



Machine Learning

Árvore de decisão: overfitting

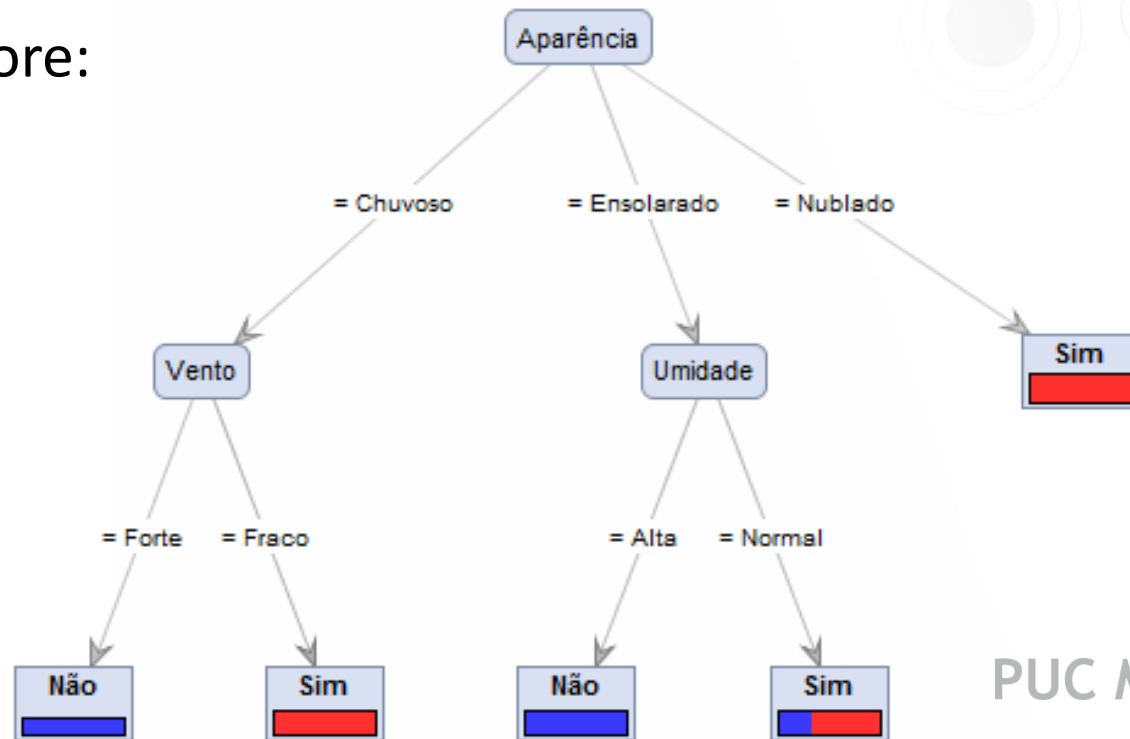
Prof. Hugo de Paula

Overfitting (superajustamento) em árvores de decisão

Considere o seguinte ruído na base de treinamento:

<Ensolarado, Quente, Normal, Forte, Não>

Nova árvore:



Overfitting

Considere uma hipótese h e:

- Taxa de erro sobre o conjunto de treinamento: $err_{train}(h)$
- Erro real sobre todo conjunto de dados: $err_{real}(h)$

