Savitzky–Golay filter

Um filtro Savitzky-Golay é um filtro digital que pode ser aplicado a um conjunto de pontos de dados digitais com a finalidade de suavizar os dados, ou seja, aumentar a precisão dos dados sem distorcer a tendência do sinal.

A ideia básica por trás do filtro Savitzky-Golay é ajustar polinômios locais aos dados ao longo de uma janela de dados móvel e, em seguida, usar esses polinômios para estimar o valor suavizado em um determinado ponto. Apresenta as propriedades de baixíssimo custo computacional do algoritmo de filtragem e invariância da fase para não deformar picos. O filtro opera em uma janela de dados de tamanho fixo que se move ao longo do conjunto de dados. A largura desta janela é um parâmetro importante que afeta o grau de suavização. Dentro de cada janela de dados, o filtro ajusta um polinômio de determinado grau aos pontos de dados. Isso é feito usando uma técnica de mínimos quadrados, que ajusta o polinômio de forma a minimizar o erro quadrático entre os dados reais e os valores estimados pelo polinômio. Uma vez que o polinômio é ajustado, ele é usado para estimar o valor suavizado em um ponto central dentro da janela de dados. Este valor suavizado substitui o valor original do ponto de dados. A janela de dados então se move para o próximo conjunto de pontos de dados, e o processo é repetido até que toda a série de dados seja suavizada.

Suponha os dados consistem de um conjunto de pontos , com valores de igualmente espaçados por . O filtro consiste em ajustar uma janela de (5, no exemplo a seguir) pontos por um polinômio de grau (3 i.e.).

- Benefícios do Filtro de Savitzky-Golay:

* Preservação de Características: Mantém picos e vales nos dados, diferentemente de outros métodos de suavização que podem suavizar demais esses detalhes.
* Redução de Ruído: Reduz o ruído nos dados, tornando-os mais adequados para análise subsequente.

- Aplicações Típicas:

* Espectroscopia: Melhorar a qualidade dos espectros antes de análises quantitativas.
* Processamento de Sinais: Suavizar dados de sensores ou sinais biomédicos.
* Análise de Séries Temporais: Suavizar flutuações em dados de séries temporais.

Correção multiplicativa de espalhamento (MSC)

A correção de espalhamento multiplicativo é um método de transformação utilizado para compensar os efeitos aditivos e/ou multiplicativos em dados espectrais. Este método remove a influência de efeitos físicos nos espectros, tais como o tamanho de partícula, a rugosidade e opacidade, os quais não trazem informações químicas sobre as amostras e introduz variações espectrais como o deslocamento da linha de base. Para isto fazer a correção, o método MSC assume que cada espectro é determinado pelas características químicas da amostra somadas às características físicas indesejadas.

Como Funciona a MSC:

- Média do Espectro de Referência: Primeiro, calcula-se o espectro médio de todos os espectros de amostra. Este espectro médio serve como referência.

- Ajuste Linear: Para cada espectro de amostra, ajusta-se uma regressão linear do espectro de referência para o espectro da amostra. Isso gera coeficientes de inclinação (b) e interceptação (a) para cada espectro de amostra.

- Correção dos espectros:

* é o espectro corrigido
* é o espectro original
* é o intercepto da regressão linear
* é o coeficiente de inclinação da regressão linear

- Benefícios da MSC:

* Redução de Variabilidade: Reduz a variabilidade entre os espectros causada pelo espalhamento de luz, que não está relacionado à composição química das amostras.
* Melhora na Análise: Melhora a precisão das análises quantitativas ao minimizar os efeitos de espalhamento de luz e outras interferências.

Padronização normal de sinal (SNV)

A Padronização Normal de Sinal (SNV, do inglês Standard Normal Variate) é uma técnica de pré-processamento de dados usada principalmente em espectroscopia e análise de dados multivariados para corrigir variabilidade de dispersão e efeitos de espessura em amostras. Essa técnica visa melhorar a qualidade dos dados, tornando-os mais comparáveis e, assim, facilitando a análise subsequente.

**- Como funciona:**

1. **Centralização dos Dados:** Para cada ponto de dado ​, subtrai-se a média da série de dados.
2. **Normalização dos Dados:** Em seguida, divide-se a diferença obtida pelo desvio padrão da série de dados.

Matematicamente, a transformação SNV para um ponto de dado é dada por:

* é o valor normalizado
* é o valor original do dado
* é a média dos valores da série de dados
* é o desvio padrão da série

- Benefícios do SNV:

* Correção de Dispersão: Reduz a variabilidade causada por efeitos de dispersão, tornando os dados mais uniformes.
* Correção de Espessura: Mitiga os efeitos causados pela variação na espessura da amostra.
* Melhoria na Comparabilidade: Facilita a comparação entre diferentes amostras, eliminando variações não relacionadas às propriedades químicas ou físicas de interesse.

