UNIVERSIDADE FEDERAL DO PARÁ - UFPA
CAMPUS UNIVERSITÁRIO DE TUCURUÍ - CAMTUC
FACULDADE DE ENGENHARIA ELÉTRICA – FEE

Edimar Fernandes Dias - 201933940004

Aprendizagem em conjunto

Relatório acadêmico apresentado como requisito para obtenção da nota final, na disciplina de aprendizado de máquina, no curso de graduação em Engenharia Elétrica, na Universidade Federal do Pará – Campus Tucuruí.

Prof. cleison Silva

TUCURUÍ/PA
2023

SUMÁRIO

1.1 Objetivos Específicos

Este relatório tem como objetivo fazer um resumo da apresentação dos alunos da pós graduação, do tópico 5-Aprendizagem em conjunto.

2. APRENDIZAGEM EM CONJUNTO

Na apresentação, foi enfatizado pelos alunos de pós-graduação a importância dos pontos discutidos sobre aprendizado em conjunto. Eles destacaram que, ao lidar com modelos que cometem erros distintos, a estratégia de tirar a média das previsões pode ser eficaz. Isso foi detalhado com a implementação de um classificador de votação, considerando votos rígidos e suaves.

A diversidade dos modelos foi ressaltada como fundamental para o sucesso do aprendizado em conjunto, permitindo que diferentes modelos se especializem em partes distintas dos dados. A capacidade de compensar erros individuais, especialmente quando os modelos tendem a sobreajustar, foi apontada como uma vantagem notável dessa abordagem.

No contexto da análise de viés e variância, os alunos destacaram as estratégias de combinar modelos de acordo com suas características. Para modelos com sub ajuste, a combinação com outros de baixa variância, usando técnicas como Boosting, foi sugerida. Por outro lado, para modelos propensos a sobre ajustar, a combinação com modelos de baixo viés, utilizando Bagging, foi considerada mais eficaz.

A necessidade de modelos não correlacionados e a técnica avançada de Stacking para aprender a combinar previsões foram mencionadas como pontos cruciais para o sucesso do aprendizado em conjunto.

2.1. Árvores de decisão (recapitulação)

Durante a apresentação, os alunos de pós-graduação sublinharam pontos cruciais na recapitulação sobre árvores de decisão. Foi enfatizado que essas árvores representam visualmente decisões através de divisões em folhas baseadas em testes. A avaliação da pureza das folhas foi destacada, utilizando heurísticas como o índice de Gini ou entropia.

Os alunos reconheceram a importância do processo de otimização, que envolve pesquisa recursiva e uma abordagem gananciosa através do algoritmo de Hunt. Além disso, destacaram a consideração de todas as possíveis divisões entre pontos de dados adjacentes para cada recurso como uma parte fundamental da construção da árvore de decisão.

2.2. Avaliação (função de perda para classificação)

Na apresentação, foram destacados conceitos-chave relacionados à avaliação em problemas de classificação. Cada folha em uma árvore de decisão prevê a probabilidade de classe, representada por $\hat{p}_c$. Duas medidas de impureza de folha foram mencionadas: o **Índice de Gini**, proposto por Gini, e a **Entropia**, uma medida mais custosa introduzida por outro autor. A fórmula para a melhor divisão, maximizando o ganho de informação, também foi enfatizada como crucial no processo de construção da árvore de decisão.

2.3. Árvores de regressão

Na apresentação, foram destacados conceitos fundamentais relacionados às árvores de regressão. Cada folha em uma árvore de regressão prevê o valor médio alvo, representado como mu, para todos os pontos contidos naquela folha. A escolha da divisão é baseada na minimização do erro quadrático das folhas, calculado como $\sum xi \epsilon L(yi - \mu)^2$. Esse enfoque visa proporcionar previsões não suavizadas em etapas, resultando em um modelo que não é capaz de extrapolar além dos dados utilizados na construção da árvore. Essas características tornam as árvores de regressão especialmente úteis para problemas nos quais a natureza dos dados se encaixa com previsões por etapas e não é necessário realizar projeções além do escopo dos dados originais.

2.4. Importância do recurso baseado em impureza/entropia

Foi salientada a avaliação da importância dos recursos para o modelo, considerando as decisões de divisão. A relevância é influenciada pelos recursos escolhidos para divisões e pela posição na árvore em que essas divisões ocorrem, com destaque para a maior importância atribuída às primeiras divisões.

