technische universität
dortmund

Zusatzfolien

K-Means

Im Rahmen der Proseminar-Vortragsreihe
"Grundlagen des Data-Minings für strukturierte Daten"
Dr. Nils M. Kriege

Antonie Vietor

5. Februar 2018

Fakultät für Informatik
Algorithm Engineering (Ls11)
Technische Universität Dortmund

Herleitung der Varianz beim BFR-Algorithmus

$$s^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2})
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right)
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^{\,2}\right)
\end{aligned}
$$

$$\sum_{i=1}^{n} 2x_i\bar{x}$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\sum_{i=1}^{n} 2x_i\bar{x} = 2\bar{x}\sum_{i=1}^{n} x_i
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n} \bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n}2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n}x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n}x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x}
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\cdot n \cdot \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x} \\
&= 2n\bar{x}^{\,2}
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^{\,2} + n\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x} \\
&= 2n\bar{x}^{\,2}
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^{\,2} + n\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^{\,2}\right)
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x} \\
&= 2n\bar{x}^{\,2}
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x} \\
&= 2n\bar{x}^2
\end{aligned}
$$

Herleitung der Varianz beim BFR-Algorithmus

$$
\begin{aligned}
s^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}\,)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^{\,2}) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\bar{x} + \sum_{i=1}^{n}\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^{\,2} + n\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^{\,2}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^{\,2} = \frac{SUMSQ}{N} - \left(\frac{SUM}{N}\right)^2
\end{aligned}
$$

$$
\begin{aligned}
\sum_{i=1}^{n} 2x_i\bar{x} &= 2\bar{x}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\,\frac{n}{n}\sum_{i=1}^{n} x_i \\
&= 2\bar{x}\cdot n\cdot\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= 2\,\bar{x}\,n\,\bar{x} \\
&= 2n\bar{x}^{\,2}
\end{aligned}
$$

Clusterzentrum als Mittelwert der Datenpunktkoordinaten

Koordinaten des Clusterzentrums entsprechen für jede Dimension den Mittelwerten der Datenpunktkoordinaten für diese Dimension, also:

$$c = \frac{1}{n} \sum_{i=1}^{n} a_i$$

Clusterzentrum als Mittelwert der Datenpunktkoordinaten

Koordinaten des Clusterzentrums entsprechen für jede Dimension den Mittelwerten der Datenpunktkoordinaten für diese Dimension, also:

$$
\begin{aligned}
c &= \frac{1}{n} \sum_{i=1}^{n} a_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix}
\end{aligned}
$$

Clusterzentrum als Mittelwert der Datenpunktkoordinaten

Koordinaten des Clusterzentrums entsprechen für jede Dimension den Mittelwerten der Datenpunktkoordinaten für diese Dimension, also:

$$
\begin{aligned}
c &= \frac{1}{n} \sum_{i=1}^{n} a_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix} \\
&= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} a_{i1} \\ \sum_{i=1}^{n} a_{i2} \end{pmatrix}
\end{aligned}
$$

Clusterzentrum als Mittelwert der Datenpunktkoordinaten

Koordinaten des Clusterzentrums entsprechen für jede Dimension den Mittelwerten der Datenpunktkoordinaten für diese Dimension, also:

$$
\begin{aligned}
c &= \frac{1}{n} \sum_{i=1}^{n} a_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix} \\
&= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} a_{i1} \\ \sum_{i=1}^{n} a_{i2} \end{pmatrix} \\
&= \begin{pmatrix} \overline{a_1} \\ \overline{a_2} \end{pmatrix}
\end{aligned}
$$

Clusterzentrum als Mittelwert der Datenpunktkoordinaten

Koordinaten des Clusterzentrums entsprechen für jede Dimension den Mittelwerten der Datenpunktkoordinaten für diese Dimension, also:

$$
\begin{aligned}
c &= \frac{1}{n} \sum_{i=1}^{n} a_i \\
&= \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} a_{i1} \\ a_{i2} \end{pmatrix} \\
&= \frac{1}{n} \begin{pmatrix} \sum_{i=1}^{n} a_{i1} \\ \sum_{i=1}^{n} a_{i2} \end{pmatrix} \\
&= \begin{pmatrix} \overline{a_1} \\ \overline{a_2} \end{pmatrix} \\
&= \bar{a}
\end{aligned}
$$

Gegeben sei:

$$a_i = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

$$\rightarrow \bar{a} = 3$$

Gegeben sei:

$$a_i = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

$$\rightarrow \bar{a} = 3$$

Es gilt $\sum(a_i - \bar{a}) = 0$:

$$\sum_{i=1}^{n}(a_i - \bar{a}) = -2 + (-1) + 0 + 1 + 2 = 0$$

Gegeben sei:

$$a_i = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{pmatrix}$$

$$\rightarrow \bar{a} = 3$$

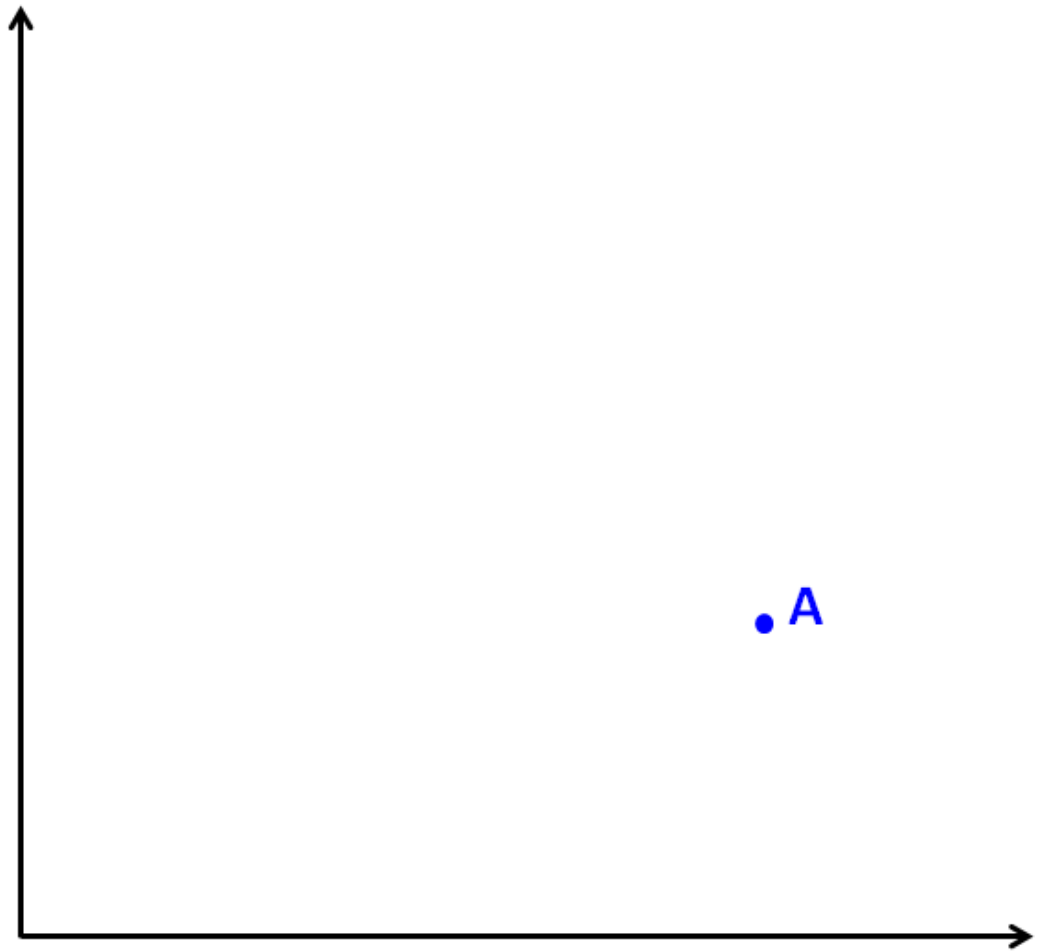
Es gilt $\sum(a_i - \bar{a}) = 0$:

$$\sum_{i=1}^{n}(a_i - \bar{a}) = -2 + (-1) + 0 + 1 + 2 = 0$$

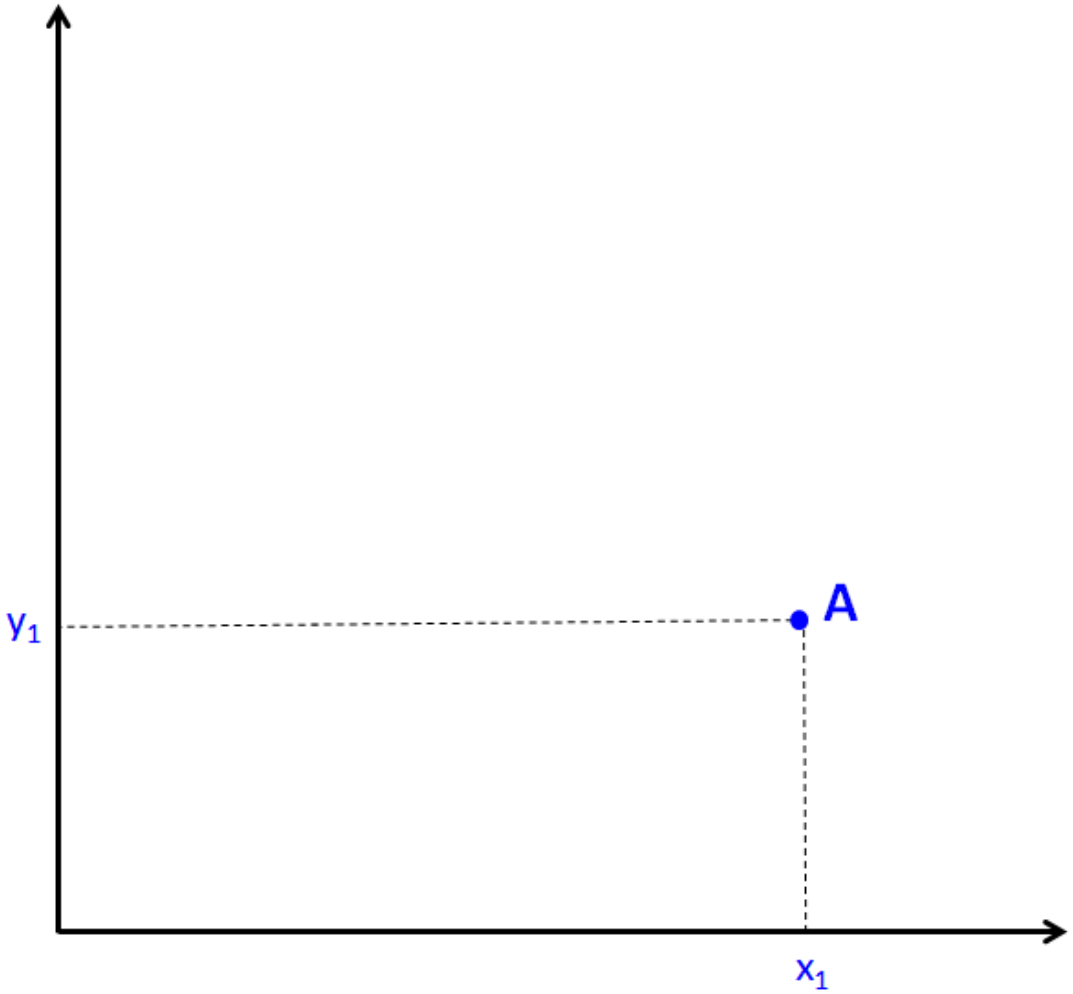
Aber $\sum(a_i - \bar{a})^2 \neq 0$:

$$\sum_{i=1}^{n}(a_i - \bar{a})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

