Gruppetime 12.02

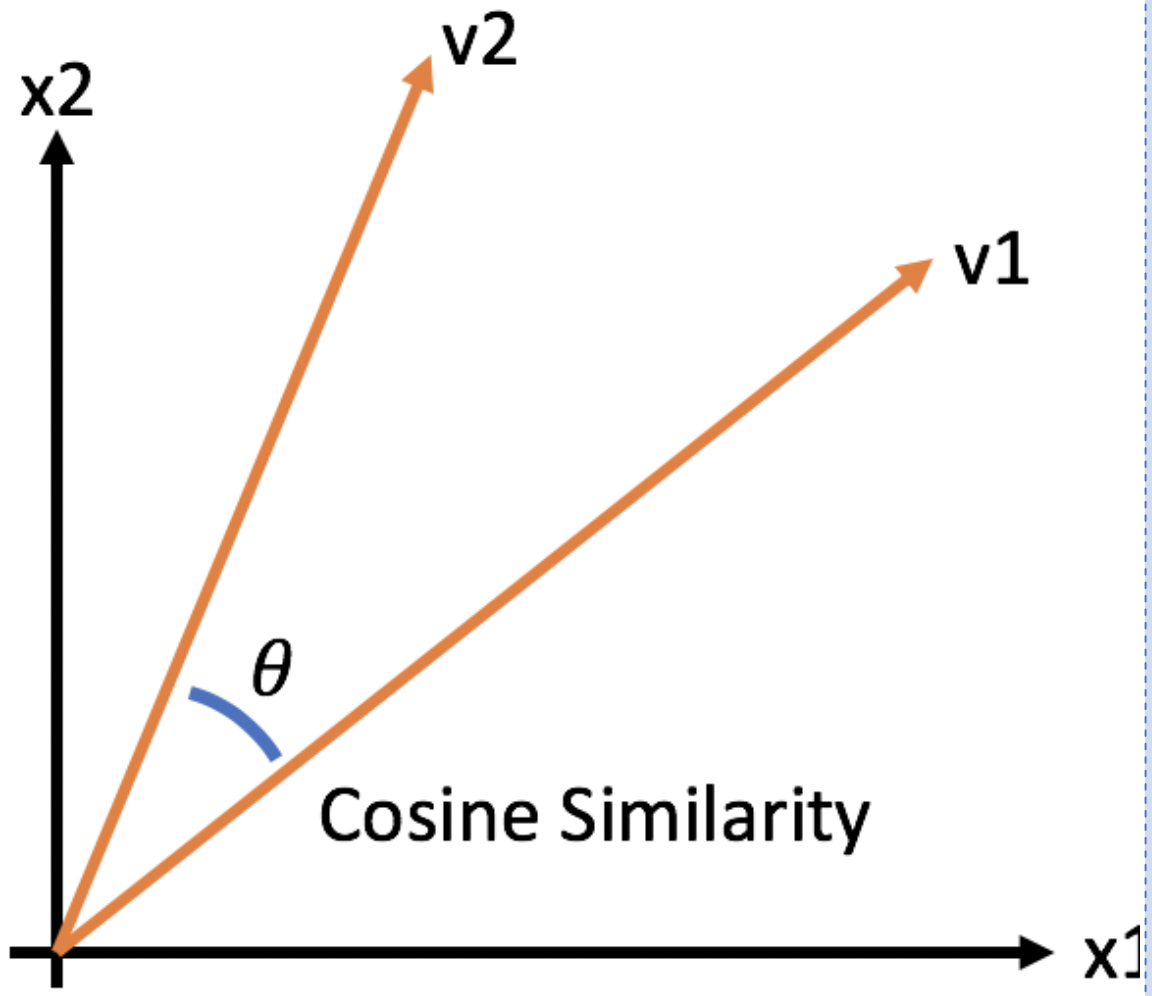
Agenda

- Cosine similarity & Euclidean distance
- K-Nearest Neighbor (KNN) classification
- Evaluation measures

