Rješenje zadatka 2.1 predmeta Strojno učenje

Siniša Biđin

8. prosinca 2012.

(a) Kod Bayesovog klasifikatora, optimalna je klasifikacijska odluka, odnosno hipoteza, ona koja maksimizira aposteriornu vjerojatnost $P(C_j|\mathbf{x})$:

$$h(\mathbf{x}) = \underset{C_j}{\operatorname{argmax}}\, p(C_j|\mathbf{x})P(C_j)$$

Bayesov je klasifikator generativan i parametarski model; generativan jer računa aposteriornu vjerojatnost na temelju zajedničke vjerojatnosti $P(C_j, \mathbf{x})$, a parametarski jer učenje svodi na optimizaciju fiksnog broja parametara pretpostavljene distribucije (složenost modela ne ovisi o broju primjera za učenje).

(b) Naivan Bayesov klasifikator pretpostavlja da su varijable međusobno uvjetno nezavisne za zadanu klasu, odnosno:

$$P(x_i|x_j, C) = P(x_i|C)$$

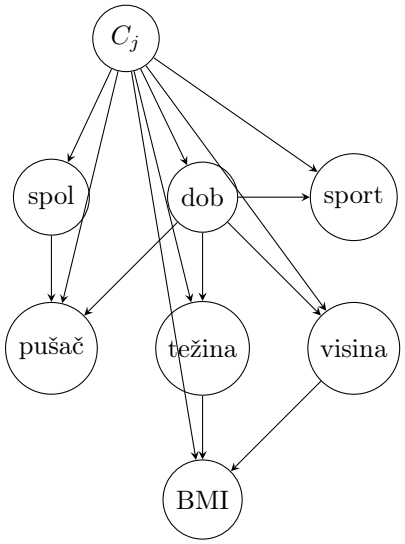
Pošto ta pretpostavka u praksi uglavnom ne vrijedi, klasifikator nazivamo naivnim. Radi se o pristranosti ograničavanjem, jer hipoteze unaprijed ograničavamo samo na one koje pretpostavljaju uvjetnu nezavisnost varijabli:

$$h(x_1, \ldots, x_n|C_j) = \underset{j}{\operatorname{argmax}}\, P(C_j) \prod_{k=1}^{n} P(x_k|C_j)$$

Pretpostavku moramo uvesti jer inače model ima previše parametara, vrlo visoku varijancu, sklon je prenaučenosti te stoga loše generalizira. Nakon pretpostavke o uvjetnoj nezavisnosti, broj parametara modela više ne ovisi eksponencijalno o dimenziji n, već linearno.

(c) Naivan model u ovom slučaju ne bi bio dobar izbor, jer pretpostavlja nezavisnosti varijabli koje su nisu nezavisne. Na primjer, indeks tjelesne mase zasigurno ovisi (i to potpuno) o visini i težini osobe.

Polunaivan model mogao bi ispravno predstaviti takve ovisnosti među varijablama. Na primjer: BMI ovisi o visini i težini; visina i težina ovise o dobi; indikacija je li osoba pušač ovisi o spolu i dobi; indikacija bavi li se osoba sportom ovisi o dobi. Odgovarajuća Bayesova mreža mogla bi izgledati ovako:



Uz varijable imenovane $x_1 = \text{spol}, x_2 = \text{dob}, x_3 = \text{visina}, x_4 = \text{težina}, x_5 = \text{BMI}, x_6 = \text{pušač}, x_7 = \text{sport}$, faktorizacija zajedničke vjerojatnosti bila bi tada:

$$\begin{aligned} P(\mathbf{x}, C_j) = {} & P(x_1|C_j)P(x_2|C_j)P(x_6|x_1, x_2, C_j)P(x_3|x_2, C_j) \\ & P(x_4, x_2, C_j)P(x_5|x_3, x_4, C_j)P(x_7|x_2, C_j)P(C_j) \end{aligned}$$