Data Mining and Machine Learning: Techniques and Algorithms

Eneldo Loza Mencía

eneldo@ke.tu-darmstadt.de

Knowledge Engineering Group, TU Darmstadt

International Week 2019, 21.1. – 24.1.
University of Economics, Prague

Outline

- Preprocessing
  - Vector space model
  - Text preprocessing pipeline
  - Similarity of Documents
- Text Classification Algorithms
  - Rocchio Classifer
  - Naïve Bayes classifier
  - Linear classification
  - Support Vector Machines
- Occam's Razor and Overfitting Avoidance

Text Classification: Examples

Text Categorization: Assign (class) labels to each document

- Labels are most often topics such as Yahoo-categories
  - *e.g., "finance," "sports," "news::world::asia::business"*
- Labels may be genres
  - *e.g., "editorials" "movie-reviews" "news“*
- Labels may be opinion
  - e.g., *“like”, “hate”, “neutral”*
- Labels may be binary concepts
  - *e.g., "interesting-to-me" : "not-interesting-to-me”*
  - e.g., *“spam” : “not-spam”*
  - e.g., *“contains adult language” :“doesn’t”*

More than one learning task could be defined over the same documents

Manning and Raghavan

News categorization

```xml
<?xml version="1.0" encoding="iso-8859-1" ?>
<newsitem itemid="477551" id="root" date="1997-03-31" xml:lang="en">
<title>SPAIN: Spain's Banesto issue $150 mln in subordinated loaN.</title>
<headline>Spain's Banesto issue $150 mln in subordinated loaN.</headline>
<dateline>MADRID 1997-03-31</dateline>
<text>
<p>Banco Espanol de Credito Banesto said on Monday it issued $150 million in subordinated 10-year 7.5 percent debt. Lead manager is Lehman Brothers.</p>
<p>The statement added that this is the first international issue Banesto has launched since 1993.</p>
<p>Banco Santander has a 50 percent stake in Banesto.</p>
<p>– Madrid Newsroom, + 341 585 8340</p>
</text>
<code code="C17">
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
<code code="C172">
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
<code code="CCAT">
  <editdetail attribution="Reuters BIP Coding Group" action="confirmed" date="1997-03-31"/>
</code>
```

Funding/Capital

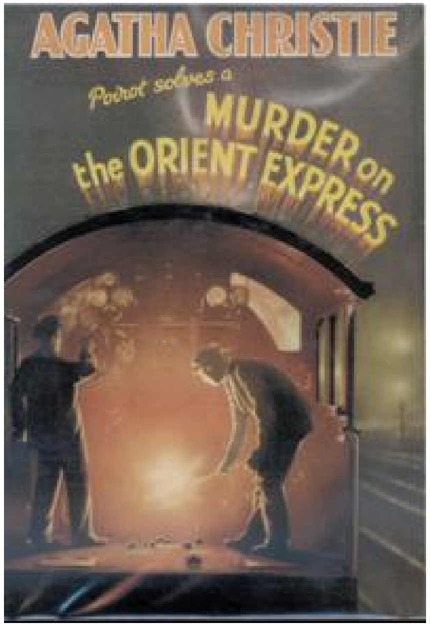
Bonds/Debt issues

Corporate/Industrial

The Reuters RCV1 dataset has in total 103 assignable news categories for 804.414 news articles

David Dolan LEWIS, Yiming YANG, Tony G. ROSE, Fan LI: RCV1: A New Benchmark Collection for Text Categorization Research. In: Journal of Machine Learning Research, 2004.

Book Scenario



Summary: Returning from an important case in Syria, Hercule Poirot boards the Orient Express in Istanbul. The train is unusually crowded for the time of year. Poirot secures a berth only with ...

Text: It was five o'clock on a winter's morning in Syria. … "Then," said Poirot, "having placed my solution before you, I have the honour to retire from the case."

