Random forest

Приемы улучшения классификаторов

- Stacking
- Bagging (bootstrap aggregation).
- Boosting

Приемы улучшения классификаторов

- Bagging (bootstrap aggregation).

  Подмножества обучающего множества с повторением
- Pasting

  Подмножества обучающего множества без повторениея

Stacking

- укладка в штабель; пакет

Random Forest

- Предложен
- Leo Breiman в статье
- «Random Forests». Machine Learning v 45 (1): 5–32 (2001)
-

Обозначения

- обучающая выборка состоит из N примеров,
- размерность пространства признаков равна M

- Строится ntree деревьев.
- Голосование: побеждает класс, за который проголосовало наибольшее число деревьев.

•Деревья разные благодаря bagging'y

- Каждое дерево обучается на своей подвыборке исходных данных.
- Подвыборку составляет sampsize наблюдений и mtry переменных.

Ключевые параметры модели

- ntree Число деревьев - сделай много, потом сокращай!
  (можно не обучать заново)
- mtry Число переменных в подвыборке
  - (Breiman: sqrt(M))
- sampsize Число наблюдений в подвыборке
  (Breiman: ceiling(.632*N))
- nodesize Минимальное число наблюдений в узле
  (Breiman: 1 наблюдение,
  У меня - 10 наблюдений...)
  replace Подвыборка с возвращением или нет (с возвращением)

Комментарий

- (1 — 1/e) = .632
- 2. Отбирая часть переменных добиваемся декорреляции результатов разных деревьев.
- 3. Леса могут использоваться в задачах регрессии
- 4. Ошибку можно измерять используя т. н. out-of-bag error

Информативность переменных

Importance переменной вычисляется как разность двух out-of-bag ошибок до и после перестановки значений в столбце.

Soft и hard голосование

- **Hard голосование** когда модели распознают класс объекта, а затем устраваем голосование
- **Soft голосование** когда модели распознают вероятности принадлежать классам, а затем вероятности усредняют по всем моделям
- **Но не все** модели имеют predict_proba() method
- soft голосование часто лучше, чем hard голосование
- soft голосование придает больший вес моделям, которые уверены в результате распознавания.