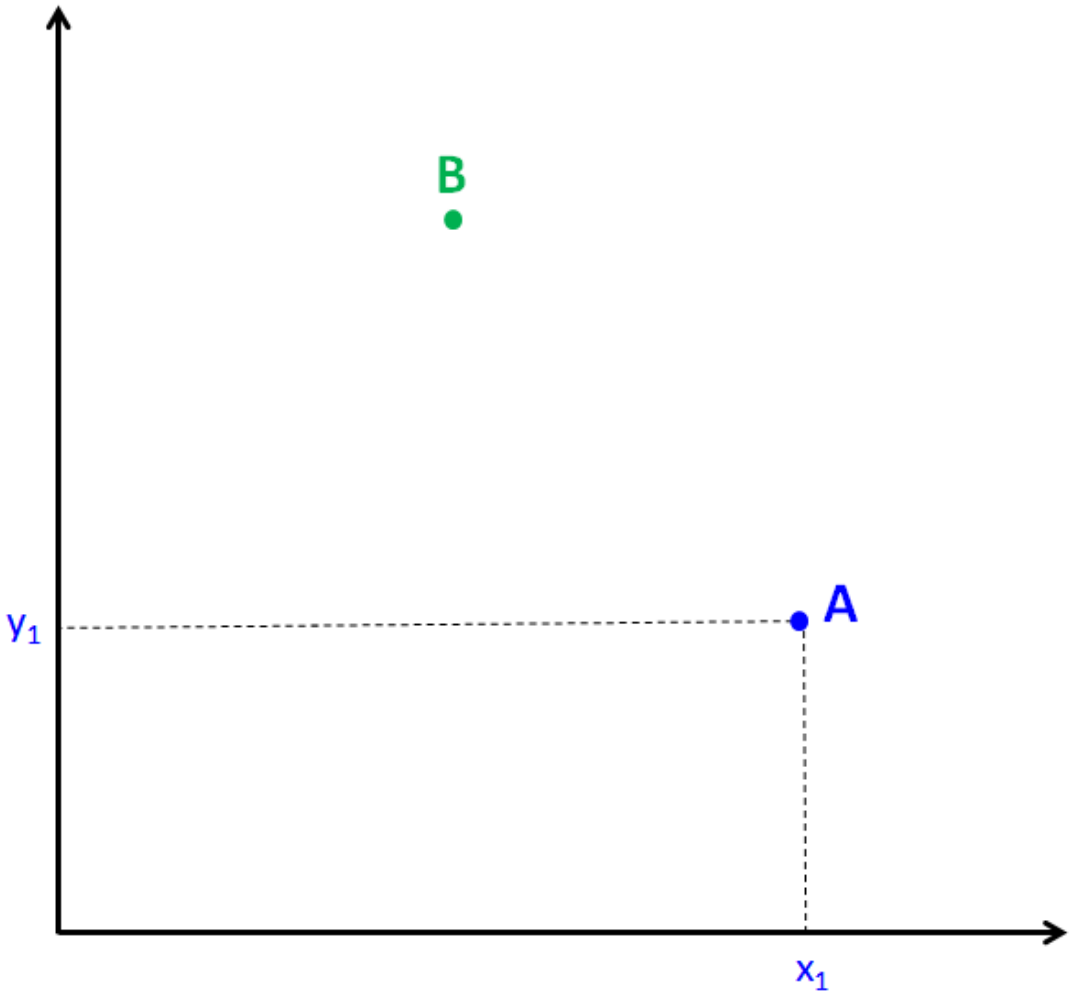
Abstand zwischen 2 Punkten



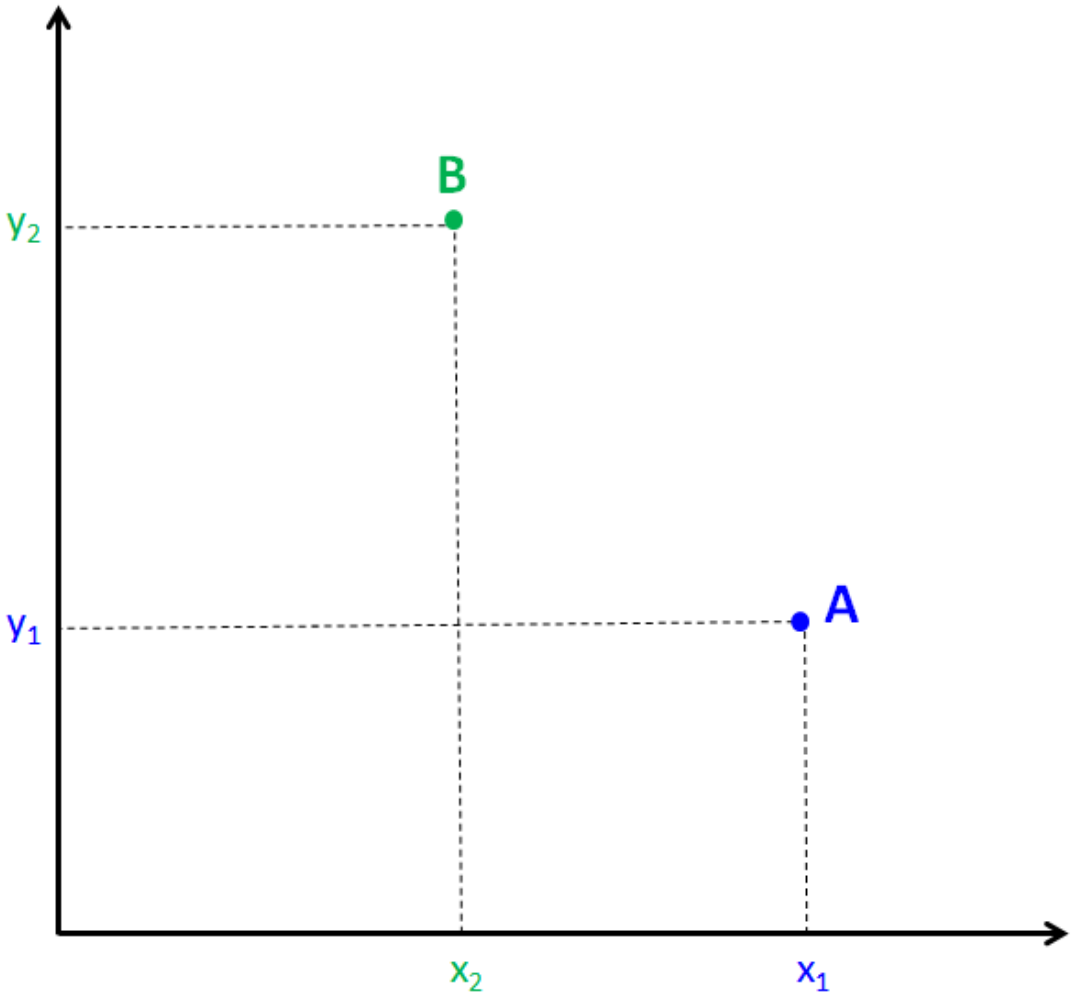
Abstand zwischen 2 Punkten



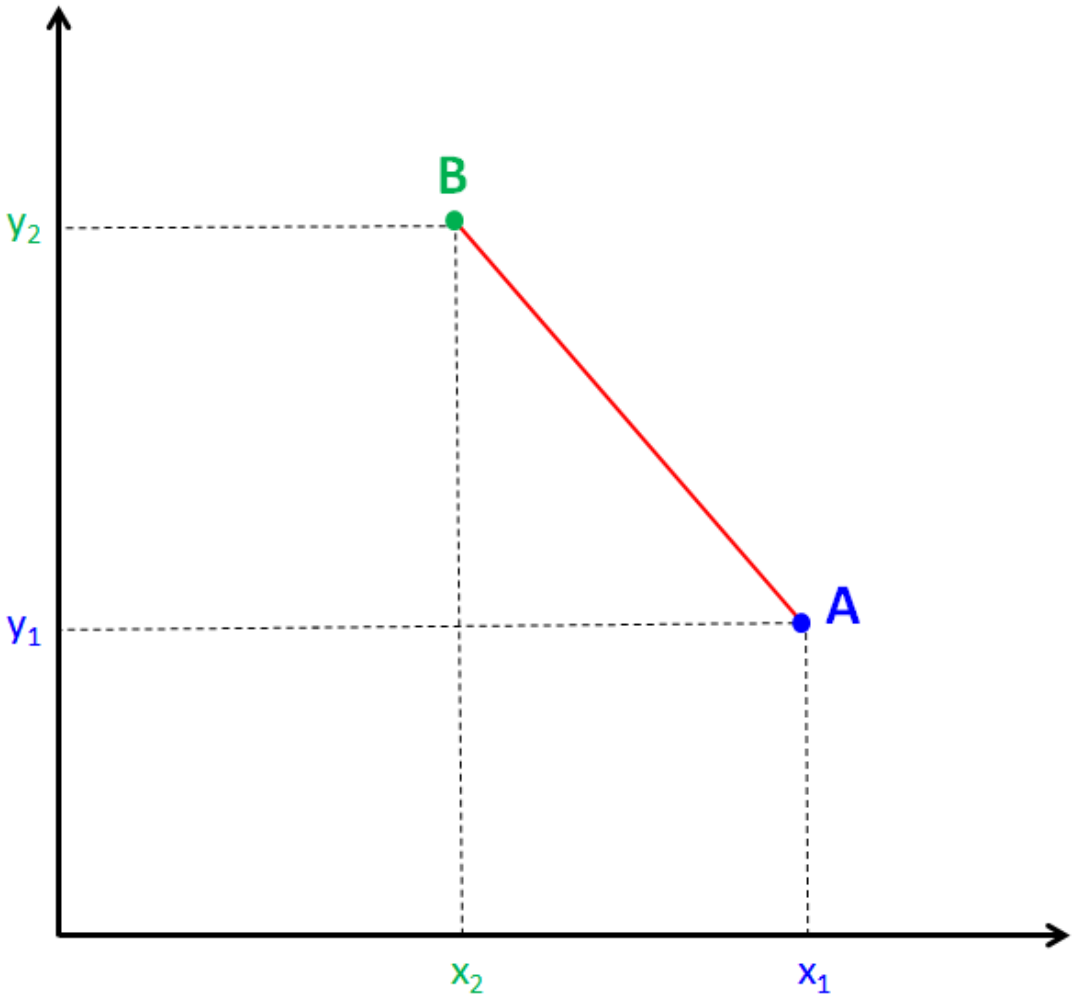
Abstand zwischen 2 Punkten



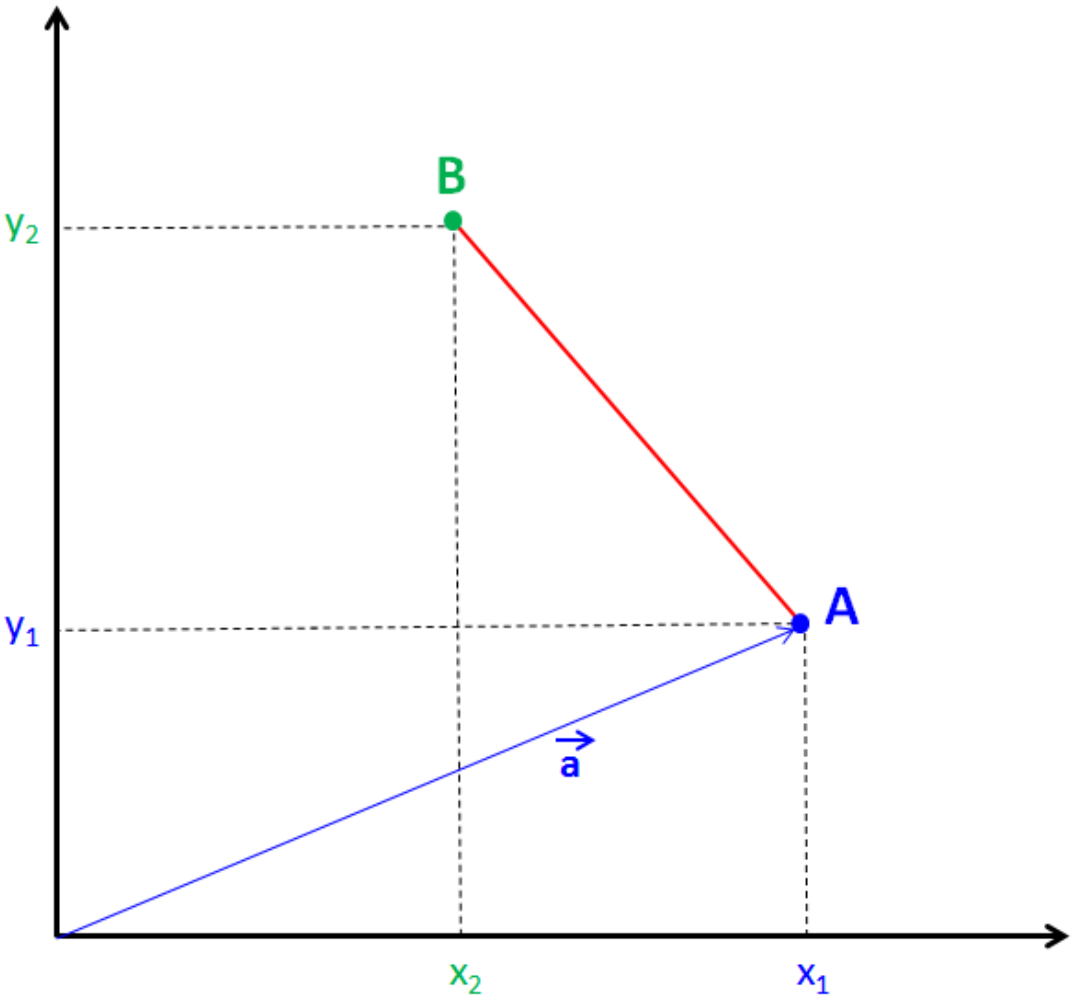
Abstand zwischen 2 Punkten



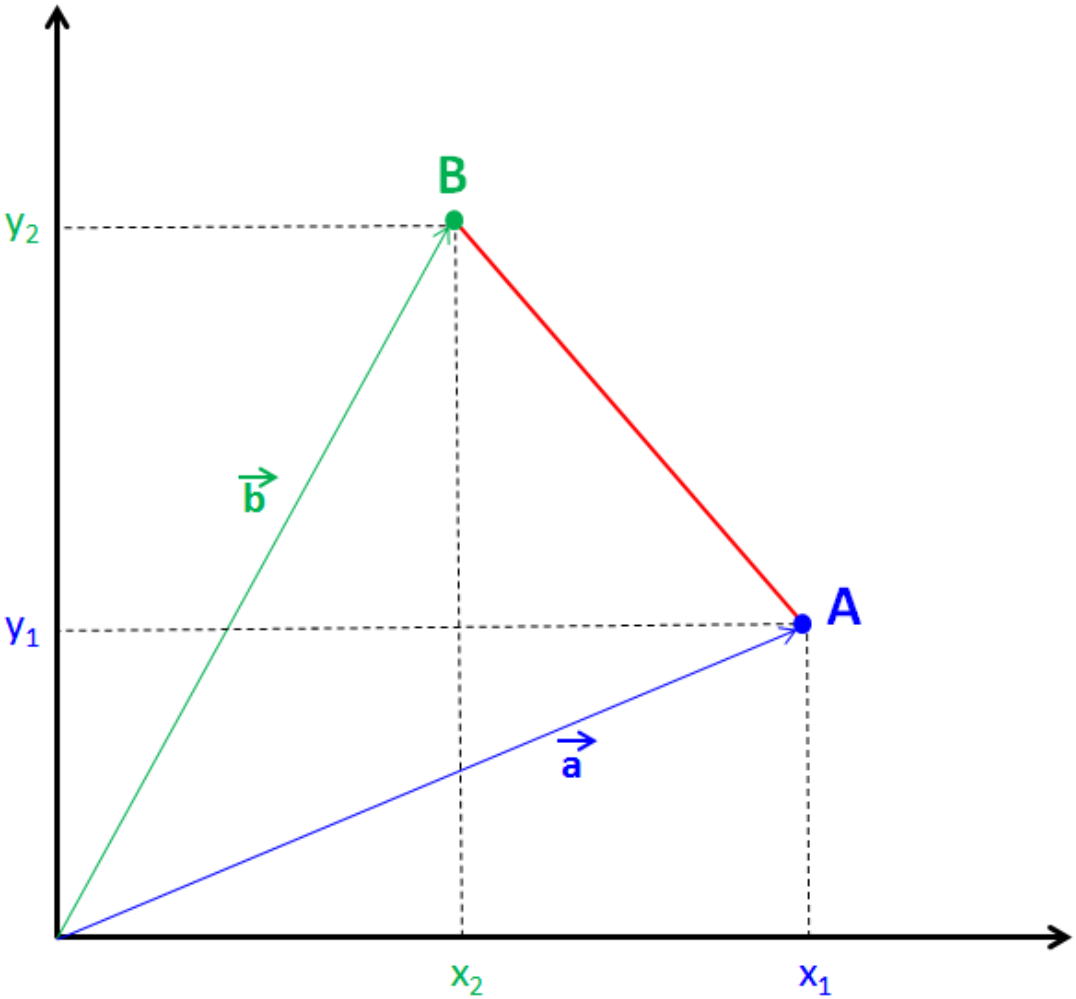
Abstand zwischen 2 Punkten



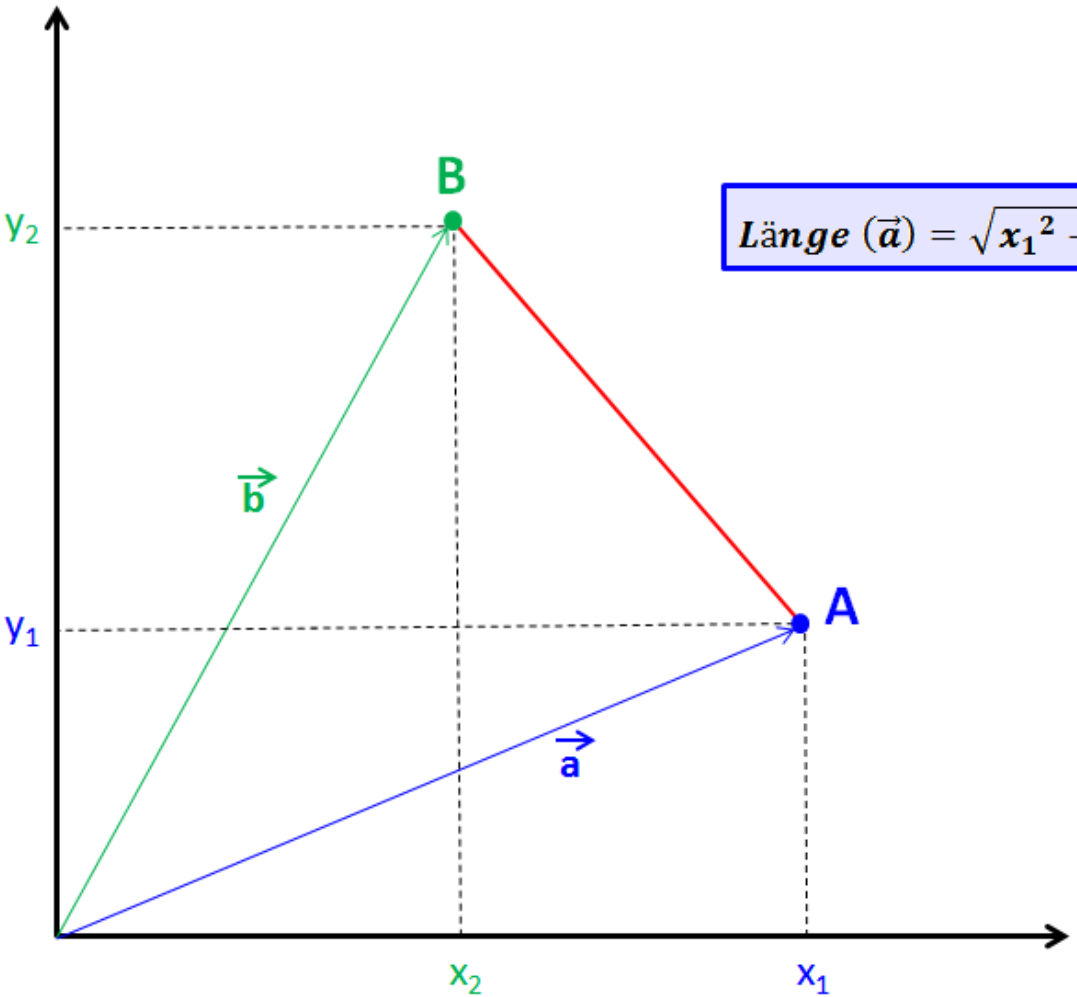
Abstand zwischen 2 Punkten



Abstand zwischen 2 Punkten

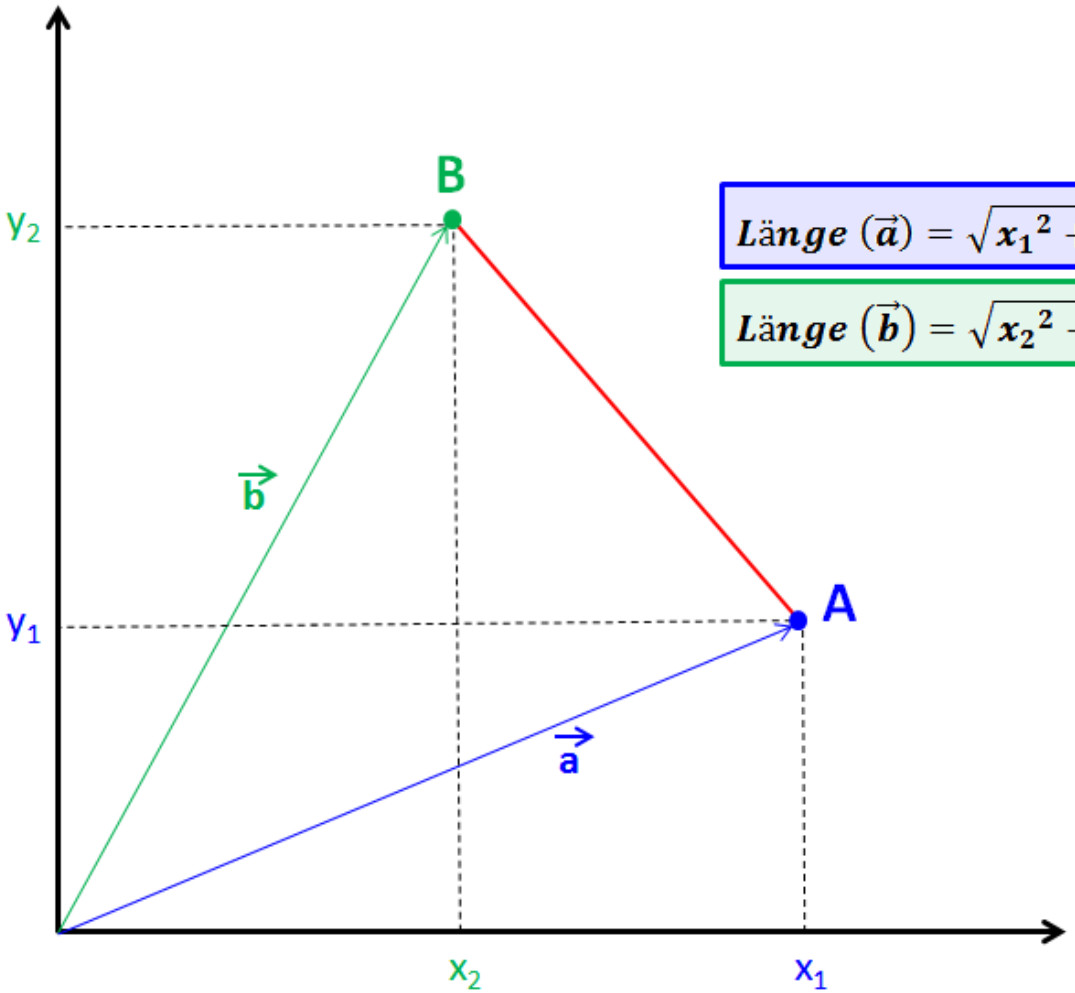


Abstand zwischen 2 Punkten



$$\text{Länge } (\vec{a}) = \sqrt{x_1{}^2 + y_1{}^2}$$

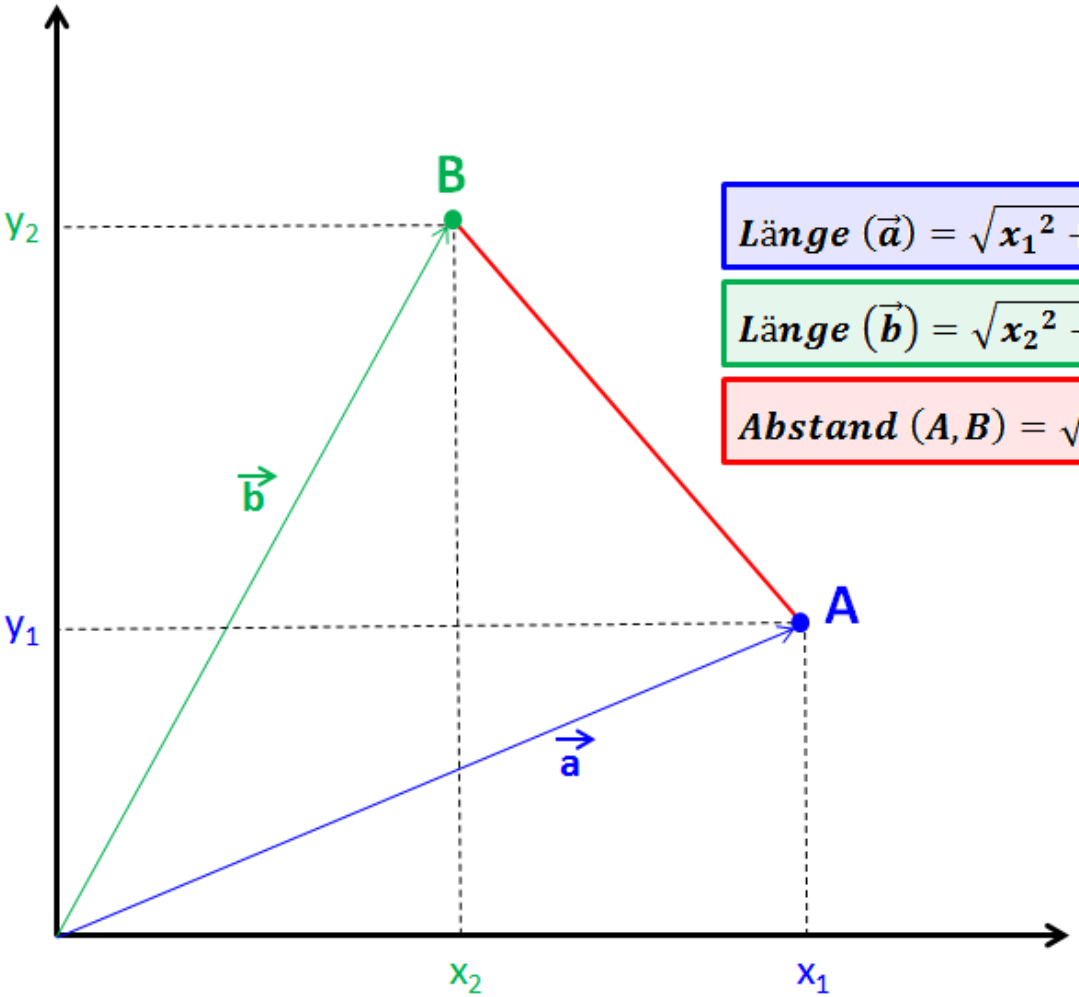
Abstand zwischen 2 Punkten



$$\text{Länge}\,(\vec{a}) = \sqrt{x_1{}^2 + y_1{}^2}$$