Author:
Agatha Christie

Genres:
Crime, Mystery, Thriller

Subjects (LOC):
Private Investigators, Orient Express, ...

Keywords:
mystery, fiction, crime, murder, british, poirot, ...

Rate:
4 of 5 stars

Epoch:
1930ies

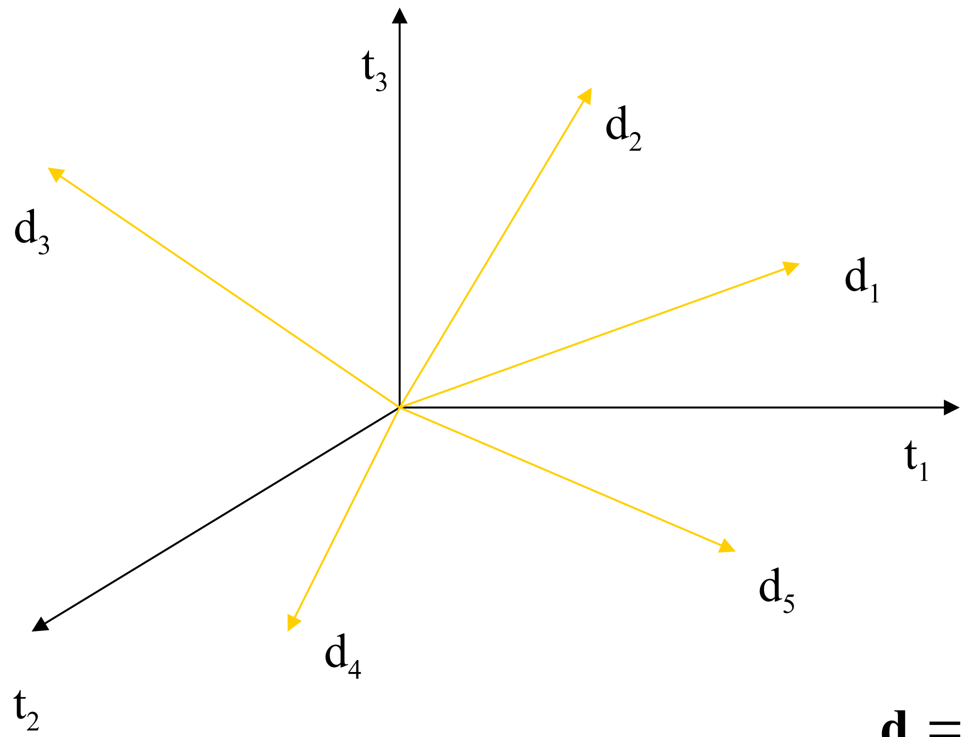
Country:
UK

...

The Vector Space Model

- Origin:
  Information Retrieval, SMART system (Salton et al.)
- Basic idea:
  - A document is regarded as a vector in an *n*-dimensional space
    - 1 dimension for each possible word (*feature, token*)
    - the value in each dimension is (in the simplest case) the number of times the word occurs in the document (*term frequency – TF*)
  - a document is a linear combination of the base vectors
  - linear algebra can be used for various computations

Intuition



$$\mathbf{d}_i = (d_{i,1}, d_{i,2}, d_{i,3})$$

Postulate: Documents that are “close together” in the vector space talk about the same things.

Manning and Raghavan

Document Representation

TECHNISCHE UNIVERSITÄT DARMSTADT

- The vector space models allows to transform a text into a document-term table

- In the simplest case
  - Rows:
    - training documents
  - Columns:
    - words in the training documents
  - More complex representation possible

- Most machine learning and data mining algorithms need this type of representation
  - they can now be applied to, e.g., text classification

Document Representation


[CS414: Systems Programming and Operating Systems]

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule ***(Last Changed: 9/14/95)***

Groups

Handouts

- **Handout 1**
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers ***(Last Changed: 10/23/95)***

