Modelagem Preditiva

Atividade Prática Supervisionada

Professor: Paulo C. Marques F.

Segundo semestre de 2022

Informações gerais

A atividade deverá ser feita em grupos de até 4 integrantes, podendo ser feita individualmente se desejado. Os grupos farão uma apresentação do relatório final da atividade na aula do dia 23 de Novembro. Todos os integrantes do grupo devem ser capazes de discutir o conteúdo do relatório. Cada grupo deve desenvolver sua atividade sem compartilhar informações com os demais grupos. O relatório da atividade deve apresentar todo o código R utilizado nas análises.

Utilize os materiais de aula, códigos, slides e o PDF do livro-texto, conforme forem necessários. Os dois vídeos abaixo cobrem os materiais de *Bagging* e *Random Forests* discutidos em um oferecimento prévio da disciplina no Insper.

- https://www.youtube.com/watch?v=ZEaTyD8lCXI
- https://www.youtube.com/watch?v=5C-tgr-vGp4

Parte teórica

- Forneça uma descrição teórica bastante completa das *Random Forests* no contexto de regressão.
- Comece pela descrição do funcionamento das árvores de regressão individuais.
- Discuta os mecanismos de *Bagging* e de seleção aleatória das preditoras nos *splits*.
- Apresente exemplos e figuras.
- **Muito importante: não utilize nenhuma figura pronta da Internet. Crie suas próprias figuras.**
- Deixe clara a intuição envolvida no ganho de performance preditiva das *Random Forests*.
- Pesquise e explique em detalhe o que é a estimativa "out-of-bag" do erro de generalização produzida por uma *Random Forest*.
- A nota da parte teórica será proporcional à qualidade e à completude da exposição.

Aplicação 1: Um problema de churn

O primeiro conjunto de dados `churn.csv` contém informações sobre os clientes de uma instituição bancária. Os nomes das colunas são auto explicativos. O objetivo é prever a variável `Exited`, que determina se o cliente cancelará o serviço (*churned*) ou não.

Divida os dados em conjuntos de treinamento e de teste de mesmo tamanho e construa modelos preditivos de classificação utilizando Regressão Logística, Árvore de Classificação, *Random Forest* e *Boosting*. Compare todos os métodos em relação à sua acurácia no conjunto de teste. Construa as curva ROC de todos os métodos e compare as áreas embaixo das curvas. Discuta os resultados.

Aplicação 2: Preço de automóveis usados

O segundo conjunto de dados `used_cars.csv` contém preços de veículos usados da marca Mercedes. Os nomes das colunas são auto explicativos (`trim` é o modelo do veículo). O objetivo é prever a variável `price`.

Divida os dados em conjuntos de treinamento e de teste de mesmo tamanho e construa modelos preditivos utilizando os métodos de Regressão Linear Múltipla, Árvore de Regressão, *Random Forest* e *Boosting*. Compare os métodos em relação à raiz quadrada do erro quadrático médio de teste. Para cada método, construa um gráfico cruzando os valores previstos e observados no conjunto de teste. Discuta os resultados. Obs: para a Regressão Linear, tabalhe com o logaritmo do preço.