4. U ovom zadatku potrebno je napraviti programsku izvedbu naivnog Bayesovog klasifikatora i provesti eksperimente nad stvarnim podatcima.

(a) Napravite programsku izvedbu naivnog Bayesovog klasifikatora za diskretne ulaze u programskome jeziku po izboru. Klasifikator treba podržavati izračun ML-procjena te izračun zaglađenih (Laplaeovih) procjena.

Napravio u MATlabu, zad4.m, nalazi se u arhivi u mapi programska podrška.

(b) Skinite skup podataka *Iris* s adrese

`http://archive.ics.uci.edu/ml/datasets/iris`

U njemu se opisuju biometrijske značajke cvjetova triju vrsta perunike odnosno irisa (*virginica*, *setosa* i *versicolor*). Skup najprije razdijelite slučajnim odabirom na podskup za učenje i ispitivanje u omjeru 2:1. Diskretizirajte značajke na skupu za učenje tako da približno trećina vrijednosti u svakom stupcu bude preslikana u jednu diskretnu vrijednost čime ćemo dobiti skup s diskretnim značajkama koje imaju po tri vrijednosti (radi jednostavnosti, možemo ih označiti brojevima 0, 1 i 2).

(c) Naučite model naivnog Bayesovog klasifikatora na skupu za učenje. Ispišite empirijsku pogrešku i pogrešku generalizacije.

Pogrešku sam računao kao broj krivo klasificiranih primjera podijeljen sa brojem ukupno klasificiranih primjera. Greške su otprilike 0.05 (5%) na skupu za učenje (empirijska pogreška) i 0.06 (6%) na skupu za provjeru (pogreška generalizacije).

(d) Ponovite eksperimente koristeći Laplaceovo zaglađivanje. Komentirajte rezultate.

Pogreške opet dođu na otprilike isto kao kod ML-procjenitelja, empirijska pogreška oko 0.05 i pogreška generalizacije oko 0.06 (ove pogreške, kao i one navedene u podzadatku c, variraju ovisno o pokretanju programa zbog nasumičnog odabira skupova za učenje i provjeru).

Budući da su greške iste sa i bez korištenja Laplaceovog zaglađivanja, možemo zaključiti da nam nema prevelikog utjecaja na odluku klasifikatora ako je vjerojatnost da se neka značajka inicijalizirala na određenu vrijednost 0 ili neki mali broj.

Valjda možemo reći da su primjeri dobro raspoređeni, da nemamo veći broj outlinera koji bi značajnije utjecali na pogrešku generalizacije ako se ne bi pojavljivali u skupu za učenje?