Rješenje zadatka 2.4 predmeta Strojno učenje

Siniša Biđin

9. prosinca 2012.

(a) (i) Slijedi ispis redundantnog skupa svih parametara.

$$
\begin{aligned}
P(x_0 = \text{high}|y = \text{acc}) &= 0.250 & P(x_0 = \text{high}|y = \text{unacc}) &= 0.271\\
P(x_0 = \text{low}|y = \text{acc}) &= 0.271 & P(x_0 = \text{low}|y = \text{unacc}) &= 0.211\\
P(x_0 = \text{med}|y = \text{acc}) &= 0.312 & P(x_0 = \text{med}|y = \text{unacc}) &= 0.223\\
P(x_0 = \text{vhigh}|y = \text{acc}) &= 0.167 & P(x_0 = \text{vhigh}|y = \text{unacc}) &= 0.295\\
P(x_1 = \text{high}|y = \text{acc}) &= 0.271 & P(x_1 = \text{high}|y = \text{unacc}) &= 0.259\\
P(x_1 = \text{low}|y = \text{acc}) &= 0.250 & P(x_1 = \text{low}|y = \text{unacc}) &= 0.223\\
P(x_1 = \text{med}|y = \text{acc}) &= 0.312 & P(x_1 = \text{med}|y = \text{unacc}) &= 0.223\\
P(x_1 = \text{vhigh}|y = \text{acc}) &= 0.167 & P(x_1 = \text{vhigh}|y = \text{unacc}) &= 0.295\\
P(x_2 = 2|y = \text{acc}) &= 0.188 & P(x_2 = 2|y = \text{unacc}) &= 0.278\\
P(x_2 = 3|y = \text{acc}) &= 0.262 & P(x_2 = 3|y = \text{unacc}) &= 0.247\\
P(x_2 = 4|y = \text{acc}) &= 0.275 & P(x_2 = 4|y = \text{unacc}) &= 0.238\\
P(x_2 = \text{5more}|y = \text{acc}) &= 0.275 & P(x_2 = \text{5more}|y = \text{unacc}) &= 0.238\\
P(x_3 = 2|y = \text{acc}) &= 0.000 & P(x_3 = 2|y = \text{unacc}) &= 0.456\\
P(x_3 = 4|y = \text{acc}) &= 0.525 & P(x_3 = 4|y = \text{unacc}) &= 0.266\\
P(x_3 = \text{more}|y = \text{acc}) &= 0.475 & P(x_3 = \text{more}|y = \text{unacc}) &= 0.278\\
P(x_4 = \text{big}|y = \text{acc}) &= 0.000 & P(x_4 = \text{big}|y = \text{unacc}) &= 0.000\\
P(x_4 = \text{med}|y = \text{acc}) &= 0.562 & P(x_4 = \text{med}|y = \text{unacc}) &= 0.466\\
P(x_4 = \text{small}|y = \text{acc}) &= 0.438 & P(x_4 = \text{small}|y = \text{unacc}) &= 0.534\\
P(x_5 = \text{high}|y = \text{acc}) &= 0.583 & P(x_5 = \text{high}|y = \text{unacc}) &= 0.224\\
P(x_5 = \text{low}|y = \text{acc}) &= 0.000 & P(x_5 = \text{low}|y = \text{unacc}) &= 0.456\\
P(x_5 = \text{med}|y = \text{acc}) &= 0.417 & P(x_5 = \text{med}|y = \text{unacc}) &= 0.319
\end{aligned}
$$

