

Validação Cruzada e Ajuste Fino dos Parâmetros

Validação Cruzada K-Fold

A validação cruzada k-fold é uma técnica de validação de modelos de aprendizado de máquina que é usada para avaliar a precisão de um modelo. Ela é chamada de "cruzada" porque os dados de treinamento são "divididos" em várias "folds" (ou seções), e cada fold é usado como um conjunto de dados de validação em várias iterações. "k" refere-se ao número de "folds" (ou seções) que os dados são divididos.

A validação cruzada k-fold funciona da seguinte maneira:

1. Os dados de treinamento são divididos em k "folds" (ou seções) de tamanhos iguais.
2. Em cada iteração, um fold é selecionado como o conjunto de dados de validação e os outros k-1 folds são usados como dados de treinamento.
3. O modelo é treinado com os dados de treinamento e é avaliado com os dados de validação.
4. Esse processo é repetido k vezes, com cada fold sendo usado como conjunto de validação uma vez.
5. As métricas de desempenho do modelo são calculadas como a média dos coeficientes de determinação R² obtidos em cada iteração.



A validação cruzada k-fold é uma boa maneira de avaliar a precisão de um modelo pois, dessa maneira, todos os dados são usados tanto para treinamento quanto para validação, o que gera uma melhor avaliação do modelo.

Outra observação muito importante é que a Validação-Cruzada: K-Fold não retorna um modelo (Por exemplo, Regressão Linear) pronto para nós utilizarmos. Ele retorna os scores de cada subdivisão, ou seja, quão performático cada uma é.

Isso é interessante para comparar a performance de vários modelos e ver qual é mais performático.

Validação Cruzada StratifiedKFold

A validação cruzada estratificada k-fold é um método utilizado para avaliar o desempenho de um modelo de aprendizado de máquina, geralmente em conjunto com a busca de hiperparâmetros. Ela é semelhante à validação cruzada k-fold comum, mas garante que cada fold contenha uma proporção similar de exemplos de cada classe.

Passo a passo:

1. Dividir o conjunto de dados em k folds de forma estratificada. Isso significa que o número de exemplos de cada classe será o mesmo em cada fold
2. Para cada fold i, usar os outros k-1 folds como conjunto de treinamento e o fold i como conjunto de validação.
3. Treinar o modelo usando o conjunto de treinamento
4. Avaliar o desempenho do modelo usando o conjunto de validação.
5. Calcular a média e desvio padrão do desempenho do modelo em todos os folds.
6. Escolher o modelo com o melhor desempenho médio.

A diferença entre a validação cruzada estratificada k-fold e a validação cruzada k-fold comum é que a validação cruzada estratificada garante que cada fold contenha uma proporção similar de exemplos de cada classe, enquanto a validação cruzada comum não garante isso. Isso é importante quando a distribuição de classes não é balanceada, ou seja, quando há desproporção entre o número de exemplos de cada classe no conjunto de dados. Neste caso, a validação cruzada estratificada pode dar uma melhor representação do desempenho do modelo no conjunto de dados geral.

Randomized SearchCV e Grid SearchCV

Randomized Search CV e Grid SearchCV são dois métodos utilizados para a busca de hiperparâmetros de um modelo de aprendizado de máquina. Eles permitem encontrar uma combinação de hiperparâmetros que melhor se adequa ao conjunto de dados e ao problema em questão.

Randomized SearchCV:

- é um método de busca de hiperparâmetros que seleciona aleatoriamente uma amostra de combinações de hiperparâmetros para serem testadas.
- Ele é útil quando há muitos hiperparâmetros e o espaço de busca é muito grande.
- Ele é menos custoso computacionalmente do que o Grid SearchCV, pois testa menos combinações de hiperparâmetros.

Grid SearchCV:

- é um método de busca de hiperparâmetros que testa todas as combinações possíveis de hiperparâmetros dentro de um intervalo especificado.

- Ele é útil quando há poucos hiperparâmetros e o espaço de busca é pequeno.
- Ele é mais custoso computacionalmente do que o Randomized SearchCV, pois testa todas as combinações de hiperparâmetros.

Ambos os métodos são utilizados para encontrar a melhor combinação de hiperparâmetros para um determinado modelo de aprendizado de máquina, e são frequentemente usados em conjunto com a validação cruzada para avaliar o desempenho do modelo com diferentes combinações de hiperparâmetros.