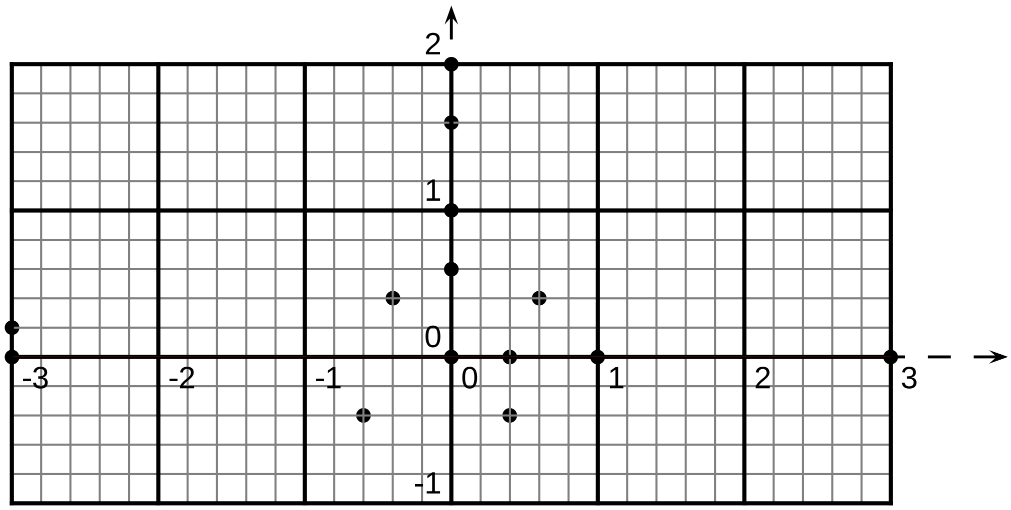
Universidade do Minho (DMA)
janeiro 2019

Machine Learning:: Fundamental and Application
Teste Modelo

Nome ____________________

Número ____________________ Curso ____________________

Nos métodos de clustering, a existência de *outliers* pode afectar a definição dos clusters é produzir resultados errados devido a grande sensibilidade da media aos valores extremos. Por outro lado, os *outliers* são indicadores de deviações importantes que pode ser interpretadas (fraude). O objetivo deste estudo é avaliar diferentes metodologias para determianr os *outliers*. Consideramos o caso simplificado de dois atributos reais $\mathcal{A} = \mathbb{R} \times \mathbb{R}$ e $D = (x^n)_{n=1,\cdots,N}$ um conjunto de eventos com $x_1^n, x_2^n \in \mathbb{R}$. A distância entre dois eventos $d(x, x')$ é da Manhattan. A figura representa uma configuração que usarmos nas aplicações.



Parte A.

Seja ε um parametro (que fixaremos mais a frente). Dizemos que um evento $x \in D$ é um ε-*outlier* de D se $d(x, x') > \varepsilon$ para qualquer $x' \in D$, $x' \neq x$.

1) No gráfico, determinar os ε-*outliers* com $\varepsilon = 0.1$, $\varepsilon = 0.8$, $\varepsilon = 1.4$.

2) Determinar os valores de ε_m e ε_M tais como se $\varepsilon < \varepsilon_m$ todos os elementos de D são *outliers* e se $\varepsilon > \varepsilon_M$, não existe nenhum *outlier*.

3) Porque a definição de ε-*outliers* não é muito útil. Quais são as principais contras?

Parte B.

Seja $x \in D$ e $\varepsilon > 0$. Notamos por ε-vizinhança de x o sub-conjunto

$$V(x; \varepsilon) = \{x' \in D, \text{ tal como }, d(x, x') \leq \varepsilon\}$$

e $|V(x; \varepsilon)|$ representa o número de elementos do conjunto.

1) Determinar $|V((0,0); 0.8)|$, $|V((-3,0); 0.8)|$, $|V((0,2); 0.8)|$.

2) Seja $\beta \in \mathbb{R}$. Um ponto é um *outlier* se

$$|V(x; \varepsilon)| \leq (1 + \beta)|V(x; \varepsilon/2)|.$$

Mostrar que se $\varepsilon > 0$ e $\beta < 0$, não pode haver *outlier*. Qual condição devemos impor ao β para que a definição faz sentido? O que se passa quando $beta > N - 1$?

3) Seja $\beta = 0.5$, determinar se os pontos $x = (0,0)$, $x = (-3,0)$ e $x = (0,2)$ são *outliers* com $\varepsilon = 0.6$ e $\varepsilon = 1.0$. Procurar os outros *outliers*.

4) O que melhora em relação à Parte A. O que fica ainda a resolver?

Parte C.

Seja $x \in D$, calculamos todas as distancias $d(x, x')$, $x' \in D$, que ordenamos como

$$d_0(x) = d(x, x) \leq d_1(x) \leq d_2(x) \leq \cdots \leq d_{N-1}(x),$$

onde notamos $x^{n_i} \in D$ tal como $d_i(x) = d(x, x^{n_i})$, $i = 0, \cdots, N-1$.

1) Determinar os 5 primeiros valores de $d_i(x)$ com $x = (0, 0)$ e $x = (-3, 0)$. O que caracteriza um *outlier*?