Cosine similarity

- Måles mellom to vektorer
- Similarity vs. Distance
- Lengdenormaliser

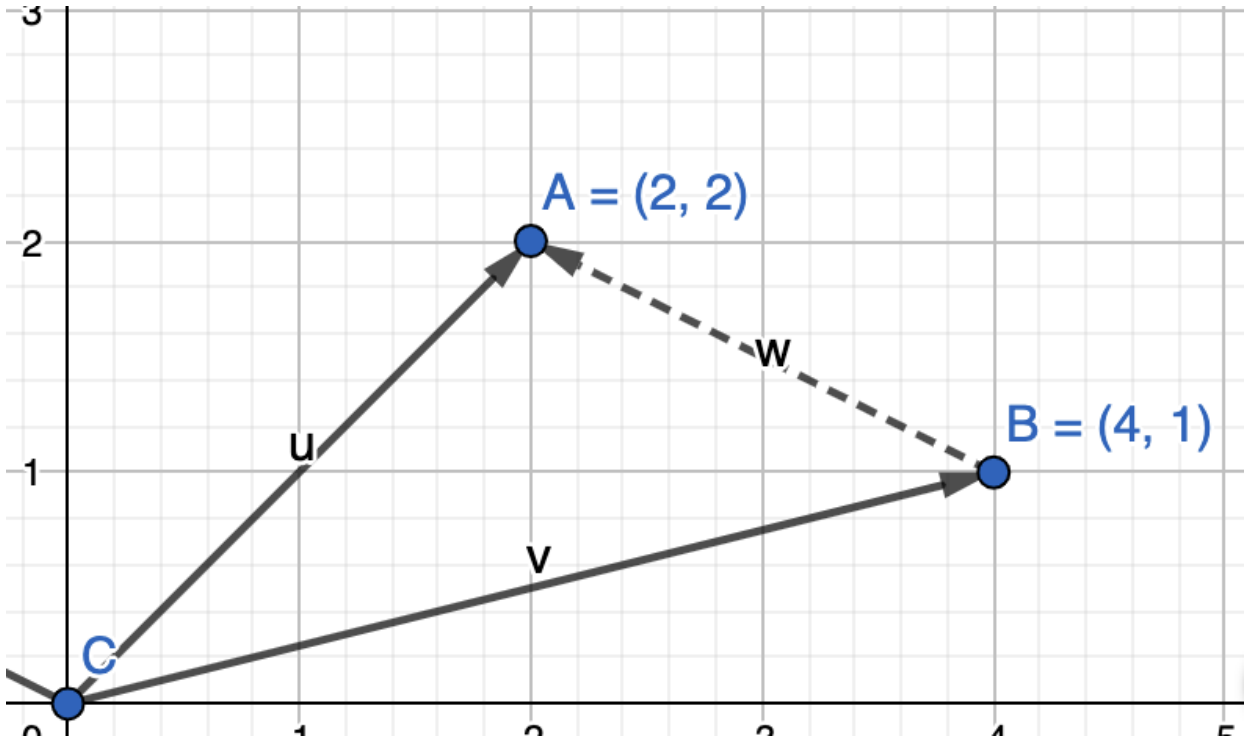
$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{b}_i$$



