INF-0615 – Aprendizado de Máquina Supervisionado I

Exercícios 8 e 9

Análise de Campanha de Marketing de um Banco

1 Descrição do Dataset

Neste exercício, vocês irão predizer se uma pessoa irá aceitar ou não um produto oferecido pela equipe de marketing do banco do qual é cliente. Para isso, uma base de dados com atributos numéricos discretos e atributos categóricos é disponibilizada. Os atributos se referem às informações pessoais dos clientes e histórico de adesão a produtos oferecidos em outras campanhas.

- **age**: Idade do cliente.
- **job**: Profissão exercida pelo cliente. Possíveis valores: *admin*, *blue-collar*, *entrepreneur*, *housemaid*, *management*, *retired*, *self-employed*, *services*, *student*, *technician*, *unemployed* ou *unknown*.
- **marital**: Estado civil do cliente. Possíveis valores: *single*, *divorced* ou *married*.
- **education**: Nível de escolaridade do cliente. Possíveis valores são: *primary*, *secondary*, *tertiary* ou *unknown*.
- **default**: Informa se o cliente é inadimplente. Possíveis valores são: *yes* ou *no*.
- **balance**: Balanço médio anual de movimentação da conta medido em euros.
- **housing**: Informa se o cliente realizou empréstimo para financiamento. Possíveis valores: *yes* ou *no*.
- **loan**: Informa se o cliente fez empréstimo pessoal. Possíveis valores: *yes* ou *no*.
- **contact**: Forma de contato realizada com o cliente: Possíveis valores são: *cellular*, *telephone* ou *unknown*.
- **day**: Último dia do mês que o banco fez contato com o cliente. Atributo númerico de 1 a 31.
- **month**: Úlitmo mês do ano que o banco fez contato com o cliente. Apresenta valores nominais referentes aos meses.
- **campaign**: Número de contatos feito com o cliente sobre esta campanha de marketing.
- **pdays**: Números de dias que se passaram desde que o cliente foi contatado por outra campanha de marketing promovida pelo banco. Valor numérico. Apresenta -1 indicando que o cliente não foi previamente contatado.
- **previous**: Número total de vezes que o cliente foi contatado previamente em outras campanhas.
- **poutcome**: Resultado se o cliente aderiu ou não a última campanha de marketing. Possíveis valores: *unknown*, *other*, *failure* ou *success*.
- **y**: Atributo alvo que indica se o cliente aderiu ao produto oferecido na campanha de marketing atual. Possíveis valores: *yes* ou *no*. **Este é o atributo alvo que iremos predizer**.

2 Tarefas

1. Inspecionem os dados. Quantos exemplos nós temos? Há features sem anoatções?
2. Treine uma árvore de decisão com todas as features.
3. Realize a poda da árvore tomando o CP (*Complexity Parameter*) baseado no menor erro do *cross-validation* realizado no treinamento.
4. Plote a acurácia no conjunto de treinamento e de validação pelo tamanho da árvore de decisão. Houve overfitting?
5. Treine uma Floresta Aleatória para predizer se o cliente irá aderir ou não ao produto oferecido.
6. Plote a acurácia no conjunto de treinamento e de validação pelo número de árvores na floresta.
7. Treine um ensemble de árvores de decisão utilizando o protocolo *Bagging* e *Pasting*.
8. Treine um ensemble de árvores de decisão utilizando o protocolo *Boosting*.
9. Compare a performance dos ensembles. Qual delas lidam melhor com o desbalanceamento? Por quê?

3 Arquivos

O arquivo disponível no Moodle é:

- *bank_full.csv*: conjunto de dados que será dividido em treinamento, validação e teste.
- *Ex08.R*: código que implementa a solução do exercício com Árvores de Decisão e Florestas Aleatórias.
- *Ex09.R*: código que implementa a solução do exercício utilizando as três técnicas de Ensemble (*Bagging, Pasting, Boosting*. Bem como uma versão modificada da Floresta Aleatória com árvores balanceadas.

4 Referências

1. Bank Market Data Set. UCI Machine Learning Repository. *https://archive.ics.uci.edu/ml/datasets/Bank+Marketing*.
2. (Moro et al., 2014) S. Moro, P. Cortez and P. Rita. *A Data-Driven Approach to Predict the Success of Bank Telemarketing*. Decision Support Systems, Elsevier, 62:22-31, June 2014.