2.5. Sub e overfitting

Com a abordagem dos alunos na apresentação, foi possível compreender a gestão de sobreajuste (underfitting) e sobreajuste (overfitting) em árvores de decisão. Destacou-se a facilidade de controlar a profundidade máxima das árvores como um hiperparâmetro chave.

A análise de viés-variância foi discutida, indicando que árvores rasas têm viés alto, mas variância muito baixa, resultando em ajuste insuficiente. Por outro lado, árvores profundas apresentam alta variância, mas baixo viés, indicando overfitting. A flexibilidade no controle da complexidade das árvores as torna ideais para ajuste.

Árvores profundas, quando combinadas com Bagging, ajudam a manter o viés baixo e reduzir a variação. Por sua vez, árvores rasas, ao serem combinadas com Boosting, auxiliam em manter a variação baixa e reduzir o viés.

2. 6. Bagging (agregação Bootstrap)

Bagging como uma técnica de redução de sobreajuste, obtendo modelos diversos ao treinar o mesmo modelo em diferentes amostras de treinamento (bootstrap). A estratégia visa reduzir a variância, calculando a média das previsões individuais.

Na prática, várias amostras de bootstrap são utilizadas para treinar modelos distintos. O aumento no número de modelos implica em maior suavização, mas também resulta em treinamento e previsões mais lentas. Os modelos básicos devem ser instáveis, como árvores de decisão profundas ou até mesmo árvores de decisão aleatórias.

O Bagging é benéfico não apenas para árvores de decisão, mas também para redes neurais profundas, resultando em conjuntos profundos. A previsão é realizada calculando a média das previsões dos modelos básicos. Para classificação, é adotada uma votação suave, possivelmente ponderada, enquanto para regressão, é utilizado o valor médio.

Uma característica adicional do Bagging é a capacidade de produzir estimativas de incerteza, combinando probabilidades de classe de modelos individuais, ou variações para problemas de regressão.

2. 7. Florestas Aleatórias

Na apresentação, os alunos ressaltaram as Florestas Aleatórias, uma abordagem que utiliza árvores aleatórias para tornar os modelos ainda menos correlacionados, resultando em maior instabilidade. Em cada divisão, apenas um conjunto aleatório de `max features` é considerado.

Além disso, mencionaram a variante das "Árvores Extremamente Aleatórias", que considera um único limite aleatório para um conjunto também aleatório de recursos, proporcionando maior rapidez no processo de construção.

2. 8. Efeito no viés e na variância

Foi destacado que aumentar o número de modelos (árvores) em Florestas Aleatórias resulta em uma diminuição da variância, reduzindo assim o overfitting. O viés praticamente não é afetado, embora possa aumentar se a floresta se tornar excessivamente grande, o que pode resultar em suavização excessiva.

8. CONCLUSÃO

Na apresentação, os alunos destacaram conceitos fundamentais sobre árvores de decisão e aprendizado em conjunto. Inicialmente, abordaram estratégias de ensemble, como o uso de Voting Classifier, Bagging e Boosting, salientando a importância da diversidade entre modelos para superar sub ajuste e sobreajuste.

A análise de árvores de decisão incluiu a representação visual de decisões, avaliação de impureza por meio do índice de Gini e Entropia, além da otimização usando algoritmo de Hunt. A flexibilidade dessas árvores foi destacada, sendo eficazes para problemas específicos, mas suscetíveis a sub- ajuste e sobreajuste.

Abordou-se a importância dos recursos na construção de árvores e a mensuração de sua relevância. O tópico de Bagging enfatizou a redução de sobreajuste através de modelos diversos, enquanto Random Forests aumentam a instabilidade para melhor desempenho.

Na gestão de desajuste e sobreajuste, os alunos ressaltaram que o controle da complexidade das árvores é crucial. Árvores rasas apresentam alto viés e baixa variância, enquanto árvores profundas têm baixo viés e alta variância. Bagging e Boosting foram indicados como estratégias eficazes para árvores profundas e rasas, respectivamente.

Por fim, Florestas Aleatórias foram discutidas como uma extensão que utilizava árvores aleatórias para aumentar a instabilidade. Aumentar o número de modelos reduz a variância, mas deve ser gerenciado para evitar suavização excessiva.

Em conclusão, os alunos destacaram a versatilidade e eficácia das árvores de decisão e técnicas de ensemble, ressaltando a necessidade de uma abordagem equilibrada para gerenciar sub ajuste e sobreajuste, adaptando-se aos requisitos específicos do problema em questão.