Random Forests (Florestas Aleatórias)

Floresta aleatória é um método de aprendizado de máquina pertencente a classe de métodos de aprendizado conjunto (ensemble learning methods) pois não é um modelo unitário, mas uma combinação de outros modelos.

Mas, antes de falarmos de florestas aleatórias em si ...

Vamos falar brevemente do modelo que dá origem a uma floresta aleatória...as arvores de decisão (ou decision trees).

Árvores de Decisão Vamos imaginar um exemplo, uma pessoa que gosta de jogar futebol ao ar livre e, ao longo do tempo foi observado que o comparecimento dessa pessoa no jogo depende de uma série de fatores associados ao clima, como temperatura, umidade, chuva, vento, etc. Podemos representar isso com a seguinte estrutura:

Esse e um exemplo relativamente rudimentar, mas que exemplifica muito bem arvores de decisão, mas só para colocar um pouco de formalidade nos termos : Os pontos em que é tomada uma decisão entre qual caminho seguir são os nós (nas caixas violeta), a conexão entre esses nos são os ramos (nas setas azul marinho, com o titulo do ramo nas caixas da mesma cor), o nó inicial que inicia o processo de divisão é a raiz (Clima), os últimos nos em ramos são as folhas (nas caixas em laranja) e por fim a distância entre as folhas e a raiz é a altura da arvore (legal né?).

Observemos que intuitivamente já vemos o funcionamento básico de uma arvore, a partir de um conjunto de informações, vamos tomando decisões a partir de parâmetros que descrevem esse conjunto e por fim chegamos em uma resposta (ou classificação) do que queremos responder.

Mas vamos incrementar um pouco as coisas e imaginar que não queremos responder se apenas uma pessoa especifica vai ao jogo, mas sim queremos um modelo que responda isso para diversas pessoas que estão em condições climáticas diferentes, e, que o nosso modelo de arvore agora é um pouco mais elaborado, como:

Sim, um pouco exagerado, mas, agora que os nós em que decidimos qual ramo seguir ficam tão específicos, conforme a altura da arvore aumentou, que acabamos ajustando nosso modelo para uma situação par locais muito específicos, e com isso acabamos falhando no objetivo que era um modelo genérico .

Então, como corrigir isso?

Bom existem diversas maneiras, mas vamos seguir por uma delas, e, imaginar um modelo que não se baseasse em uma única arvore, mas várias, em que cada uma vai ser parametrizada com dados (nós) aleatórios que dividem bem nosso conjunto de dados, ou seja, não vamos ter a classificação baseada em uma única arvores, mas um conjunto delas, ou ... uma floresta .

Florestas Aleatórias Finalmente chegamos então ao modelo de florestas aleatórias, que assim como arvores de decisão pode ser usado para classificação ou regressão, e, para criarmos um modelo assim, uma das maneiras é quebrar nosso conjunto de dados em vários conjuntos menores, e, a partir deles criar uma série de arvores de decisão distintas, e a resposta desse modelo é geralmente a média das respostas das arvores que o compõe.

E para reduzirmos o ajuste excessivo dos dados a uma única situação, na criação das arvores de decisão usamos um método estocástico (uma palavra bonita pra aleatório) na escolha dos parâmetros que formam os nós, ou seja, escolhemos aleatoriamente, mas de maneira que os dados ficam bem divididos (para isso tem algumas funções matemáticas que conduzem esse processo), os pontos em que tomamos a decisão de qual ramo seguir.

Referências:

- Artigos Web:
  - Turing Talks - Modelos de Predição: Random Forest
  - Turing Talks - Modelos de Predição: Introdução à Predição
  - Turing Talks - Modelos de Predição: Decision Tree
- Livro:
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.) [PDF]. Springer. Cap. 8 (Tree-Based Methods)
- Fonte das imagens:
  - Figura 1
  - Figura 2
  - Figuras 3 e 4 eu criei no canvas baseado no modelo do livro citado acima
  - Figura 5