$$
\begin{aligned}
P(x_0 = \text{high}|y = \text{good}) &= 0.000 & P(x_0 = \text{high}|y = \text{vgood}) &= 0.000 \\
P(x_0 = \text{low}|y = \text{good}) &= 0.667 & P(x_0 = \text{low}|y = \text{vgood}) &= 0.600 \\
P(x_0 = \text{med}|y = \text{good}) &= 0.333 & P(x_0 = \text{med}|y = \text{vgood}) &= 0.400 \\
P(x_0 = \text{vhigh}|y = \text{good}) &= 0.000 & P(x_0 = \text{vhigh}|y = \text{vgood}) &= 0.000 \\
P(x_1 = \text{high}|y = \text{good}) &= 0.000 & P(x_1 = \text{high}|y = \text{vgood}) &= 0.200 \\
P(x_1 = \text{low}|y = \text{good}) &= 0.667 & P(x_1 = \text{low}|y = \text{vgood}) &= 0.400 \\
P(x_1 = \text{med}|y = \text{good}) &= 0.333 & P(x_1 = \text{med}|y = \text{vgood}) &= 0.400 \\
P(x_1 = \text{vhigh}|y = \text{good}) &= 0.000 & P(x_1 = \text{vhigh}|y = \text{vgood}) &= 0.000 \\
P(x_2 = 2|y = \text{good}) &= 0.200 & P(x_2 = 2|y = \text{vgood}) &= 0.000 \\
P(x_2 = 3|y = \text{good}) &= 0.267 & P(x_2 = 3|y = \text{vgood}) &= 0.200 \\
P(x_2 = 4|y = \text{good}) &= 0.267 & P(x_2 = 4|y = \text{vgood}) &= 0.400 \\
P(x_2 = 5\text{more}|y = \text{good}) &= 0.267 & P(x_2 = 5\text{more}|y = \text{vgood}) &= 0.400 \\
P(x_3 = 2|y = \text{good}) &= 0.000 & P(x_3 = 2|y = \text{vgood}) &= 0.000 \\
P(x_3 = 4|y = \text{good}) &= 0.533 & P(x_3 = 4|y = \text{vgood}) &= 0.400 \\
P(x_3 = \text{more}|y = \text{good}) &= 0.467 & P(x_3 = \text{more}|y = \text{vgood}) &= 0.600 \\
P(x_4 = \text{big}|y = \text{good}) &= 0.000 & P(x_4 = \text{big}|y = \text{vgood}) &= 0.000 \\
P(x_4 = \text{med}|y = \text{good}) &= 0.533 & P(x_4 = \text{med}|y = \text{vgood}) &= 1.000 \\
P(x_4 = \text{small}|y = \text{good}) &= 0.467 & P(x_4 = \text{small}|y = \text{vgood}) &= 0.000 \\
P(x_5 = \text{high}|y = \text{good}) &= 0.667 & P(x_5 = \text{high}|y = \text{vgood}) &= 1.000 \\
P(x_5 = \text{low}|y = \text{good}) &= 0.000 & P(x_5 = \text{low}|y = \text{vgood}) &= 0.000 \\
P(x_5 = \text{med}|y = \text{good}) &= 0.333 & P(x_5 = \text{med}|y = \text{vgood}) &= 0.000
\end{aligned}
$$

(ii) Klasificiramo primjer $\mathbf{x} = \{\text{high, vhigh, 3, 2, big, med}\}$:

```
ghci> import Bayes
Bayes> d <- loadDataset "CarEvaluation2.txt" carFeats
Bayes> naive d (words "high vhigh 3 2 big med")
[(0.0,"vgood"),(0.0,"unacc"),(0.0,"good"),(0.0,"acc")]
```

Aposteriorne vjerojatnosti klasa za zadani primjer sve iznose 0.0, jer za svaku klasu postoji barem jedna značajka čija se vrijednost u skupu primjera nije realizirala. Model je prenaučen.

(iii) Ponovno ispisujemo redundantan skup svih parametara.

$$\begin{aligned}
P(x_0 = \text{high}|y = \text{acc}) &= 0.250 & P(x_0 = \text{high}|y = \text{unacc}) &= 0.271 \\
P(x_0 = \text{low}|y = \text{acc}) &= 0.270 & P(x_0 = \text{low}|y = \text{unacc}) &= 0.212 \\
P(x_0 = \text{med}|y = \text{acc}) &= 0.311 & P(x_0 = \text{med}|y = \text{unacc}) &= 0.223 \\
P(x_0 = \text{vhigh}|y = \text{acc}) &= 0.168 & P(x_0 = \text{vhigh}|y = \text{unacc}) &= 0.294 \\
P(x_1 = \text{high}|y = \text{acc}) &= 0.270 & P(x_1 = \text{high}|y = \text{unacc}) &= 0.259 \\
P(x_1 = \text{low}|y = \text{acc}) &= 0.250 & P(x_1 = \text{low}|y = \text{unacc}) &= 0.223 \\
P(x_1 = \text{med}|y = \text{acc}) &= 0.311 & P(x_1 = \text{med}|y = \text{unacc}) &= 0.223 \\
P(x_1 = \text{vhigh}|y = \text{acc}) &= 0.168 & P(x_1 = \text{vhigh}|y = \text{unacc}) &= 0.294 \\
P(x_2 = 2|y = \text{acc}) &= 0.189 & P(x_2 = 2|y = \text{unacc}) &= 0.278 \\
P(x_2 = 3|y = \text{acc}) &= 0.262 & P(x_2 = 3|y = \text{unacc}) &= 0.247 \\
P(x_2 = 4|y = \text{acc}) &= 0.275 & P(x_2 = 4|y = \text{unacc}) &= 0.238 \\
P(x_2 = \text{5more}|y = \text{acc}) &= 0.275 & P(x_2 = \text{5more}|y = \text{unacc}) &= 0.238 \\
P(x_3 = 2|y = \text{acc}) &= 0.004 & P(x_3 = 2|y = \text{unacc}) &= 0.456 \\
P(x_3 = 4|y = \text{acc}) &= 0.523 & P(x_3 = 4|y = \text{unacc}) &= 0.266 \\
P(x_3 = \text{more}|y = \text{acc}) &= 0.473 & P(x_3 = \text{more}|y = \text{unacc}) &= 0.278 \\
P(x_4 = \text{big}|y = \text{acc}) &= 0.004 & P(x_4 = \text{big}|y = \text{unacc}) &= 0.001 \\
P(x_4 = \text{med}|y = \text{acc}) &= 0.560 & P(x_4 = \text{med}|y = \text{unacc}) &= 0.465 \\
P(x_4 = \text{small}|y = \text{acc}) &= 0.436 & P(x_4 = \text{small}|y = \text{unacc}) &= 0.534 \\
P(x_5 = \text{high}|y = \text{acc}) &= 0.580 & P(x_5 = \text{high}|y = \text{unacc}) &= 0.225 \\
P(x_5 = \text{low}|y = \text{acc}) &= 0.004 & P(x_5 = \text{low}|y = \text{unacc}) &= 0.456 \\
P(x_5 = \text{med}|y = \text{acc}) &= 0.416 & P(x_5 = \text{med}|y = \text{unacc}) &= 0.320
\end{aligned}$$

