Raport do zadania 7

Naiwny klasyfikator Bayes’a

Kacper Siemionek

Numer indeksu: 331430

1. Wpływ atrybutów na jakość klasyfikatora

Oceniono dokładność klasyfikatora, stosując wszystkie kombinacje 5 z 7 zdyskretyzowanych wcześniej atrybutów, które miały największy wpływ na wyniki. Pozostałe atrybuty (Name, Ticket, Cabin) nie miały znaczenia w kontekście przetrwania. Średnia dokładność została policzona przez walidację krzyżową z podziałem na 10 części.

Tabela 1.1 Dokładność klasyfikatora dla kombinacji różnych atrybutów.

Pclass	Sex	Age	Fare	SibSp	Parch	Embarked	Średnia dokładność	Różnica
✓	✓	✓	✓	✓			0.7755	0.0787
✓	✓	✓	✓		✓		0.7643	0.1011
✓	✓	✓	✓			✓	0.7779	0.1316
✓	✓	✓		✓	✓		0.7811	0.1910
✓	✓	✓		✓		✓	0.7721	0.1124
✓	✓	✓			✓	✓	0.7722	0.1348
✓	✓		✓	✓	✓		0.7768	0.1986
✓	✓		✓	✓		✓	0.7688	0.1032
✓	✓		✓		✓	✓	0.7665	0.1685
✓	✓			✓	✓	✓	0.7923	0.0787
✓		✓	✓	✓	✓		0.7082	0.0974
✓		✓	✓	✓		✓	0.7026	0.0526
✓		✓	✓		✓	✓	0.7205	0.1124
✓		✓		✓	✓	✓	0.6903	0.1461
✓			✓	✓	✓	✓	0.7060	0.1798
	✓	✓	✓	✓	✓		0.7655	0.1236
	✓	✓	✓	✓		✓	0.7699	0.1236
	✓	✓	✓		✓	✓	0.7856	0.1124
	✓	✓		✓	✓	✓	0.7790	0.1461
	✓		✓	✓	✓	✓	0.7721	0.1573
		✓	✓	✓	✓	✓	0.6869	0.1348

Średnie wyniki walidacji krzyżowej dla powyższych kombinacji mieszczą się w przedziale 0,68 – 0,79. Największą dokładność osiągnął klasyfikator z atrybutami *Pclass Sex, SibSp, Parch, Embarked* (0,7923), a najmniejszą z atrybutami *Age, Fare, SibSp, Parch, Embarked* (0,6869).

Większe różnice pomiędzy wynikami w poszczególnych częściach walidacji krzyżowej mogą świadczyć o niedouczeniu klasyfikatora dla danych atrybutów, zwłaszcza przy niższej średniej dokładności. Oznacza to, że model w większości testów osiąga stosunkowo niskie wyniki i nie jest w stanie nauczyć się prawidłowych wzorców.

2. Poprawienie działania klasyfikatora

Jak poprawić działanie klasyfikatora?

- **Obsługa ciągłych atrybutów** – zamiast dyskretyzować niektóre atrybuty, możemy liczyć ich prawdopodobieństwo, zakładając, że mają rozkład normalny.
- **Laplace smoothing** – dodając wygładzenie α, pozbywamy się zerowych prawdopodobieństw, które mogły wystąpić w zbiorze treningowym, wtedy klasyfikator nie wyklucza kompletnie danej sytuacji podczas testów.
- **Testowanie różnej liczby atrybutów** – ładując do klasyfikatora więcej danych, możemy polepszyć jego dopasowanie do danych, jednak może się to wiązać z przeuczeniem, szczególnie gdy użyjemy za dużo atrybutów. Analogicznie zmniejszając liczbę atrybutów, zwiększamy ryzyko niedouczenia. Kluczowe jest dobranie najbardziej istotnych cech dla konkretnego przypadku.