Miniprojekt 2

Plik `rp.data` zawiera wyniki badań diagnostycznych pozwalających stwierdzać, czy wykryty rak piersi jest łagodny czy złośliwy. Każdy wiersz pliku zawiera 10 liczb: pierwsze dziewięć z nich odpowiada wynikom dziewięciu różnych parametrów (każdy o wartości od 1 do 10), natomiast ostatnia dana jest prawdziwą informacją o nowotworze: 2 oznacza, że jest on łagodny, natomiast 4 świadczy o złośliwości.

Porównaj działanie regresji logistycznej oraz naiwnego klasyfikatora bayesowskiego **implementując** oba algorytmy i korzystając z danych zawartych w pliku `rp.data`. Załóż, że wszystkie dziewięć cech jest od siebie niezależnych.

Implementując klasyfikator bayesowski zastosuj wygładzenie Laplace'a oraz załóż, że dla cech $x_j \in \{1, \ldots, 10\}$, gdzie $j = 1, \ldots, 9$, zachodzi

$$p(x_j = d|y = c) = \phi^j_{c,d}, \text{ gdzie } \sum_{d=1}^{10} \phi^j_{c,d} = 1,$$

czyli że zmienne $x_j|y = c$ mają rozkład wielomianowy. Implementując regresję logistyczną możesz dodać składnik regularyzujący.

W obu algorytmach wykorzystaj zbiór 2/3 obserwacji jako zbiór treningowy, natomiast pozostałą 1/3 jako zbiór testowy. Zwróć uwagę na to, żeby wybrać 2/3 obserwacji dla każdej z klas.

Dla obu algorytmów **stwórz wykres** zawierający krzywe uczenia, czyli przedstawiające błąd klasyfikacji jako funkcję rozmiaru zbioru treningowego. Zaznacz punkty odpowiadające błędom obliczonym na całym zbiorze testowym po zastosowaniu algorytmu na następujących frakcjach zbioru treningowego: $0.01, 0.02, 0.03, 0.125, 0.625, 1$. Aby uwiarygodnić wyniki, uśrednij co najmniej 5 przebiegów algorytmu na losowych wyborach obserwacji do zbioru treningowego i testowego.

Opisz wnioski, jakie możesz wyciągnąć na podstawie osiągniętych wyników. Co możesz powiedzieć o zbieżności krzywej uczenia w obu przypadkach?

Zapoznaj się z artykułem *On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes* Andrew Nga i Michaela Jordana. Czy otrzymane przez Ciebie wyniki zgadzają się z teoretycznymi wynikami opisanymi w tej pracy?