$$\begin{aligned}
P(x_0 = \text{high}|y = \text{good}) &= 0.020 & P(x_0 = \text{high}|y = \text{vgood}) &= 0.034\\
P(x_0 = \text{low}|y = \text{good}) &= 0.633 & P(x_0 = \text{low}|y = \text{vgood}) &= 0.552\\
P(x_0 = \text{med}|y = \text{good}) &= 0.327 & P(x_0 = \text{med}|y = \text{vgood}) &= 0.379\\
P(x_0 = \text{vhigh}|y = \text{good}) &= 0.020 & P(x_0 = \text{vhigh}|y = \text{vgood}) &= 0.034\\
P(x_1 = \text{high}|y = \text{good}) &= 0.020 & P(x_1 = \text{high}|y = \text{vgood}) &= 0.207\\
P(x_1 = \text{low}|y = \text{good}) &= 0.633 & P(x_1 = \text{low}|y = \text{vgood}) &= 0.379\\
P(x_1 = \text{med}|y = \text{good}) &= 0.327 & P(x_1 = \text{med}|y = \text{vgood}) &= 0.379\\
P(x_1 = \text{vhigh}|y = \text{good}) &= 0.020 & P(x_1 = \text{vhigh}|y = \text{vgood}) &= 0.034\\
P(x_2 = 2|y = \text{good}) &= 0.204 & P(x_2 = 2|y = \text{vgood}) &= 0.034\\
P(x_2 = 3|y = \text{good}) &= 0.265 & P(x_2 = 3|y = \text{vgood}) &= 0.207\\
P(x_2 = 4|y = \text{good}) &= 0.265 & P(x_2 = 4|y = \text{vgood}) &= 0.379\\
P(x_2 = \text{5more}|y = \text{good}) &= 0.265 & P(x_2 = \text{5more}|y = \text{vgood}) &= 0.379\\
P(x_3 = 2|y = \text{good}) &= 0.021 & P(x_3 = 2|y = \text{vgood}) &= 0.036\\
P(x_3 = 4|y = \text{good}) &= 0.521 & P(x_3 = 4|y = \text{vgood}) &= 0.393\\
P(x_3 = \text{more}|y = \text{good}) &= 0.458 & P(x_3 = \text{more}|y = \text{vgood}) &= 0.571\\
P(x_4 = \text{big}|y = \text{good}) &= 0.021 & P(x_4 = \text{big}|y = \text{vgood}) &= 0.036\\
P(x_4 = \text{med}|y = \text{good}) &= 0.521 & P(x_4 = \text{med}|y = \text{vgood}) &= 0.929\\
P(x_4 = \text{small}|y = \text{good}) &= 0.458 & P(x_4 = \text{small}|y = \text{vgood}) &= 0.036\\
P(x_5 = \text{high}|y = \text{good}) &= 0.646 & P(x_5 = \text{high}|y = \text{vgood}) &= 0.929\\
P(x_5 = \text{low}|y = \text{good}) &= 0.021 & P(x_5 = \text{low}|y = \text{vgood}) &= 0.036\\
P(x_5 = \text{med}|y = \text{good}) &= 0.333 & P(x_5 = \text{med}|y = \text{vgood}) &= 0.036
\end{aligned}$$

Ponovno klasificiramo prethodan primjer, no ovaj put koristeći Laplaceovo zaglađivanje:

```
Bayes> d' <- loadDataset' "CarEvaluation2.txt" carFeats
Bayes> naive d' (words "high vhigh 3 2 big med")
[(2.472e-6,"unacc"),(1.616e-8,"acc"),
 (6.361e-10,"good"),(2.52e-10,"vgood")]
```

Pošto je $P(y = \text{unacc}|\mathbf{x})$ najveći, primjer klasificiramo u klasu unacc.

(iv) Empirijska pogreška iznosi 0.1175, a pogreška generalizacije 0.13.

(b) *Preskočeno.*