$$\text{Länge}\,(\vec{b}) = \sqrt{x_2{}^2 + y_2{}^2}$$

Abstand zwischen 2 Punkten



$$\textbf{Länge}\ (\vec{a}) = \sqrt{x_1{}^2 + y_1{}^2}$$

$$\textbf{Länge}\ (\vec{b}) = \sqrt{x_2{}^2 + y_2{}^2}$$

$$\textbf{Abstand}\ (A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

- **Nichtnegativität**: Für alle $v \in V$ gilt $\|v\| \geq 0$.

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

- **Nichtnegativität**: Für alle $v \in V$ gilt $\|v\| \geq 0$.
- **Definiertheit**: Für alle $v \in V$ gilt $\|v\| = 0 \Leftrightarrow v = 0$.

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

- **Nichtnegativität**: Für alle $v \in V$ gilt $\|v\| \geq 0$.
- **Definiertheit**: Für alle $v \in V$ gilt $\|v\| = 0 \Leftrightarrow v = 0$.
- **Homogenität**: Für alle $v \in V$ und alle $\alpha \in \mathbb{K}$ gilt $\|\alpha v\| = |\alpha|\|v\|$.

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

- **Nichtnegativität**: Für alle $v \in V$ gilt $\|v\| \geq 0$.
- **Definiertheit**: Für alle $v \in V$ gilt $\|v\| = 0 \Leftrightarrow v = 0$.
- **Homogenität**: Für alle $v \in V$ und alle $\alpha \in \mathbb{K}$ gilt $\|\alpha v\| = |\alpha|\|v\|$.
- **Dreiecksungleichung**: Für alle $v, w \in V$ gilt $\|v + w\| \leq \|v\| + \|w\|$.

Norm

Definition (Norm)

Sei V ein Vektorraum über $\mathbb{K}$. Eine Funktion

$$V \to \mathbb{R}, v \mapsto \|v\|$$