2) Seja $\varepsilon > 0$, definimos o índice m tal como

$$d_i - d_{i-1} \leq \varepsilon,\ i = 1, \cdots, m \qquad d\cancel{_{m+1} - d_m} > \cancel{\varepsilon}.$$

Se não existir tal m, pomos $m = N - 1$. Determinar o índice m para os pontos $x = (0, 0)$, $x = \cancel{(-3, 0)}$ ~~e~~ $x = \cancel{(0, 1.6})$ com $\varepsilon =\cancel{0.4}$.

3) Em prática, para identificar os conjuntos de *outliers*, introduzimos um número limite `nbOut` e consideramos que o ponto x é um *outlier* se $m \leq$`nbOut`. Usamos aqui `nbOut`~~=3~~. Determinar quais são, entre os eventos $x = (0, 0)$, $x = (-3, 0)$ e $x = (0, 1.6)$, os *outliers* com $\varepsilon = 0.4$. Existem outros *outliers*?

4) Determinar de novo os *outliers* com $\varepsilon = 1.0$.

Correção

Parte A.

1) Se $\varepsilon = 0.1$ todos os pontos são *outliers*. Se $\varepsilon = 0.8$, temos $x = (-0.6, -0.4)$ e $x = (3, 0)$. Se $\varepsilon = 1.4$, apenas o evento $x = (3, 0)$ é um *outlier*.

2) A distância entre dois pontos distintos é $\varepsilon_m = 0.2$ logo se $\varepsilon < \varepsilon_m$ todos os eventos serão considerados como *outliers*. Por outro lado, para $x = (3, 0)$,a distância mínima é 2 com os outros pontos. Logo seja $\varepsilon_M = 2$, para qualquer $\varepsilon > \varepsilon_M$, não há *outlier*.

3) Esta definição não permite de identificar os pontos $x = (-3, 0)$ e $(-3, 0.2)$ como *outliers* assim que os dois pontos $x = (0, 2)$ e $x = (0, 1.8)$. Apenas os *outliers* bem isolados são detectados.

Parte B.

1) $|V((0, 0); 0.8)| = 5$, $|V((-3, 0); 0.8)| = 2$, $|V((0, 2); 0.8)| = 2$.

2) Se $\beta < 0$ temos de procurar os x que verificam a relação $|V(x; \varepsilon)| < |V(x; \varepsilon/2)|$. Mas, como $V(x; \varepsilon/2) \subset V(x; \varepsilon)$, temos necessaramente $|V(x; \varepsilon)| \geq |V(x; \varepsilon/2)|$ logo não existe evento x que pode ser *outlier* se β negativo. Concluimos que é necessario $\beta \geq 0$ para detectar os *outliers*. Se $\beta > N - 1$, e notando que pelo menos $|V(x; \varepsilon/2)| = 1$ (o conjunto contém pelo menos x) a condição torna $|V(x; \varepsilon)| \leq N$ o que é automaticamente verificado porque $|D| = N$. Deduzimos que $\beta \in [0, N - 1]$ e que os valores de β próximos de 0 correspondem aos *outliers* mais isolados.

3) Com $\varepsilon = 0.6$ e $\beta = 0.5$, temos a condição $|V(x; 0.6)| \leq 1.5|V(x; 0.3)|$. Os pontos $(0, 0)$, $(0, 2)$ não são *outliers* enquanto $(-3, 0)$ é um *outlier*. Podemos também observar que os pontos $(-0.6, -0.4)$ e $(3, 0)$ são *outliers*.
Quando $\varepsilon = 1.0$, a condição torna $|V(x; 1.0)| \leq 1.5|V(x; 0.5)|$. $x = (0, 0)$ não é um *outlier*, mas $(2, 0)$ é agora um *outlier*. Os outros *outliers* são ainda $(-0.6, -0.4)$ e $(3, 0)$.

4) Conseguimos identificar o evento $(2, 0)$ como *outlier* mas a definição inclue $(-0.6, -0.4)$ como *outlier* enquanto parece pertenecer ao cluster.

Parte C.

1) Com $x = (0, 0)$, temos $d_0 = 0$, $d_1 = 0.4$, $d_2 = 0.6$, $d_3 = d_4 = 0.8$, $d_5 = 1.0$. Como $x = (-3, 2)$ temos $d_0 = 0$,$d_1 = 0.2$, $d_2 = 2.8$,$d_3 = d_4 = 3.3$, $d_5 = 3.4$. Um *outlier* é caracterizado por uma brutal variação do valor de d.

2) Para o ponto $x = (0, 0)$ temos $\cancel{m = 9.}$ Para $x = (-3, 0)$, $\cancel{m = 2}$ ~~e~~ para $x = (0, 1.6)$ temos $\cancel{m = 2.}$

3) Os pontos $x \cancel{= (-3, 0)}$ ~~e~~ $x = \cancel{(0, 1.6)}$ são *outliers* assim que $(\cancel{-3, 0.2})$, $\cancel{(0, 2)}$ ~~e~~ $(\cancel{3, 0})$. O ponto $(-0.6, -0.4)$ é também um *outlier*.

4) Com $\varepsilon = 1.0$, os eventos $(-3, 0.2)$, $(-3, 0.0)$ e $(3, 0)$ são *outliers*.