Overfitting

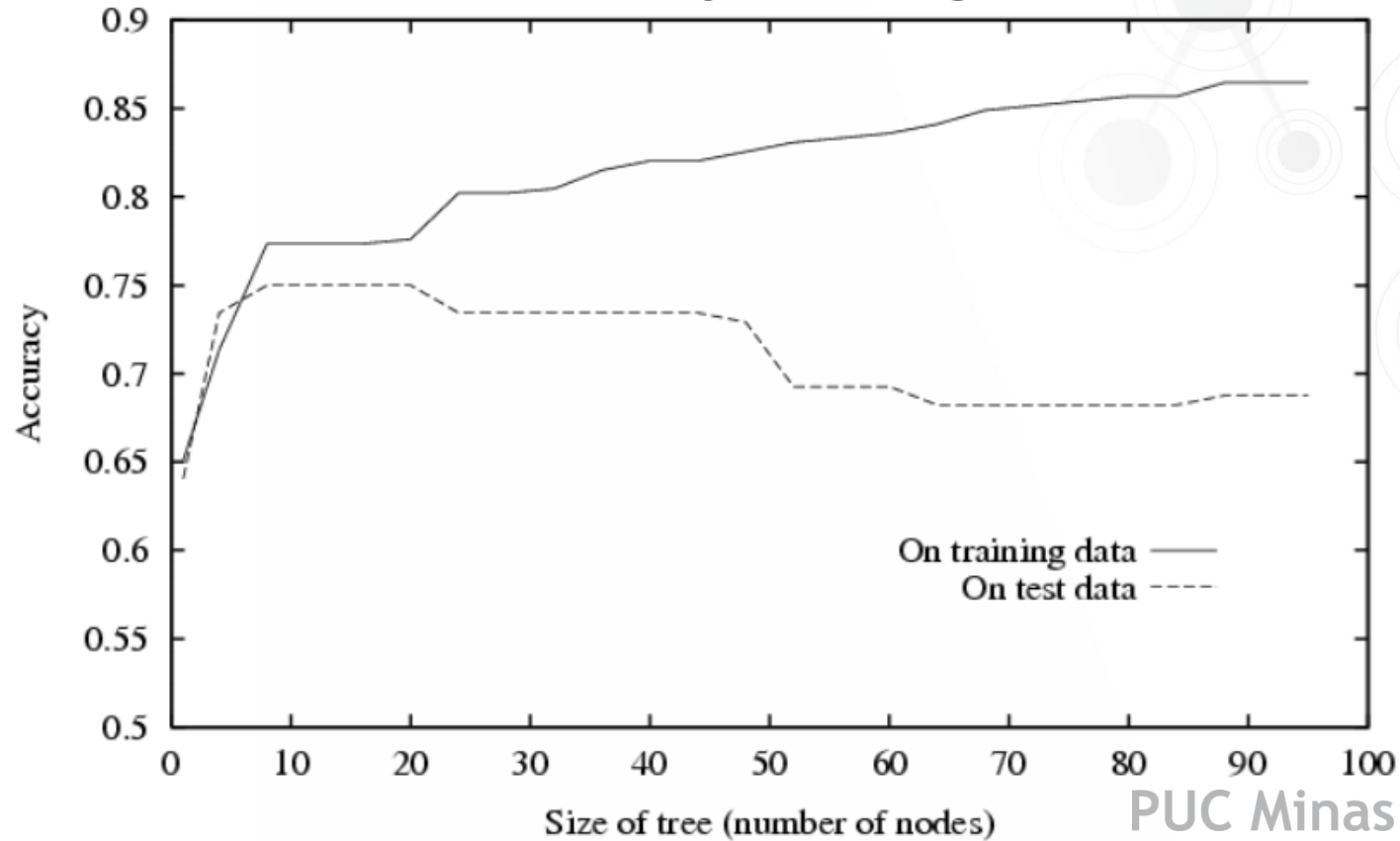
Diz-se que h sofre *overfitting* referente aos dados de treinamento se:

$$\mathbf{err}_{real}(h) > \mathbf{err}_{train}(h)$$

Quantidade de overfitting

$$\mathbf{err}_{real}(h) - \mathbf{err}_{train}(h)$$

Overfitting



Evitando *overfitting*

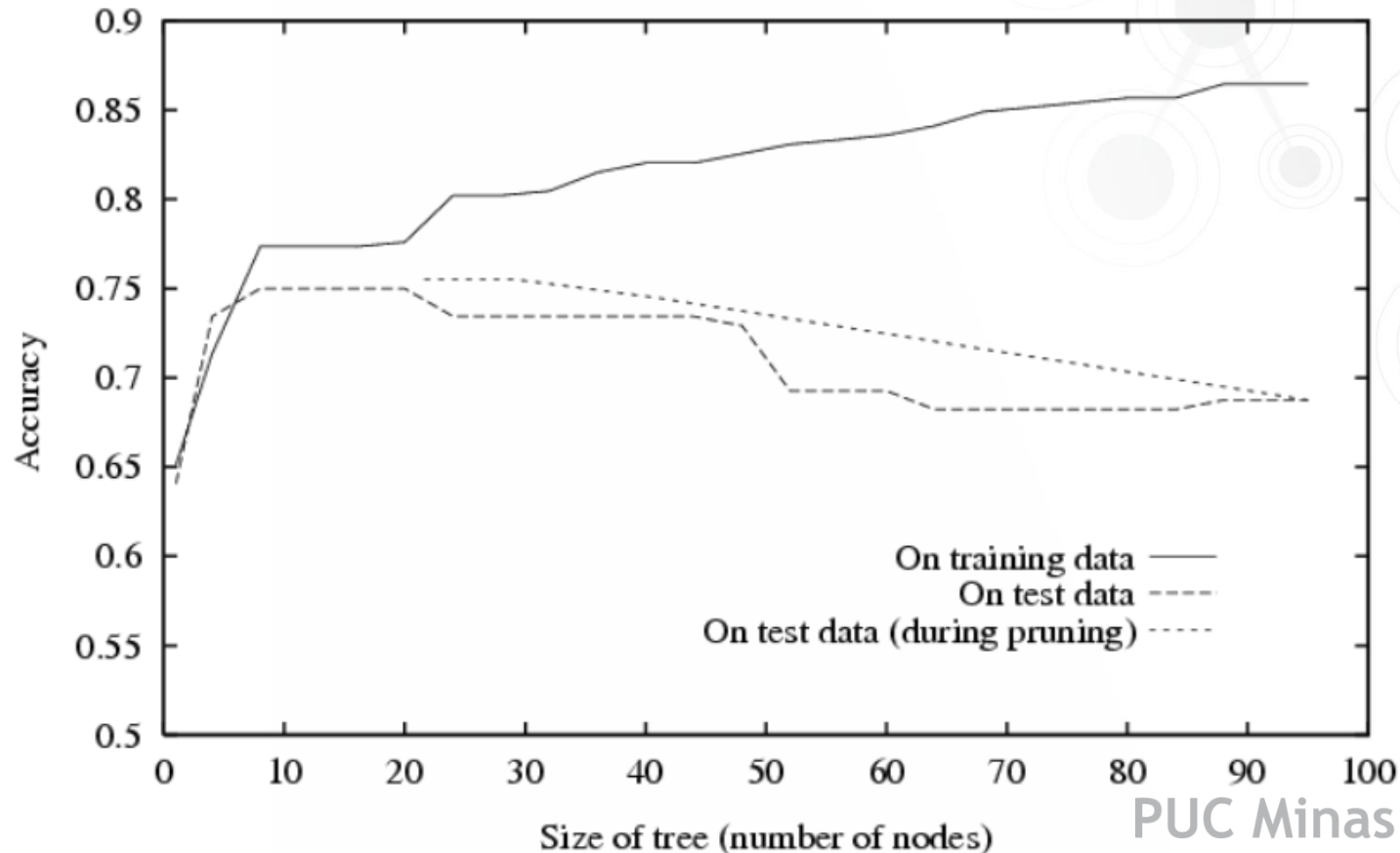
Parar de crescer a árvore quando não for estatisticamente relevante.

