UFMG UNIVERSIDADE FEDERAL DE MINAS GERAIS

PROPOSTA DE TRABALHO

ANÁLISE DE EFICIÊNCIA DE MÉTODOS DE OTIMIZAÇÃO DE HIPERPARÂMETROS USANDO ANOVA EM BLOCOS

Equipe F
Bernardo Bacha de Resende
Gustavo Augusto Faria dos Reis
Marília Macêdo de Melo

Planejamento e Análise de Experimentos (EEE933)
2025/2

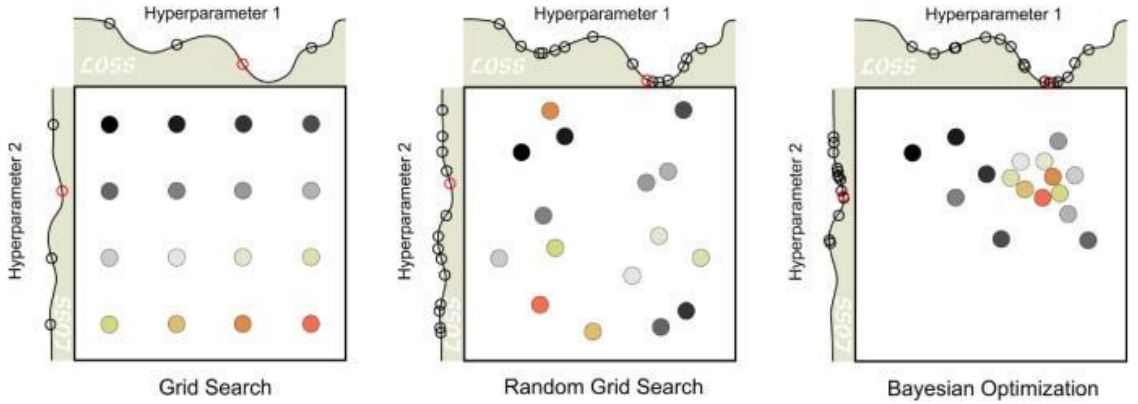
Otimização de Hiperparâmetros (HPs): O Conflito entre Performance e Custo

A otimização de HPs é um passo crítico para extrair a performance máxima de modelos de Machine Learning (ML). Os métodos disponíveis para esta tarefa apresentam um claro *trade-off* entre custo computacional e eficiência de busca.

Contextualização do Desafio

1. ***Grid Search* (GS):**
   - Abordagem de força bruta que testa todas as combinações de uma grade pré-definida. Embora garanta a cobertura, sofre com a "maldição da dimensionalidade", apresentando custo computacional exponencial.
2. ***Random Search* (RS):**
   - Realiza uma amostragem estatística do espaço. Estudos demonstram que é capaz de encontrar resultados competitivos com o GS utilizando apenas uma fração do orçamento computacional.
3. **Otimização Bayesiana (BO):**
   - Método sequencial inteligente que utiliza modelos probabilísticos para decidir o próximo teste. O objetivo é maximizar a eficiência da busca, convergindo para o ótimo com o menor número possível de iterações.

O Problema Central - O problema central é determinar se a eficiência teórica dos métodos modernos (RS e BO) se traduz em vantagem prática robusta sobre o método tradicional (GS), ou se há perda de acurácia ou perda de eficiência computacional.



A Complexidade Estatística da Comparação

Comparar os métodos de *tuning* de forma justa é estatisticamente complexo. Uma análise rasa pode levar a conclusões fundamentalmente erradas.

Justificativa

1. **O Fator Indesejável (*Nuisance Factor*)**
   - A performance (Acurácia e Tempo) é altamente dependente do bloco (*dataset*) escolhido. Datasets fáceis (ex: Iris) terão acurácia alta para todos os métodos, enquanto datasets difíceis (ex: MNIST) terão acurácia baixa para todos. Essa variabilidade entre *datasets* não é o nosso foco de estudo, sendo um Fator Indesejável (*Nuisance Factor*).

