Random Forest

Plan prezentacji

- Drzewa Decyzyjne
- Random Forest
- Extra Trees

Drzewa decyzyjne

- Iteracyjnie dziel zbiór treningowy, według wartości na wybranej w każdym kroku zmiennej
  - Cięcia dobieraj tak by rozdzielać obserwacje należące do różnych klas (np. kryterium Gini)
- Przestań dzielić gdy warunek stopu zostanie spełniony (np. zbyt mały zbiór danych) i oblicz predykcje dla obszaru.

Drzewa decyzyjne cd.

- Zalety:
  - Mogą reprezentować złożone interakcje między zmiennymi
- Wady:
  - Słabo radzą sobie gdy trzeba wziąć pod uwagę wiele słabych predyktorów
- Inne:
  - Biorą pod uwagę tylko porządek punktów na prostej liczb rzeczywistych

Lasy Losowe

- Zbuduj wiele drzew decyzyjnych, uśrednij ich predykcje.
- Jak uzyskać zróżnicowane drzewa biorące pod uwagę różne predyktory?

Losowość w Lasach

- Bootstraping
  - Buduj każde drzewo na nieco innym 'podzbiorze' danych
- Losowy wybór predyktorów
  - Do każdego cięcia wylosuj zbiór cech z których następnie wybierzesz najlepszą. - Typowo pierwiastek z liczby wszystkich ficzerów.

Kryterium Stopu

- Klasyfikacja – budujemy tak głębokie drzewo jak to możliwe (liście wielkości 1)
- Regresja – Tu wybór parametru ma większe znaczenie – polecany minimalny rozmiar liścia 5.

- Każde takie drzewo dość mocno overfituje się do danych – ale liczymy (i wiemy z praktyki) że uśrednianie predykcji zredukuje nam ten błąd.

Lasy Losowe - OOB

- OOB error – out of bag error
- Jeśli stosujemy bootstraping to dla każdego elementu zbioru treningowego istnieje około 37 % procent drzew które go nie widziały. Można ich użyć do estymacji błędu na tym elemencie.
- Dzięki temu dobrze szacujemy błąd całego modelu.

Lasy Losowe – ile drzew posadzić?

- Ile drzew posadzić? - Tak dużo by otrzymać zbieżność wyników.
  - np. jeśli 300 i 600 drzew daje ten sam wynik (np. OOB) to 600 drzew powinno w zupełności wystarczyć.

Extra Trees

- Extra Trees (lub Extreemly Randomized Trees)
  - Losowe punkty podziału – losowo wybieramy nie tylko zmienne które mogą nam służyć do podziału w każdym węźle drzewa ale też konkretne punkty podziału (zamiast kierować się np. współczynnikiem Gini)
  - Wyniki są zazwyczaj trochę lepsze od RF.
  - Szkolenie drzewa jest znacznie krótsze (aczkolwiek trzeba wyszkolić ich trochę więcej)

Extra Trees – Parametry

- Domyślnie parametry dla klasyfikacji podobne do RF (liście wielkości 1, pierwiastek z predyktorów przy każdym splicie)
- Bootstraping nie poprawia rezultatów!
  - To sugeruje że wartość bootstrapingu w RF polegała właśnie na randomizacji optymalnych cięć.

Extra Trees – losowe punkty podziału

- Twórcy algorytmu sugerują kryterium losowania cięć z rozkładu jednostajnego między największym i najmniejszym punktem w zbiorze (taka jest też implementacja w scikit-learn)
- To powoduje że nie tylko porządek punktów na zmiennej ma znaczenie (trzeba wziąć to pod uwagę gdy nasz rozkład ma np. bardzo ciężki ogon i rozważyć zrangowanie zmiennych)

Jeszcze bardziej losowe drzewa?

- Nienadzorowana procedura budowy drzew:
  - Jeśli dla Extra Trees do każdego punktu podziału będziemy losować tylko jednego kandydata (zamiast pierwiastka z wszystkich) to otrzymamy nienadzorowaną procedurę budowy drzew.
  - Nadzorowane jest tylko przydzielenie liściom predykcji
  - Ta metoda daje czasem lepsze wyniki!
  - Nie działa gdy wiele zmiennych jest niezwiązanych z predykowaną.

Random Forest – podsumowanie

- Bardzo dobry, niemal uniwersalny predyktor
- Nie ma konieczności porównywania różnych hiperparametrów
- OOB (jeśli używa się bootstapingu)
- Warto też spróbować Extra Trees