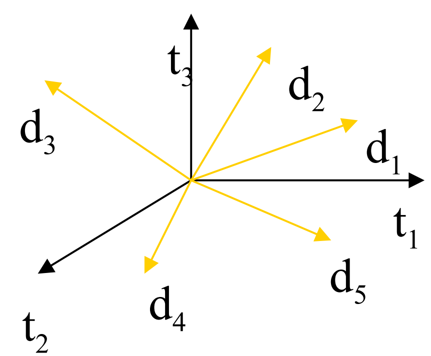
The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

	baseball	specs	graphics	....	quicktime	computer
D1	0	3	0	....	2	0
D2	1	2	0	...	0	0
D3	0	0	2	...	1	5
.....	....	....	....	....	....	....



Text Preprocessing Pipeline
Tokenization

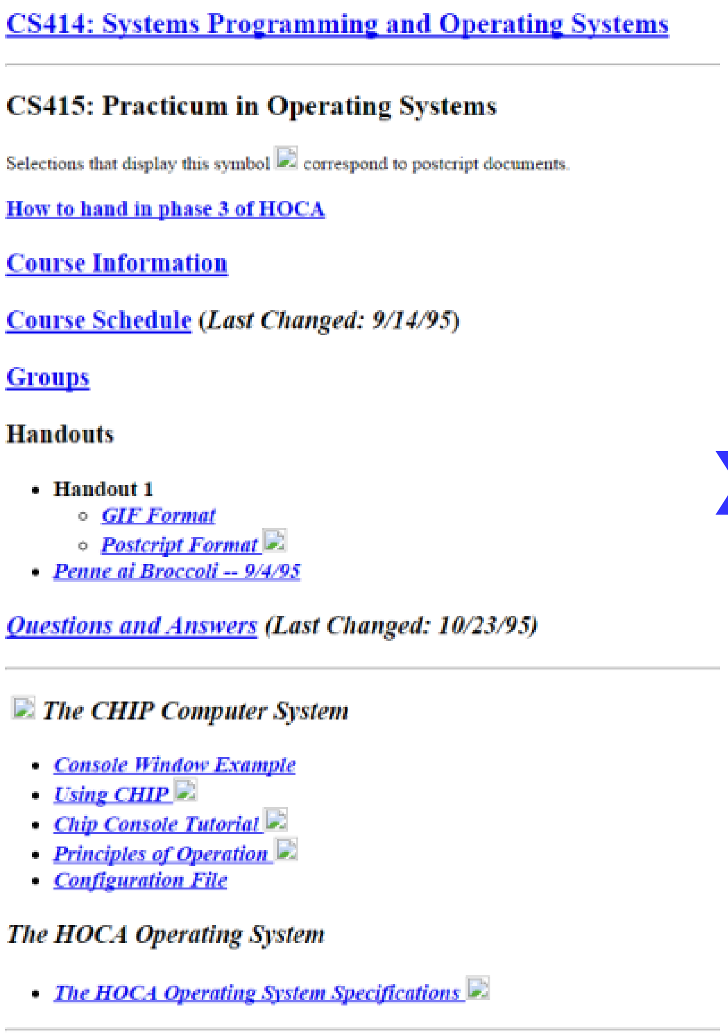
- Identification of basic document entities („words“)
  - typically performed in indexing phase
- Issues in tokenization:
  - ***Finland’s capital*** →

    Finland? Finlands? Finland’s?
  - ***Hewlett-Packard*** → ***Hewlett*** and ***Packard*** as two tokens?
    - ***State-of-the-art***: break up hyphenated sequence.
    - ***co-education*** ?
    - ***the hold-him-back-and-drag-him-away-maneuver*** ?
    - It’s effective to get the user to put in possible hyphens
  - ***San Francisco***: one token or two? How do you decide it is one token?