Gerar a árvore completa e depois podá-la.

Poda com erro reduzido

- Dividir dado em *treinamento* e *validação*
- Criar árvore que classifica *treinamento* corretamente
- Repetir até que seja prejudicial ao modelo
 - Avaliar o impacto da poda de cada nó (e seus descendentes) da árvore na *validação*.
 - Remover nó que mais aumenta a acurácia na *validação* (algoritmo guloso).

Efeito da poda com erro reduzido no *Overfitting*



Aviso legal

O material presente nesta apresentação foi produzido a partir de informações próprias e coletadas de documentos obtidos publicamente a partir da Internet. Este material contém ilustrações adquiridas de bancos de imagens de origem privada ou pública, não possuindo a intenção de violar qualquer direito pertencente à terceiros e sendo voltado para fins acadêmicos ou meramente ilustrativos. Portanto, os textos, fotografias, imagens, logomarcas e sons presentes nesta apresentação se encontram protegidos por direitos autorais ou outros direitos de propriedade intelectual.

Ao usar este material, o usuário deverá respeitar todos os direitos de propriedade intelectual e industrial, os decorrentes da proteção de marcas registradas da mesma, bem como todos os direitos referentes a terceiros que por ventura estejam, ou estiveram, de alguma forma disponíveis nos slides. O simples acesso a este conteúdo não confere ao usuário qualquer direito de uso dos nomes, títulos, palavras, frases, marcas, dentre outras, que nele estejam, ou estiveram, disponíveis.

É vedada sua utilização para finalidades comerciais, publicitárias ou qualquer outra que contrarie a realidade para a qual foi concebido. Sendo que é proibida sua reprodução, distribuição, transmissão, exibição, publicação ou divulgação, total ou parcial, dos textos, figuras, gráficos e demais conteúdos descritos anteriormente, que compõem o presente material, sem prévia e expressa autorização de seu titular, sendo permitida somente a impressão de cópias para uso acadêmico e arquivo pessoal, sem que sejam separadas as partes, permitindo dar o fiel e real entendimento de seu conteúdo e objetivo. Em hipótese alguma o usuário adquirirá quaisquer direitos sobre os mesmos.

O usuário assume toda e qualquer responsabilidade, de caráter civil e/ou criminal, pela utilização indevida das informações, textos, gráficos, marcas, enfim, todo e qualquer direito de propriedade intelectual ou industrial deste material.



PUC Minas
Virtual

© PUC Minas • Todos os direitos reservados, de acordo com o art. 184 do Código Penal e com a lei 9.610 de 19 de fevereiro de 1998.
Proibidas a reprodução, a distribuição, a difusão, a execução pública, a locação e quaisquer outras modalidades de utilização sem a devida autorização da Pontifícia Universidade Católica de Minas Gerais.