2. **A Variabilidade Dominante:**
   - A variabilidade entre os blocos (*datasets*) (ex: a diferença de acurácia entre Iris e MNIST) é provavelmente muito maior do que a variabilidade entre os métodos de *tuning* (ex: a diferença entre RS e BO).

3. **O Risco da Análise Ingênua (ANOVA Simples):**
   - Ao ignorar-se o efeito do dataset, toda a variabilidade entre os métodos será incorporada ao termo de erro residual do modelo. Um erro residual inflado diminui o F-*value* do nosso fator de interesse (Método de Tuning) e reduz drasticamente o poder do teste.

4. **O Efeito Mascarado**
   - Pode levar ao Erro Tipo II: falhar em rejeitar a hipótese nula H_0 e concluir erroneamente que "não há diferença" entre os métodos, quando na verdade a diferença existe, mas foi mascarada pela variabilidade dos *datasets*.

Objetivo do Experimento

O objetivo central deste experimento é quantificar o *trade-off* de eficiência (custo vs. benefício) dos três principais métodos de otimização de hiperparâmetros. Para fazer isso, vamos responder a duas questões de pesquisa separadas:

1. **Análise de Acurácia**
   - **Questão**: Qual é o benefício real de cada método?
   - **Análise**: Testa a diferença estatisticamente significativa na acurácia média que o Grid Search (GS), o Random Search (RS) e a Otimização Bayesiana (BO) conseguem alcançar.
   - **Hipótese Central**: A complexidade do BO encontra uma acurácia realmente superior à do RS, ou eles são, no final, estatisticamente equivalentes?

2. **Análise de Custo (Tempo):**
   - **Questão**: Qual é o custo computacional para alcançar esse benefício?
   - **Análise**: Testa a diferença estatisticamente significativa no tempo de execução médio que GS, RS e BO levam para encontrar seus respectivos ótimos.
   - **Hipótese Central**: O Grid Search é significativamente mais lento que o RS e o BO, justificando o uso de métodos modernos.

Design Experimental: Experimento Completamente Aleatorizado com Blocos (RCBD)

Adotamos este planejamento para isolar o efeito de interesse da variabilidade indesejada. O experimento é definido pelos seguintes componentes:

1. **Fator de Interesse (Tratamento)**
   - O fator principal cuja influência queremos testar é o *Metodo_Tuning*.
   - Níveis *(a=3)*: *Grid Search* (GS), *Random Search* (RS) e Otimização Bayesiana (BO).

2. **Fator de Bloco (Fator Indesejável)**
   - O fator cuja variabilidade queremos controlar e remover estatisticamente do erro.
   - Níveis *(b=5)*: Os 5 datasets selecionados (Iris, *Wine*, *Breast Cancer, Heart Disease*, MNIST).

3. **Controle Experimental (Configuração Fixa)**
   - Para uma comparação justa e isolar o efeito dos métodos de tuning, o algoritmo de ML será mantido constante.
   - Algoritmo Fixo: Support Vector Machine (SVM) com Kernel RBF.
   - Otimização dos parâmetros *c* e *gamma.*

4. **Variáveis de Resposta (Métricas de Eficiência)**
   - Duas métricas serão coletadas em cada um dos $a \times b = 15$ experimentos para permitir a análise de custo-benefício.
   - Resposta 1 (Benefício): Acurácia do modelo.
   - Resposta 2 (Custo): Tempo de Execução do *tuning*.

Seleção dos níveis do fator de bloco

Os cinco blocos (*datasets*) foram selecionados de repositórios clássicos (UCI e Kaggle), que representam os 5 níveis do nosso fator indesejável, conforme o planejamento RCBD.

A escolha foi baseada na diversidade de características, criando blocos que representam uma gama de desafios (de fácil a difícil). Isso é crucial para validar se as conclusões sobre os métodos de tuning são generalizáveis.

