Modelagem Preditiva

Atividade Prática Supervisionada

Professor: Paulo C. Marques F.

Segundo semestre de 2025

Formação dos grupos

Esta atividade prática supervisionada deverá ser feita em grupos de até 4 integrantes, sem exceções para a formação de grupos maiores. Se desejado, a atividade poderá ser feita individualmente. Um dos integrantes do grupo deverá enviar um e-mail para o professor paulocmf1@insper.edu.br até o dia 19 de outubro informando os nomes de todos os integrantes do grupo. Não confie em ChatGPT e afins, pois você terá que explicar detalhadamente todos os conceitos do relatório no dia da apresentação. Escolha bem os integrantes do seu grupo, pelos motivos detalhados abaixo na seção "Apresentação". Havendo problemas de instalação de quaisquer bibliotecas, entre em contato com o professor assistente em guilhermeeev1@insper.edu.br e utilize o horário de monitoria para sanar eventuais dificuldades. Não deixe para resolver problemas em cima da hora. Antecipe o trabalho.

Apresentação

Os grupos apresentarão o conteúdo da APS na aula do dia 05 de novembro. Todos os integrantes do grupo deverão ser capazes de discutir o conteúdo completo do relatório, sob o risco de penalização da nota de todo o grupo. Para cada grupo que fará a apresentação, o professor sorteará um item diferente da APS para cada integrante do grupo, que deverá ser explicado em detalhe na lousa. Portanto, é fundamental que todos os integrantes do grupo estudem e entendam **todo** o conteúdo da APS. Não "divida" o conteúdo e o entendimento dos itens da APS.

Formato do relatório e entrega

Observe os seguintes pontos na composição e entrega do relatório.

a. O relatório deverá ser escrito no Word com a fonte Times New Roman, no tamanho 12pt.
b. A formatação da página deve ser padrão.
c. As linhas devem ter espacejamento simples.
d. Todas as figuras (com exceção daquelas figuras geradas ao rodar os modelos no R) devem ser feitas à mão, fotografadas com o celular e coladas no relatório. A qualidade e criatividade das figuras faz parte da nota do relatório. Capriche.
e. Todo o código deve ser escrito em R. Não utilize Python nesta APS.
f. O código R da aplicação não deve fazer parte do corpo do relatório.
g. Todas as figuras e demais saídas / valores gerados pelo código R devem constar do relatório, com as devidas discussões.
h. Apenas um dos integrantes do grupo deverá fazer a entrega via link do Blackboard no dia 5 de novembro até 23h59. Crie um arquivo .zip com os dois arquivos: o Word do relatório e o arquivo com o código R da aplicação solicitada.

Parte teórica (7,0 pontos)

O relatório tratará do entendimento teórico dos modelos estudados no curso e suas aplicações a dados reais. Seja criativo na elaboração de seus exemplos e figuras. Não crie exemplos complicados demais, ou simples demais. Garanta um entendimento impecável dos conceitos envolvidos. No final do processo de escrita, revise seu texto quanto à corretude gramatical. Atenção: o ChatGPT costuma utilizar o termo "árvore de decisão", que nunca utilizamos neste curso. Tome cuidado. Entenda **todo** o conteúdo do relatório para se preparar para a apresentação em classe.

1. (2,0 pontos) Treine um modelo de regressão logística e um modelo de árvore de classificação utilizando os dados *Q1_training.csv*. Compare a AUC dos dois modelos no conjunto de teste

Q1_test.csv. Investigue e explique detalhadamente o motivo de você ter obtido estes resultados. É possível modificar a regressão logística para melhorar os resultados? Explique detalhadamente seu argumento e exiba o código correspondente.

2. (2,0 pontos) Explique detalhadamente o método de bagging de árvores de regressão (não mencione árvores de classificação neste item). Discuta cuidadosamente o que é o bootstrap e como este é utilizado no bagging. Construa um ou mais exemplos com figuras para apoiar suas explicações. Utilizando apenas a library(tree), programe um bagging para o conjunto de dados "California housing", utilizado em aula. Quebre os dados 50/50 para treinamento e teste. Verifique se o método de subespaço aleatório gera um ganho de performance para este conjunto de dados.

3. (1,5 ponto) Discuta o mecanismo de aleatorização dos splits introduzido por Breiman que levou à definição de uma Random Forest. Utilize exemplos. Deixe clara a intuição envolvida no ganho de performance das Random Forests no que tange ao trade-off entre viés e variância. É correto afirmar que Breiman apenas utilizou o método do subespaço aleatório inventado por Ho em 1998?

4. (1,5 ponto) Explique detalhadamente o cálculo do erro out-of-bag nas Random Forests em problemas de regressão. Invente figura para explicar o conceito. Faça um estudo com o conjunto de dados "California housing" (utilizado em aula) mostrando que o erro out-of-bag se aproxima do erro de teste conforme aumentamos o número de árvores da floresta. Quebre os dados 50/50 para treinamento e teste. Utilize a library(ranger).

Aplicação 1: Um problema de churn (1,5 ponto)

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 cancelou o serviço (churned) ou não. Divida os dados (50/50) em conjuntos de treinamento e de teste e construa modelos preditivos de classificação utilizando k-NN, Regressão Logística, Árvore de Classificação, Random Forest e CatBoost. Escolha o parâmetro k do método k-NN utilizando uma "regra de bolso" (pesquise!). Não é necessário otimizar os parâmetros do CatBoost nesta aplicação. Inicialmente, compare todos os métodos em relação à sua acurácia no conjunto de teste (ou seja, utilizando "probabilidade de corte" igual a 50%). Construa as curvas ROC de todos os métodos e compare as áreas embaixo das curvas. Discuta os resultados.

Aplicação 2: Preço de automóveis usados (1,5 ponto)

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 (50/50) em conjuntos de treinamento e de teste e construa modelos preditivos utilizando os métodos de Regressão Linear Múltipla, Árvore de Regressão, Random Forest e CatBoost. Não é necessário otimizar os parâmetros do CatBoost nesta aplicação. Compare os métodos em relação à raiz quadrada do erro quadrático médio de teste. Faça gráficos com os valores previstos e observados dos preços. Discuta os resultados.