Florestas Aleatórias (*Random Forests*)

1. Visão Geral

- Vimos que, no contexto de comitês, busca-se gerar um conjunto diversificado de máquinas para resolver determinado problema. No caso de ensembles, a ideia é que cada máquina busque lidar com a tarefa em mãos "a partir de perspectivas diferentes", gerando respostas que, combinadas, podem levar a uma melhor generalização.
- A aplicação dessa ideia no contexto de árvores de decisão traz algumas perspectivas interessantes, que merecem um tratamento em separado. Elas dão origem ao conceito de floresta aleatória (*random forest*).

- Antes, porém, começaremos de um ponto mais familiar: a construção de um ensemble "convencional" de árvores. Uma forma de construir um ensemble de árvores poderia ser, por exemplo, através de *bagging*. Nesse caso, um conjunto de dados de N amostras é amostrado (com reposição), gerando M "novos" conjuntos que, por sua vez, são usados para construir M árvores. As respostas dessas árvores são combinadas para gerar a saída do ensemble.
- A extensão desse ensemble para a noção de *random forest* tipicamente envolve a construção de árvores *a partir de subconjuntos de características*, de modo a promover diversificação.
- Em [Ho, 1995], discute-se que, se houver m atributos, haverá 2^m possíveis subconjuntos. É possível então pensar que, usando diferentes subconjuntos, gera-se diversidade.

- Em [Breiman, 2001], coloca-se, de fato, como estratégia simples de construção de uma *random forest* a definição de certo número de subconjuntos (de mesma cardinalidade) e a geração de árvores a partir deles.
- Em seu site [Breiman, 2019], Breiman apresenta uma estratégia básica sistemática:
  - Se há N amostras de treinamento, gere N conjuntos de treinamento com amostragem com reposição (*bootstrapping*).
  - Se há m atributos, escolha um número k de atributos ($k << m$) tal que, a cada nó, k atributos são escolhidos (de m) para particioná-lo. O valor k é mantido constante à medida que se constrói a floresta.
  - Cada árvore cresce sem limitação externa e sem poda.

- Há dois fatores cruciais nessa construção: a *capacidade* (*strength*) de cada árvore (que tem a ver com sua acurácia – árvores boas têm taxa de erro reduzida) e a *correlação* entre árvores.
- Um menor valor de k diminui a capacidade e a correlação. Um maior valor de k melhora a capacidade, mas aumenta também a correlação.
- O aumento do número de árvores, segundo [Breiman, 2019], não gera sobreajuste no ensemble, o que é uma característica interessante.
- Um grande número de árvores pode ser gerado em tempo relativamente curto (usando virtualmente qualquer algoritmo de indução de árvores).
- Em [Breiman, 2001], mostra-se que o método se beneficia do uso de *out-of-bag estimates*. Para que possamos entender o que isso significa, pensemos no processo de *bootstrapping*: amostra-se o conjunto com reposição até o tamanho

original. Naturalmente, devem ocorrer repetições e, desse modo, certos dados estarão ausentes de um determinado conjunto.

- Imagine que, para certo ponto ($\mathbf{x}_i$, d_i), obtenhamos a resposta de cada árvore do ensemble. Podemos então considerar apenas aquelas que não tiveram o referido ponto em seu conjunto de treinamento – essas árvores compõem um *out-of-bag classifier*. A resposta desse classificador pode servir como um indicador da qualidade da generalização do modelo.
- Uma aplicação possível desse conceito é a determinação da importância de atributos para certo problema. Gera-se a floresta como visto acima, e obtém-se o erro *out-of-bag* para cada dado. Quando se analisa um atributo *j*, permuta-se o valor desse atributo ao longo do conjunto de treinamento e se verifica o novo erro *out-of-bag*. A diferença entre o erro novo e o erro anterior gera um *score* que pode servir de norte.

1.1. Florestas Aleatórias, Medidas de Dissimilaridade e Clusterização

- Interessantemente, as ideias de construção de florestas aleatórias podem servir para medir similaridade / dissimilaridade [Shi e Horvath, 2006].
- No caso de dados rotulados, ou seja, num caso *supervisionado*, uma possibilidade é apresentar padrões $\mathbf{x}_i$ e $\mathbf{x}_j$ a cada árvore do ensemble e, se o nó terminal (folha) obtido for o mesmo, conta-se um valor "1" (caso contrário, conta-se "0"). Ao final, pode-se verificar o valor normalizado da soma obtida (percentual de árvores para as quais os dados tem o mesmo nó-folha). Esse índice funciona como uma medida de similaridade (ou leva a uma medida de dissimilaridade, se for o caso).
- No caso não-supervisionado, considera-se que se dispõe de um conjunto de dados não-rotulados *D*. Gera-se então um conjunto de dados sintéticos *D'* a

partir de uma distribuição de referência. Desse modo, estabelecem-se duas classes (D e D'), e passa a ser possível utilizar a medida para dados rotulados (focando no conjunto de dados "reais" D).

2. Referências bibliográficas

BREIMAN, L., “Random Forests”, Machine Learning, Vol. 45, pp. 5 – 32, 2001.

BREIMAN, L., https://www.stat.berkeley.edu/~breiman/RandomForests/, acessado em 30/06/2019.

HO, T. K., ”Random Decision Forests”, Third International Conference on Document Analysis and Recognition, Montreal, Canadá, 1995.

SHI, T., HORVATH, S., ”Unsupervised Learning with Random Forest Predictors”, Journal of Computational and Graphical Statistics, Vol. 15, No. 1, pp. 118 – 138, 2006.

WIKIPEDIA, *Artigos Diversos*, 2019.