Leren Opdracht 6

Mag in teams van 2
Deadline 15 december 2014, 12.00 uur

1. (2 punten) Implementeer een programma dat de kans op een datapunt (vector) uitrekent in een meerdimensionale Gaussian (normale) kansdichtheidsverdeling. Pas dit toe op de data van digits123 als volgt: bereken voor de eerste op de eerste twee voorbeelden gegeven de klasse waar ze bijhoren, dus P(eerste voorbeeld van klasse 1 | klasse 1), etc.

2. (2 punten) Verander nu de labels van de eerste twee voorbeelden van 1 in 2, van 2 in 3 en van 3 in 1. Deze voorbeelden zijn nu dus fout gelabeld. Implementeer de methode van Andrew om anomalies te vinden. Pas die toe op de data van elke klasse apart, inclusief de fout gelabelde voorbeelden. Doe alsof niet bekend is dat er anomalies zijn. Worden de aangebrachte anomalies gevonden? Vergelijk de kans op deze voorbeelden met die van de vorige vraag.

3. (2 punten) Implementeer k-means clustering. Haal de klasse labels uit de data digits123 weg (of althans, gebruik ze niet). Pas nu k-means clustering toe op de data digits123 met k=1, k=2, k= 3. k=4 en k=5.

   (a) Is er een elbow point? Zoja, waar. Zonee, hoe kan het aantal clusters geoptimaliseerd worden?

   (b) Vergelijk de clusters die worden gevonden met $k = 3$ met de echte labels. Stel dat we de klasse die in een cluster het meest voorkomt als klasse nemen, hoeveel voorbeelden worden dan fout voorspeld?