Aprendizado Ensemble - Florestas Aleatórias
Trabalho 1 – INF01017 – 2019/2

Prof. Bruno Castro da Silva
bsilva@inf.ufrgs.br

Profa. Mariana R. Mendoza
mrmendoza@inf.ufrgs.br

1 Objetivo

O Trabalho 1 da disciplina consiste na implementação do algoritmo de Florestas Aleatórias (*Random Forests*) para tarefas de classificação seguindo o paradigma de aprendizado *ensemble*, i.e., o uso de múltiplos modelos. O trabalho será desenvolvido em **grupos de 2 ou 3 alunos (preferível)**[1] e inclui a implementação de todas as características principais do algoritmo de Florestas Aleatórias discutidas em aula. Em particular, os grupos deverão implementar:

- O algoritmo de indução de uma árvore de decisão, usando como critério de seleção de atributos para divisão de nós o Ganho de Informação (baseado no conceito de entropia), como visto na disciplina, tratando tanto atributos categóricos quanto numéricos;
- Uma função para percorrer a árvore de decisão treinada e realizar a classificação de uma nova instância (do conjunto de teste);
- O mecanismo de bootstrap (amostragem com reposição) para geração de subconjuntos a partir do conjunto de dados de treinamento originais. Cada bootstrap será utilizado para o treinamento de uma árvore no aprendizado *ensemble*;
- O mecanismo de amostragem de m atributos a cada divisão de nó, a partir dos quais será selecionado o melhor atributo de acordo com o critério de Ganho de Informação;
- O treinamento de um *ensemble* de árvores de decisão, adotando os mecanismos de bootstrap e seleção de atributos com amostragem, como mencionados acima;
- O mecanismo de votação majoritária entre as múltiplas árvores de decisão no *ensemble*, para classificação de novas instâncias utilizando o modelo de Florestas Aleatórias;
- A técnica de validação cruzada (*cross-validation*) **estratificada**, para avaliar poder de generalização do modelo e a variação de desempenho de acordo com diferentes valores para os parâmetros do algoritmo (ex., número de árvores no *ensemble*).
- Avaliação do impacto do número de árvores no no desempenho do *ensemble*.

A divisão de nós no algoritmo básico de indução de árvores de decisão deve ser tratada tanto para atributos categóricos quanto atributos numéricos. Para atributos numéricos, sugere-se utilizar como valor de divisão do nó a média aritmética entre os valores do atributo na partição analisada, a fim de simplificar os cálculos. No entanto, os grupos podem optar por adotar outra estratégia para definição deste valor, conforme visto em aula (ver Aula 03) ou baseado em revisão da literatura. Os

[1] Recomenda-se manter a mesma formação de grupos para todos os trabalhos práticos da disciplina

grupos podem assumir que o tipo de cada atributo será informado manualmente junto com o *dataset* para classificação.

Cada grupo, ao aplicar sua implementação de Florestas Aleatórias nos problemas de classificação, deverá utilizar a metodologia de validação cruzada estratificada (*cross-validation*, a fim de avaliar o desempenho do modelo e o efeito de diferentes valores de parâmetros no aprendizado do algoritmo. Especificamente, o parâmetro a ser otimizado neste trabalho é **número de árvores no *ensemble* (*ntree*)**, tal que a implementação deve ser executada para diferentes número de árvores no *ensemble* (ex: 10, 25, 50,...) - onde a escolha do limite superior dependerá de custos computacionais envolvidos na execução do algoritmo - e os respectivos desempenhos deverão ser comparados. O **número de atributos amostrados (*m*)**, pode ser otimizado ou pode ser adotado o valor padrão sugerido na literatura para tarefas de classificação, i.e., raiz quadrada do número total de atributos no *dataset*, a critério do grupo.

Todos os trabalhos submetidos serão examinados de forma manual pelos professores, e também comparados com implementações conhecidas de algoritmos de indução de árvore e com as implementações dos demais alunos (inclusive de semestres anteriores), a fim de detectar plágios. É proibido o uso, total ou parcial, de bibliotecas ou implementações de aprendizado de máquina que resolvam partes deste projeto. Se houver quaisquer dúvidas sobre se algum recurso de software pode ou não ser utilizado, por favor consultem um dos professores antes de fazê-lo.

