Dokumentacja – laboratorium nr 7
Wprowadzenie do Sztucznej Inteligencji

Dominika Wyszyńska 318409

17 stycznia 2024

1 Wstęp

Celem zadania było zaimplementowanie naiwnego klasyfikatora Bayesa i zbadanie jego skuteczności w klasyfikacji gatunków irysów na podstawie zbioru danych Iris. Zbiór ten zawiera informacje o trzech różnych gatunkach irysów, opisanych za pomocą cech takich jak długość i szerokość płatków oraz kielichów. Należało także pamiętać o podziale zbioru danych na zbiór trenujący i uczący.

Klasyfikator ten opiera się na twierdzeniu Bayesa, które możemy przedstawić za pomocą poniższego wzoru:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

Gdzie:

- $P(y|X)$ to prawdopodobieństwo, że obserwacja X należy do klasy y,
- $P(X|y)$ to prawdopodobieństwo wystąpienia obserwacji X pod warunkiem, że należy do klasy y,
- $P(y)$ to prawdopodobieństwo wystąpienia klasy y,
- $P(X)$ to prawdopodobieństwo wystąpienia obserwacji X.

Naiwny klasyfikator Bayesa zakłada, że cechy obserwacji są wzajemnie niezależne pod warunkiem danej klasy. W naszym przypadku, przyjmujemy, że cechy są rozkładem normalnym (Gaussa) i używamy wzoru na gęstość prawdopodobieństwa rozkładu normalnego:

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}}$$

Gdzie:

- x_i to i-ta cecha obserwacji,
- μ_y to średnia wartość cechy dla klasy y,
- σ_y to odchylenie standardowe cechy dla klasy y.

Powyższe wzory są kluczowe dla obliczeń prawdopodobieństw i predykcji klas w naiwnym klasyfikatorze Bayesa.

2 Opis zaplanowanych eksperymentów

Eksperymenty zostały zaprojektowane w następujący sposób:

- Badanie wpływu stosunku podziału danych trenujących do danych uczących na działanie klasyfikatora
- Porównanie działania klasyfikatora dla różnych rozkładów prawdopodobieństwa (w tym rozkład Gaussa)

3 Wyniki i opis eksperymentów

Podczas eksperymentów przeprowadzono testy dla dwóch różnych gestości prawdopodobieństwa: gaussowskiej i jednostajnej (uniform). Gestość jednostajna o wzorze:

$$f(x) = \frac{1}{b-a}$$

gdzie: a to dolna granica przedziału, b to górna granica przedziału.

Eksperymenty polegały na równoczesnym testowaniu obu gestości prawdopodobieństwa, a także na zmienianiu stosunku wielkości danych uczących do danych trenujących. Celem było zrozumienie wpływu różnych gestości na wydajność klasyfikatora naiwnego Bayesa w zależności od proporcji danych uczących.

Podział danych: 20% - dane uczące, 80% - dane trenujące

- Naiwny klasyfikator Bayesa - rozkład jednostajny:
  - 43 poprawne trafienia, 77 pomyłek
  - Dokładność: 0.36
- Naiwny klasyfikator Bayesa - rozkład gaussowski:
  - 114 poprawnych trafień, 6 pomyłek
  - Dokładność: 0.95

Podział danych: 50% - dane uczące, 50% - dane trenujące

- Naiwny klasyfikator Bayesa - rozkład jednostajny:
  - 29 poprawnych trafień, 46 pomyłek
  - Dokładność: 0.39
- Naiwny klasyfikator Bayesa - rozkład gaussowski:
  - 74 poprawnych trafień, 1 pomyłka
  - Dokładność: 0.99

Podział danych: 70% - dane uczące, 30% - dane trenujące

- Naiwny klasyfikator Bayesa - rozkład jednostajny:
  - 19 poprawnych trafień, 26 pomyłek
  - Dokładność: 0.42
- Naiwny klasyfikator Bayesa - rozkład gaussowski:
  - 44 poprawnych trafień, 1 pomyłka
  - Dokładność: 0.98

Podział danych: 90% - dane uczące, 10% - dane trenujące

- Naiwny klasyfikator Bayesa - rozkład jednostajny:
  - 6 poprawnych trafień, 9 pomyłek
  - Dokładność: 0.4
- Naiwny klasyfikator Bayesa - rozkład gaussowski:
  - 15 poprawnych trafień, 0 pomyłek
  - Dokładność: 1.0

Obserwacje wskazują na to, iż stosunek między danymi uczącymi a danymi trenującymi może wpływać na skuteczność klasyfikatora. W niektórych przypadkach, zwłaszcza przy mniejszym zbiorze danych uczących, klasyfikator mógł wykazywać niższą dokładność. Dlatego istotne jest zbalansowanie proporcji między danymi uczącymi a trenującymi, aby uzyskać optymalne wyniki klasyfikacji.

Porównanie obu rozkładów prawdopodobieństwa

Porównując wyniki eksperymentów dla obu rozkładów prawdopodobieństwa, można zauważyć, że rozkład gaussowski osiągnął zazwyczaj o wiele wyższą dokładność w porównaniu do rozkładu jednostajnego. Rozkład gaussowski zdaje się lepiej radzić sobie z różnymi proporcjami danych uczących, uzyskując wyższe trafienia i wyższą ogólną dokładność klasyfikacji.

4 Podsumowanie

Podsumowując, w kontekście klasyfikatora naiwnego Bayesa, gestość prawdopodobieństwa o charakterze gaussowskim okazała się być bardziej skuteczna w zadaniach klasyfikacyjnych niż gestość jednostajna. Ponadto, zauważono, że optymalny stosunek między danymi uczącymi a danymi trenującymi może wpływać na dokładność klasyfikacji, szczególnie w przypadku mniejszych zbiorów danych uczących. Warto zadbać o zrównoważone proporcje między tymi dwoma rodzajami danych, aby osiągnąć najlepsze wyniki klasyfikacji.