Zadanie 2

Bartosz Kozłowski

Na podstawie podanych cech obserwacji sprawdzono efektywność klasyfikacji z użyciem naiwnego klasyfikatora Bayes'owskiego, LDA i QDA. Dane z plików zostały pobrane do tabeli, którą następnie rozdzielono na tablice parametrów i tablice odpowiadających im klas. Następnie wykorzystując funkcje ALL obliczono parametry dla metody z powtórnym podstawieniem dla wszystkich parametrów oraz dla pierwszych 2, 5 i 10. Za pomocą funkcji ALL_2 obliczono wymagane parametry dla przypadku rozdzielonych danych. Za pomocą funkcji ALL_3 wykonano obliczenia parametrów dla kroswalidacji dla k = 5. Wszystkie funkcje ALL przyjmują jako argument tablice danych. Program wyświetla następujące wyniki:

Wszystkie parametry
Powtorne podstawienie

	ACC	TP	TN	TPR	FPR
NB	0.9888	176	178	0.9888	0.0111
LDA	1.0000	178	178	1.0000	0.0000
QDA	0.9944	177	178	0.9944	0.0056

2 parametry
Powtorne podstawienie

	ACC	TP	TN	TPR	FPR
NB	0.8256	144	178	0.8090	0.1604
LDA	0.8256	144	178	0.8090	0.1604
QDA	0.8303	145	178	0.8146	0.1564

5 parametrow
Powtorne podstawienie

	ACC	TP	TN	TPR	FPR
NB	0.8639	152	178	0.8539	0.1275
LDA	0.8836	156	178	0.8764	0.1100
QDA	0.8936	158	178	0.8876	0.1010

10 parametrow
Powtorne podstawienie

	ACC	TP	TN	TPR	FPR
NB	0.9614	171	178	0.9607	0.0378
LDA	0.9888	176	178	0.9888	0.0111
QDA	0.9944	177	178	0.9944	0.0056

Podzielony zbior danych

	ACC	TP	TN	TPR	FPR
NB	0.8367	37	45	0.8222	0.1509
LDA	0.8182	36	45	0.8000	0.1667
QDA	0.8367	37	45	0.8222	0.1509

Wynik z kroswalidacja

	ACC	TP	TN	TPR	FPR
LDA	0.7845	135	178	0.7584	0.1946

Na podstawie przedstawionych wyników, widać że w przypadku powtórnego podstawienia dla wszystkich parametrów najlepszy jest klasyfikator LDA, lecz gdy trochę zmniejszymy liczbę uwzględnionych parametrów najlepszym klasyfikatorem staje się QDA. W przypadku rozdzielonych danych najlepszym klasyfikatorem jest Naive Bayes i QDA posiadające taką samą dokładność. Widać, że w przypadku metody powtórnego podstawienia wyniki klasyfikatorów są

najlepsze, ponieważ model jest dopasowany do właśnie tych danych. Można także zauważyć spadek dokładności wraz z spadkiem uwzględnionych zmiennych. Gdy zbiór danych zostanie rozdzielony na dane wejściowe i testowe klasyfikatory osiągają dokładność zbliżoną do metody powtórnego podstawienia, lecz są bardziej wiarygodne niż wyniki poprzedniej metody, ponieważ klasyfikator nie jest sprawdzany na tych samych danych na podstawie których został stworzony. W przypadku kroswalidacji wartość dokładności klasyfikatora jest najniższa, w odległości ok. 0.3 od wyniku uzyskanego przy rozdzieleniu danych. Wynika z tego, że prawdopodobnie rzeczywista wartość dokładności dla klasyfikatora LDA zawiera się w przedziale 0.7845 - 0.8182.