2 Entrega de Resultados: até 10/10/2019

- Os grupos deverão enviar seu código fonte e relatório pelo Moodle do INF até a data de entrega do trabalho (ver na seção "Critérios de avaliação" a política adotada para entregas com atraso).
- O relatório deverá estar em formato **pdf**, e deverá conter uma descrição das características gerais da implementação de cada grupo (p.ex., descrição das estruturas de dados utilizadas para armazenar as árvores, como é feita a classificação de novas instâncias, detalhes sobre possíveis otimizações feitas para tornar o algoritmo mais eficiente, etc), análise da corretude da implementação (ver item abaixo), e análise de desempenho do algoritmo implementados para diferentes valores de *ntree* em *datasets* selecionados pelos professores;
- Cada grupo deverá apresentar em seu relatório a árvore induzida por sua implementação para um conjunto de dados de *benchmark*, fornecido pelos professores juntamente com a enunciado do trabalho. O objetivo deste teste é permitir que os professores verifiquem a corretude da implementação básica de cada grupo. O conjunto de dados de *benchmark* está disponível no Moodle da disciplina. Espera-se que os grupos apresentem tanto a estrutura final da árvore induzida, como os valores de Ganho de Informação de cada divisão de nós realizada. Ao garantir a corretude da implementação do algoritmo de indução de árvores de decisão, os grupos poderão ter maior confiança no funcionamento do algoritmo de Florestas Aleatórias implementado. OBS: Os alunos devem atentar para as diferenças entre número de classes e tipos de atributos (categóricos/numéricos) entre o dataset benchmark e os datasets de avaliação
- Após verificada a corretude do algoritmo de indução de uma árvore de decisão, como descrito acima, cada grupo deverá treinar modelos de Florestas Aleatórias para os seguintes *datasets* básicos de classificação:

  1. German Credit Data Set *(34 atributos, 351 exemplos, 2 classes)*
     https://www.openml.org/d/31
     Objetivo: predizer se o risco de crédito de um cliente bancário é bom ou ruim
  2. Vertebral Column Data Set *(6 atributos, 310 exemplos, 3 classes)*
     https://www.openml.org/d/1523
     Objetivo: classificar pacientes ortopédicos em 3 classes (normal, hérnia de disco, espondilolistese) a partir de características biomecânicas
  3. Wine Data Set *(13 atributos, 178 exemplos, 3 classes)*
     https://www.openml.org/d/187
     Objetivo: predizer o tipo de um vinho baseado em sua composição química

4. (opcional; pontos extras) SPAM E-mail Data Set *(58 atributos, 4601 exemplos, 2 classes)*
https://www.openml.org/d/44
Objetivo: predizer se um e-mail é spam ou não

   OBS.: Caso o grupo deseje explorar diferentes *datasets* (incluindo *datasets* de interesse próprio), poderá fazê-lo. Experimentos adicionais serão analisados e poderão ser considerados para pontuação extra no trabalho.

- Todos os links disponibilizados acima possuem um link para download dos dados em formato .csv (recomendável). Observe na página de descrição dos data sets os tipos e distribuições dos atributos preditivos.
- Para cada *dataset* citado acima, utilize o método de **validação cruzada estratificada** para treinar e testar o algoritmo de Florestas Aleatórias. O objetivo é avaliar o desempenho deste algoritmo de classificação e comparar a *performance média* de modelos de Floresta Aleatória para diferentes valores do parâmetro *ntree*.
- Sugere-se utilizar $k = 10$ folds na validação cruzada. Destes k folds, $k - 1$ folds serão usados para construção do modelo de Florestas Aleatórias com *ntree* árvores, aplicando-se o algoritmo de indução de árvores de decisão combinado aos processos de amostragem de instâncias (bootstrap, i.e., amostragem com reposição) e de atributos (com base no parâmetro m). O fold restante (de teste) será usado para avaliar o desempenho do *ensemble*. Isto é, em cada etapa da validação cruzada, será gerada não apenas uma árvore de decisão, mas múltiplos modelos de árvore de decisão (tantos quantos forem o valor de *ntree*), os quais irão compor o modelo de Florestas Aleatórias. Perceba que como o treinamento de cada árvore envolve aleatoriedade, haverá diversidade entre as árvores geradas a partir do mesmo conjunto de $k - 1$ folds de treinamento. Este processo de treinamento e teste será repetido k vezes, de acordo com o funcionamento básico da estratégia de validação cruzada visto em aula. O uso de validação cruzada repetida é opcional.
- Como medida de performance para este trabalho, indicamos o uso da **F1-measure**, definida em função da precisão e do recall (atribuindo mesmo peso a ambas, isto é, $\beta = 1$). Observe que alguns problemas são de classificação binária, enquanto outros de classificação multiclasse.
- Mostre em seu relatório, através de gráficos, o desempenho do algoritmo conforme a variação no parâmetro *ntree*. De acordo com a teoria sobre aprendizado *ensemble*, espera-se diminuição no erro de classificação de acordo com o aumento no número de árvores no modelo de Florestas Aleatórias. Discuta acerca dos seus resultados no relatório final.
- Dica: Tendo em vista a aleatoriedade envolvida no treinamento do algoritmo de Florestas Aleatórias, recomenda-se utilizar o processo de inicialização manual de *seeds* de números aleatórios a fim de permitir a replicação dos resultados no processo de *debugging* do código.

