Uczenie statystyczne dla łańcuchów Markowa

Głównym celem naszych badań jest uogólnienie teorii uczenia statystycznego, w przypadku gdy mamy do czynienia z danymi o strukturze łańcuchów Markowa.

Ogromna ilość informacji, które otrzymujemy z mediów, Internetu czy globalnych systemów komunikacji spowodowała, że metody uczenia statystycznego stały się bardzo ważnym narzędziem w analizie danych. Ogromne ilości danych są generowane przez wszechobecne narzędzia komunikacji, urządzenia mobilne, np. telefony komórkowe, kamery systemów bezpieczeństwa, drony, platformy medyczne i komercyjne, komunikatory internetowe, etc. Uczenie z tych danych pozwoli na znaczny rozwój nauki i technologii, jak również może poprawić warunki życia ludzi na całym świecie.

Liczne zastosowania uczenia statystycznego spowodowały dynamiczny rozwój tej dziedziny. Jednakże, teoria uczenia statystycznego jest dokładnie zbadana w przypadku niezależnych obserwacji, o tym samym rozkładzie. Takie założenie o danych bardzo często nie jest jednak spełnione w rzeczywistych problemach analizy danych. Prognoza rynkowa, rozpoznawanie mowy czy obrazów mają bardzo często zależną strukturę pomiędzy obserwacjami.

Nasz projekt koncentruje się na badaniu algorytmów uczących dla danych o strukturze Markowa. Modele, gdzie obserwacje są łańcuchem Markowa to np. systemy kolejkowania i przechowywania, a także wiele modeli finansowych, ubezpieczeniowych czy też takich, gdzie badacz jest zainteresowany wykrywaniem anomalii. Jednym z naszych głównych celów jest zbadanie wydajności algorytmów uczących (w przypadku gdy dane są łańcuchem Markowa) poprzez minimalizację ryzyka empirycznego. Jesteśmy głównie zainteresowani problemami klasyfikacji w przypadku danych Markowa. W uczeniu maszynowym i uczeniu statystycznym, klasyfikacja polega na przyporządkowaniu nowej obserwacji do klas, na podstawie próby uczącej, która zawiera obserwacje (przykłady), których klasa jest już znana. Przykładami problemu klasyfikacji jest filtrowanie wiadomości email jako spam lub nie-spam lub przyporządkowanie diagnozy do pacjenta na podstawie cech tego pacjenta (płeć, ciśnienie krwi, brak lub występowanie pewnych symptomów, etc.). W naszym projekcie zbadamy wydajność wybranych algorytmów uczących, zbadamy też zgodność oraz wybrane własności dla metody k najbliższych sąsiadów lub k średnich. Nasze badania motywujemy zastosowaniami praktycznymi, jednakże nasze wyniki będą w dużej mierze teoretyczne. Większość naszych wyników udowodnimy, używając nierówności wykładniczych dla łańcuchów Markowa Harrisa, które otrzymamy na początku tego projektu.

Nasze badania motywujemy też słowami V. Vapnika

Nie ma nic bardziej praktycznego niż dobra teoria.

Nasz projekt jest bardzo dobrym tego przykładem.