Random Forest

Fabrício Barth

Agosto de 2019

Ensemble Learning

- Métodos que geram diversos modelos e agregam o seu resultado.
- No caso do Random Forest, são geradas diversas árvores e cada árvore é gerada considerando apenas um sub-conjunto do conjunto de treinamento.

Random Forest

- O algoritmo possui apenas dois parâmetros configuráveis:
  - ⋆ quantidade de atributos considerados em cada árvore (m_{try}), e;
  - ⋆ quantidade de árvores (n_{tree}).

Random Forest

Para problemas de classificação e regressão o algoritmo funciona da seguinte forma:

- Cria n_{tree} sub-conjuntos de exemplos a partir do dataset original.
- Para cada sub-conjunto de exemplos cria-se uma árvore de classificação ou regressão sem poda. A criação de cada árvore considera apenas um sub-conjunto de exemplos: m_{try} atributos selecionados aleatoriamente e $2/3$ dos exemplos também selecionados aleatoriamente.

- A predição para novos dados acontece pela agregação das predições das n_{tree} árvores.
- Para problemas de **classificação** é considerado a maioria dos votos.
- Para problemas de **regressão** é considerado a média dos votos.

Particularidades de implementação no sklearn

```
max_features : int, float, string or None, optional (default="auto")
The number of features to consider when looking for the best split:

If int, then consider max_features features at each split.
If float, then max_features is a fraction and int(max_features * n_features)
If "auto", then max_features=sqrt(n_features).
If "sqrt", then max_features=sqrt(n_features) (same as ?auto?).
If "log2", then max_features=log2(n_features).
If None, then max_features=n_features.
```

```
max_depth : integer or None, optional (default=None)
The maximum depth of the tree. If None, then nodes are expanded
until all leaves are pure or until all leaves contain less than
min_samples_split samples.

warm_start : bool, optional (default=False)
When set to True, reuse the solution of the previous call to fit
and add more estimators to the ensemble, otherwise, just fit a
whole new forest.
```

Estimativa de erro

- Uma estimativa de erro, usando apenas o conjunto de treinamento, pode ser obtida através do conjunto de treinamento. Ao invés de ser utilizado algum outro método, como *cross-validation*.
- Para cada árvore construída é usado um sub-conjunto de exemplos. $1/3$ dos exemplos são mantidos fora do conjunto de treinamento. Estes exemplos mantidos fora do conjunto de treinamento são utilizados como teste.

Exemplos

https://github.com/fbarth/ml-espm/blob/master/scripts/python/05_01_random_forest.ipynb

Material de **consulta**

- Liaw and Wiener. Classification and Regression by randomForest. R News 2 (3): 18–22 (2002)
- Breiman and Cutler. Random Forests. Acessado em https://www.stat.berkeley.edu/ breiman/RandomForests/
- http://rpubs.com/fbarth/exemploRandomForest
- https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html