3 Critérios de avaliação

- Pontualidade na entrega do trabalho. Atenção: atrasos na entrega do trabalho serão penalizados proporcionalmente ao tempo de atraso, **sendo descontado 1 (um) ponto por dia de atraso** (o trabalho como um todo vale 10 pontos).
- Apresentação oral dos resultados: qualidade da apresentação e domínio da implementação e resultados, bem como capacidade de arguição acerca dos mesmos
- Corretude do algoritmo básico de indução de árvore de decisão para os dados benchmark fornecidos, através da apresentação da árvore induzida e dos valores calculados para o Ganho de Informação no relatório final.
- Completude do trabalho e atendimento aos requisitos mínimos de implementação definidos neste enunciado (ver Seção 1).
- Avaliação correta do algoritmo implementado com os conjuntos de dados fornecidos, utilizando adequadamente as estratégias e métricas para análise de desempenho apresentadas em aula e sugeridas neste enunciado.

- Qualidade do relatório final

4 Política de Plágio

Grupos poderão **apenas** discutir questões de *alto nível* relativas a resolução do problema em questão. Poderão discutir, por exemplo, aspectos relacionados ao processo de validação cruzada ou de boostrap, bem como dificuldades gerais encontradas ao se analisar cada *dataset*. **Não** é permitido que os grupos utilizem quaisquer códigos fonte provido por outros grupos, ou encontrados na internet. Os alunos **poderão** fazer consultas na internet ou em livros **apenas** para estudar o modo de funcionamento das técnicas e algoritmos utilizados e analisar os respectivos **pseudo-códigos**. **Não** é permitida a análise ou cópia de implementações concretas (em quaisquer linguagens de programação) da técnica escolhida. O objetivo deste trabalho é justamente implementar a técnica do zero e descobrir as dificuldades envolvidas na sua utilização para resolução de um problema de aprendizado. Toda e qualquer fonte consultada pelo grupo (tanto para estudar os métodos a serem utilizados, quanto para verificar a estruturação da técnica em termos de *pseudo-código*) **precisa obrigatoriamente** ser citada no relatório final. Os professores utilizam rotineiramente um sistema anti-plágio que compara o código-fonte desenvolvido pelos grupos, bem como com implementações conhecidas e disponíveis online, ou de semestres passados.

Qualquer nível de plágio (ou seja, utilização de implementações que não tenham sido 100% desenvolvidas pelo grupo) poderá resultar em **nota zero** no trabalho. Caso a cópia tenha sido feita de outro grupo da disciplina, *todos* os alunos envolvidos (não apenas os que copiaram) serão penalizados. Esta política de avaliação **não** é aberta a debate posterior. Se você tiver quaisquer dúvidas sobre se uma determinada prática pode ou não, ser considerada plágio, ***não* assuma** nada: pergunte aos professores. Os grupos deverão desenvolver o trabalho **sozinhos**. Os professores estarão à disposição para sanar dúvidas ao longo do processo - recomendamos, no entanto, não deixar as dúvidas para o último momento!