1. ***Iris Dataset - Flores***
   - *Benchmark* clássico para classificação multiclasse (3 classes).
   - Representa um bloco de "dificuldade muito baixa" (150 amostras, 4 features).
2. ***Wine Dataset - Vinhos***
   - Problema multiclasse (3 classes) ligeiramente mais complexo que o Iris.
   - Representa um bloco de "dificuldade baixa-média" (178 amostras, 13 features).
3. ***Breast Cancer* (Wisconsin) - Medidas de tumores**
   - Problema de classificação binária (Maligno/Benigno) com maior número de features.
   - Representa um bloco de "dificuldade média" (569 amostras, 30 features).
4. ***Heart Disease* (Cleveland) - Doença cardiaca**
   - Dataset médico clássico para classificação binária (303 amostras, 13 features).
   - Representa um bloco realista de "dificuldade média".
5. **MNIST (*Digit Recognizer*) - Reconhecimento de dígitos**
   - Problema multiclasse (10 classes) de alta dimensionalidade.
   - Representa um bloco de "alta dificuldade" (70,000 amostras, 784 features), que testará o custo e a escalabilidade dos métodos.

Metodologia: Análise Estatística

A análise estatística será realizada em duas fases principais: primeiro, teste ANOVA para detectar diferenças, e segundo, uma análise *post-hoc* para identificar quais grupos são diferentes.

1. **Ajuste do modelo e Validação das premissas**
   Primeiro, ajustaremos o modelo linear RCBD (Resposta ~ *Metodo_Tuning* + *Dataset_Bloco*) para extrair os resíduos. Em seguida, faremos a Validação das Premissas:
   - **Normalidade:** Verificada graficamente (Q-Q Plot) e pelo Teste de Shapiro-Wilk.
   - **Homoscedasticidade:** Verificada pela análise do gráfico de Resíduos vs. Valores Ajustados (buscando variância constante).
2. **Decisão: Escolha do teste estatístico**
   A escolha do teste principal depende do resultado da Validação das Premissas:
   - **Cenário A**: Premissas Válidas: Usaremos a ANOVA Paramétrica. Confiaremos na estatística F (F_0) para determinar se o *Metodo_Tuning* é significativo (p-valor < a).
   - **Cenário B**: Premissas Violadas: A ANOVA paramétrica é inválida. Utilizaremos a alternativa não paramétrica para o design RCBD: o Teste de Friedman.
3. **Análise Post-Hoc**
   - **Cenário A:** Se a ANOVA (Cenário A) for significativa, usaremos o Teste de Tukey HSD (Aula 8) para comparar todos contra todos
   - **Cenário B:** (Pós-Friedman): Se o Teste de Friedman for significativo, usaremos o Teste U de Mann-Whitney para realizar as comparações par a par.

Resultados Esperados

Nossa hipótese central é que a melhor escolha de método não é definida por uma única métrica, mas sim pelo *trade-off* entre custo e benefício. Esperamos que ambas as análises de variância (ANOVA) revelem conclusões opostas:

1. **Análise de Benefício (Acurácia)**
   - **Hipótese:** Esperamos não rejeitar a Hipótese Nula (H_0) para a acurácia.
   - **Interpretação:** O p-valor do teste (seja ANOVA ou Friedman) será provavelmente maior que o nível de significância *a* ($p > 0.05$).
   - **Consequência Prática**: Isso sugere que todos os três métodos (GS, RS, BO) são estatisticamente equivalentes em sua capacidade de encontrar a acurácia máxima do modelo.

2. **Análise de Custo (Tempo de Execução)**
   - **Hipótese:** Esperamos rejeitar fortemente a Hipótese Nula (H_0) para o tempo.
   - **Interpretação:** O p-valor do teste será muito pequeno ($p < 0.001$).
   - **Consequência Prática**: Isso confirma que existe uma diferença estatisticamente significativa e prática no custo computacional de cada método.