**: A seleção de variáveis é uma parte intrínseca do processo de ajuste do modelo. A **regressão Lasso**, árvores de decisão e florestas aleatórias são citados como exemplos, pois fornecem diretamente a importância das características após o ajuste do modelo.

- **Tratamento de outliers (Outlier Handling)**: Dados de processos industriais podem conter outliers devido a falhas de sensores, erros de medição ou condições operacionais anormais. O capítulo divide os métodos de tratamento de outliers em:
    - **Métodos univariados (Univariate methods)**: Limpam cada variável separadamente. A **regra dos 3-sigma** e o **identificador de Hampel** são mencionados. O identificador de Hampel, usando mediana e desvio absoluto mediano (MAD), é apresentado como mais robusto a outliers do que a regra dos 3-sigma.
    - **Métodos multivariados (Multivariate methods)**: Levam em consideração as relações entre múltiplas variáveis. A **Distância de Mahalanobis (MD)** é explicada como uma medida de distância de uma observação do centro da distribuição de dados, considerando a covariância dos dados. O **Estimador MCD (Minimum Covariance Determinant)** é introduzido como uma abordagem robusta para estimar a covariância em dados contaminados por outliers. O **PCA (Análise de Componentes Principais)** também é mencionado como um método multivariado capaz de distinguir diferentes tipos de outliers.
    - **Métodos de data-mining**: Abordam situações onde as suposições de gaussianidade ou unimodalidade dos dados não se sustentam. Uma técnica mencionada envolve ajustar um modelo (como PCA) aos dados e, em seguida, aplicar métodos tradicionais de detecção de outliers aos resíduos, pois os outliers tendem a ter grandes resíduos.
- **Tratamento de dados faltantes (Handling Missing Data)**: Dados faltantes são comuns em processos industriais devido a falhas de sensores ou problemas de coleta de dados. O capítulo não detalha técnicas específicas para tratamento de dados faltantes neste excerto.

O capítulo conclui enfatizando que não existe um método universal para lidar com todos os tipos e graus de contaminação de dados, e uma boa compreensão do processo subjacente e da natureza dos dados é crucial para escolher a técnica apropriada. Um fluxo de trabalho natural de pré-processamento envolve primeiro a remoção de ruído, seguida pela seleção de variáveis e remoção de outliers. O capítulo

marca o fim da primeira fase da jornada de aprendizado de máquina, com as bases estabelecidas para explorar algoritmos de ML clássicos na próxima fase.