Euclidean distance

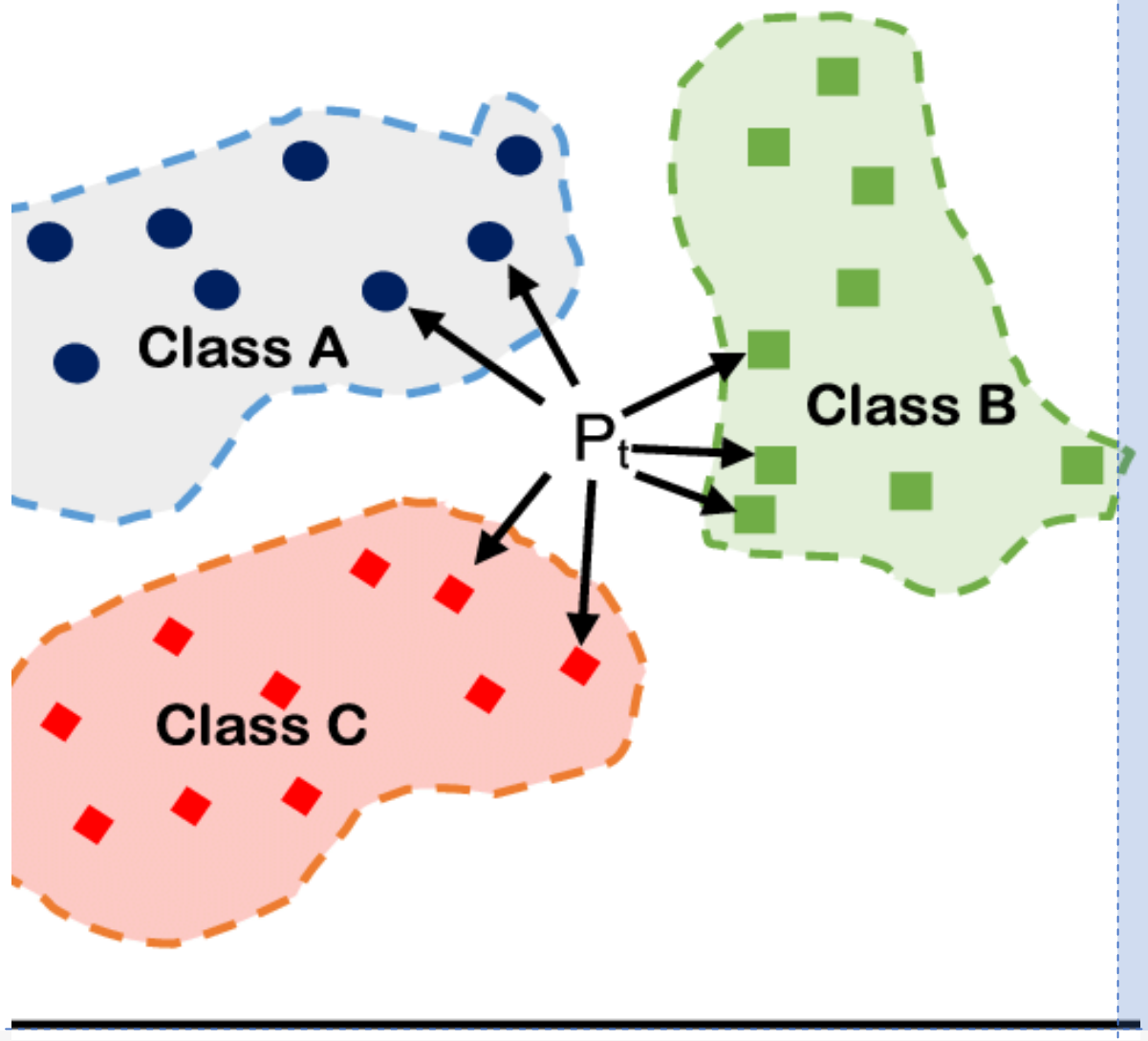
- Måles mellom ytterpunktet til to vektorer
- Lengdenormaliser

$$d(\boldsymbol{a}, \boldsymbol{b}) = \sqrt{\sum_{i=1}^{n} (\boldsymbol{a}_i - \boldsymbol{b}_i)^2}$$



K-Nearest Neighbor (KNN) classification

- *Veiledet læring*
- *Contiguity hypothesis*
- *Voronoi-tesselering*
- *K må være oddetall*



Evaluation measures

- Accuracy $= \frac{TP+TN}{N} = \frac{TP+TN}{TP+TN+FP+FN}$
  - The ratio of correct predictions.
  - Not suitable for unbalanced numbers of positive / negative examples.
- Precision $= \frac{TP}{TP+FP}$
  - The number of detected class members that were correct.
- Recall $= \frac{TP}{TP+FN}$
  - The number of actual class members that were detected.
  - Trade-off: Positive predictions for all examples would give 100% recall but (typically) terrible precision.
- F-score $= 2 \times \frac{precision \times recall}{precision + recall}$
  - Balanced measure of precision and recall (harmonic mean).