

Leren: Schrijfopdracht 4

Week 5 - Semester 1 2014/15

Individueel maken

Deadline 8 december 2014 12.00 uur

1. (1 punt) *Beslisbomen*

Doe twee iteraties van het beslisboomleeralgorithme met de hand met de data in onderstaande tabel. Maak dus een root knoop en daaronder nog 1 andere knoop daaronder. Gebruik information gain voor de keuze van de variabelen. Het is niet nodig alle splitsingen in intervallen uit te rekenen. Neem een splitsing die er op het oog het beste uitziet en bereken daarvan de information gain. NB. het is mogelijk om een interval weer op te splitsen in subintervallen met een variabele die hoger al gebruikt is. Wat voorspelt deze beslisboom voor voorbeeld 7?

ID	X1	X2	Y
1	1	3	p
2	1	4	p
3	3	3	p
4	7	5	q
5	9	2	q
6	12	8	p
7	10	5	?

2. (1 punt) *Regel van Bayes*

Stel dat in het voorbeeld dat Mitchell (ch. 6, sectie 6.2.1) geeft van de regel van Bayes de patient nog een keer naar een ziekenhuis gaat en zich opnieuw laat testen. De test is weer positief. Hoe groot is nu de kans dat hij de ziekte heeft?

3. (1 punt) *Bayes*

Stel dat we de verdeling van verkoopprijzen van huizen in Nederland hebben over 6 maanden. Deze verdeling is normaal (Gaussian) met gemiddelde μ_{NL} en standaardafwijking σ_{NL} . We hebben ook gegevens van de prijzen in Gouda van de laatste 2 maanden. Van de huizen in Gouda hebben we ook (numerieke) gegevens zoals oppervlak, aantal kamers, bouwjaar. De variabelen lijken niet echt normaal verdeeld, maar we nemen aan dat dit komt door de kleine steekproef. We willen deze twee soorten informatie met elkaar combineren tot een voorspeller van huizenprijzen in Gouda. Omdat de voorspellingen niet erg precies zijn willen we als uitkomst een kansverdeling hebben, dus $P(prijs|gegevens)$.

(a) Geef een formule of een algoritme hiervoor dat gebaseerd is op de regel van Bayes.

(b) Als we multivariate lineaire regressie toepassen op de Gouda data dan krijgen we een hypothese die de prijs voorspelt. We kunnen met de data die we hebben de verdeling schatten van $P(prijs|gegevens)$. Wat is het verband tussen deze verdeling en die van de vorige deelvraag 3 (a)?