heißt Norm auf V, wenn sie die nachfolgenden Eigenschaften erfüllt:

- **Nichtnegativität**: Für alle $v \in V$ gilt $\|v\| \geq 0$.
- **Definiertheit**: Für alle $v \in V$ gilt $\|v\| = 0 \Leftrightarrow v = 0$.
- **Homogenität**: Für alle $v \in V$ und alle $\alpha \in \mathbb{K}$ gilt $\|\alpha v\| = |\alpha|\|v\|$.
- **Dreiecksungleichung**: Für alle $v, w \in V$ gilt $\|v + w\| \leq \|v\| + \|w\|$.

Die Norm wird vereinfachend durch $\|\cdot\|$ dargestellt:

$$\|\cdot\| : V \to \mathbb{R}$$

p-Norm

Definition (p-Norm)

Für jede natürliche Zahl $n \in \mathbb{N}$ und jede reelle Zahl $p \geq 1$ definiert man auf dem Vektorraum $\mathbb{R}^n$ die sogenannte p-Norm $\|\cdot\|_p : \mathbb{K}^n \to \mathbb{R}$ durch:

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_n|^p}$$

für alle $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{K}^n$.

p-Norm

Definition (p-Norm)

Für jede natürliche Zahl $n \in \mathbb{N}$ und jede reelle Zahl $p \geq 1$ definiert man auf dem Vektorraum $\mathbb{R}^n$ die sogenannte p-Norm $\|\cdot\|_p : \mathbb{K}^n \to \mathbb{R}$ durch:

$$\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} = \sqrt[p]{|x_1|^p + |x_2|^p + \cdots + |x_n|^p}$$

für alle $x = (x_1, x_2, \ldots, x_n)^T \in \mathbb{K}^n$.

Für $p = 2$ entspricht die p-Norm genau der **euklidischen Norm** auf $\mathbb{R}^n$

Euklidische Norm

Definition (Euklidische Norm)

Die euklidische Norm entspricht der Wurzel der Summe der Betragsquadrate der Komponenten des Vektors:

$$\|x\|_2 := \sqrt{\sum_{i=1}^{n} |x_i|^2}$$

Euklidische Distanz

in 2 oder 3 Dimensionen beschreibt die euklidische Norm die Länge eines Vektors in der Ebene oder im Raum.
Für p metrische Variablen ist die **Euklidische Distanz** definiert als:

$$\sqrt[p]{\sum_{i=1}^{n} |x_{ik} - x_{ij}|^p}$$

Definition K-Means von Folie 5:

$$\begin{aligned} J(c_j) &= \sum_{a_i \in c_j} \|a_i - c_j\|^2 \\ &= \sum_{a_i \in c_j} \left(\sqrt{\sum_{i=1}^{n} |a_i - c_j|^2} \right)^2 \\ &= \sum_{a_i \in c_j} |a_i - c_j|^2 = \sum_{a_i \in c_j} d^2(a_i, c_j) \end{aligned}$$