- Aplicações:

* Espectroscopia: SNV é amplamente usado em espectroscopia para corrigir variações de fundo e dispersão de luz, melhorando a qualidade dos espectros.
* Análise Multivariada: Utilizado em métodos como PLS (Partial Least Squares) e PCA (Principal Component Analysis) para melhorar a precisão das análises.

Divisão de dados x cross-validation x k-fold

1. **Divisão de Dados (Train-Test Split)**

Divisão de dados é uma técnica simples de validação de modelos em que os dados disponíveis são divididos em duas partes:

- Conjunto de Treino (Training Set): Usado para treinar o modelo.

- Conjunto de Teste (Test Set): Usado para testar o modelo e avaliar seu desempenho.

Um exemplo comum é dividir 70% dos dados para treino e 30% para teste. É importante garantir que o modelo não esteja apenas "decorando" os dados, mas que ele seja capaz de generalizar para novos dados.

1. **Validação Cruzada (Cross-Validation)**

Validação cruzada é uma técnica mais robusta para avaliar o desempenho de um modelo, onde os dados são divididos em várias partes (ou "folds") e o modelo é treinado e testado várias vezes. K-Fold Cross-Validation é o tipo mais comum de validação cruzada:

- K-Fold significa que os dados são divididos em K partes iguais (folds).

- O modelo é treinado K vezes, cada vez usando um fold diferente como conjunto de teste e os outros K-1 folds como conjunto de treino.

No final, os resultados dos K testes são combinados para dar uma avaliação mais confiável do desempenho do modelo.

1. **Exemplos Simples**

- Divisão de Dados (Train-Test Split):

Imagine que você tem 100 dados. Com uma divisão 70-30, você usaria 70 dados para treinar o modelo e 30 dados para testá-lo.

- K-Fold Cross-Validation:

Se K=5, você divide seus 100 dados em 5 partes iguais (20 dados cada). Você treina o modelo 5 vezes, cada vez usando um fold diferente como conjunto de teste e os outros 4 folds como conjunto de treino. Isso significa que todos os dados são usados tanto para treino quanto para teste em algum momento, o que dá uma avaliação mais robusta do modelo.

1. **Vantagens e Desvantagens**

- Train-Test Split:

Vantagem: Simples e rápido.

Desvantagem: A avaliação pode ser instável porque depende de uma única divisão dos dados.

- K-Fold Cross-Validation:

Vantagem: Mais robusta e estável, pois usa múltiplas divisões dos dados.

Desvantagem: Mais demorada, pois o modelo é treinado e testado várias vezes.

1. **Resumo**

Divisão de Dados: Divide os dados uma vez em treino e teste.

K-Fold Cross-Validation: Divide os dados várias vezes para treinar e testar o modelo repetidamente.

Cross-Validation (Validação Cruzada): Geralmente se refere a técnicas como K-Fold, que fornecem uma avaliação mais confiável do desempenho do modelo ao usar múltiplas divisões dos dados.

Regressão parcial de mínimos quadrados (PSL)

**É** uma técnica de análise estatística multivariada que se utiliza para encontrar as relações fundamentais entre dois blocos de dados: um conjunto de variáveis independentes (X) e um conjunto de variáveis dependentes (Y). PLS é particularmente útil quando as preditoras são muitas e altamente colineares. A PLS não assume que os preditores são fixos, ao contrário da regressão múltipla. Isto significa que os preditores podem ser medidos com erro, tornando a PLS mais robusta à incerteza da medição.

**- Como funciona:**

PLS busca encontrar componentes latentes que maximizem a covariância entre os dados de entrada (X) e os dados de saída (Y). O processo pode ser resumido em três etapas principais:

1. Decomposição dos dados:

* PLS decompõe os dados em componentes latentes. Para o conjunto de preditores X, gera-se um conjunto de variáveis latentes T e para o conjunto de respostas Y, gera-se um conjunto de variáveis latentes U.
* Essas componentes latentes são combinações lineares dos preditores e das respostas.

1. Modelagem da Relação Entre os Latentes:

* PLS modela a relação entre os componentes latentes T e U para explicar a variabilidade em Y a partir de X.
* A relação é ajustada de forma que os componentes extraídos expliquem a maior parte da covariância entre X e Y.

1. Regressão:

* Finalmente, utiliza-se os componentes latentes para construir um modelo de regressão que prediz Y a partir de X.

**- Vantagens do PLS:**

* **Lida com Colinearidade:** PLS é eficaz quando há colinearidade entre as variáveis preditoras, o que pode ser um problema para outras técnicas de regressão.
* **Redução de Dimensionalidade:** PLS reduz a dimensionalidade do problema, condensando a informação de muitas variáveis em um menor número de componentes latentes.
* **Robustez:** É robusto contra sobreajuste quando utilizado com validação cruzada para determinar o número adequado de componentes.

**- Aplicações:**

* **Quimiometria:** Para análise de dados espectrais, previsão de propriedades químicas e físicas.