Manning and Raghavan

Text Preprocessing Pipeline

TECHNISCHE UNIVERSITÄT DARMSTADT



CS414: Systems Programming and Operating Systems

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule (*Last Changed: 9/14/95*)

Groups

Handouts

- Handout 1
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers (*Last Changed: 10/23/95*)

The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

cs415, home, page, cs414, systems, programming, and, operating, systems, cs415, practicum, in, operating, systems, selections, that, display, this, symbol, correspond, to, postcript, documents, how, to, hand, in, phase, 3, of, hoca, course, information, course, schedule, last, changed, 9, 14, 95, groups, handouts, handout, 1, gif, format, postcript, format, penne, ai, broccoli, 9, 4, 95, questions, and, answers, last, changed, 10, 23, 95, the, chip, computer, system, console, window, example, using, chip, chip, console, tutorial, principles, of, operation, configuration, file, the, hoca, operating, system, the, hoca, operating, system, specifications, this, page, is, maintained, by, lorenzo, alvisi

Text Preprocessing Pipeline
Stemming

TECHNISCHE UNIVERSITÄT DARMSTADT

- Reduce terms to their “roots” before indexing

- “Stemming” suggest crude affix chopping
  - language dependent
  - e.g., ***automate(s), automatic, automation*** all reduced to ***automat***.
- Stemming may reduce number of terms by ~35%

for example compressed and compression are both accepted as equivalent to compress.

→

for exampl compress and compress ar both accept as equival to compress

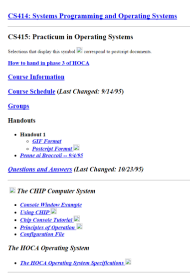
Manning and Raghavan

Text Preprocessing Pipeline
Stop Words

TECHNISCHE UNIVERSITÄT DARMSTADT

- Remove most frequent words in the (English) language
  - a, about, above, across, after, afterwards, again, against, all, almost, alone, along, already, also, although, always, am, ... yet, you, your, yours, yourself, yourselves
- Assumption:
  - These words occur in all documents and are irrelevant for retrieval
- Rule of 30: ~30 words account for ~30% of all term occurrences in written text
- Stop lists used to be popular, but are sometimes avoided, because important information may be lost
  - polysemous words: „can“ as a verb vs. „can“ as a noun
  - phrases: “Let it be”, “To be or not to be”, pop group „The The“
  - relations: “flights to London” vs. „flights from London“

Text Preprocessing Pipeline


CS414: Systems Programming and Operating Systems

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule (*Last Changed: 9/14/95*)

Groups

Handouts

- Handout 1
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers (Last Changed: 10/23/95)

The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

cs415, home, page, cs414, systems, programming, and, operating, systems, cs415, practicum, in, operating, systems, selections, that, display, this, symbol, correspond, to, postcript, documents, how, to, hand, in, phase, 3, of, hoca, course, information, course, schedule, last, changed, 9, 14, 95, groups, handouts, handout, 1, gif, format, postcript, format, penne, ai, broccoli, 9, 4, 95, questions, and, answers, last, changed, 10, 23, 95, the, chip, computer, system, console, window, example, using, chip, chip, console, tutorial, principles, of, operation, configuration, file, the, hoca, operating, system, the, hoca, operating, system, specifications, this, page, is, maintained, by, lorenzo, alvisi

cs415, home, page, cs414, system, program, oper, system, cs415, practicum, oper, system, select, displai, symbol, correspond, postcript, document, hand, phase, 3, hoca, cours, inform, cours, schedul, last, chang, 9, 14, 95, group, handout, handout, 1, gif, format, postcript, format, penn, ai, broccoli, 9, 4, 95, question, answer, last, chang, 10, 23, 95, chip, comput, system, consol, window, exampl, us, chip, chip, consol, tutori, principl, oper, configur, file, hoca, oper, system, hoca, oper, system, specif, page, maintain, lorenzo, alvisi

Text Preprocessing Pipeline
Term Weighting

TECHNISCHE UNIVERSITÄT DARMSTADT

Different ways for computing the $d_{i,j}$:

- Boolean
  - possible values are only
    - 0 (term does not occur in document)
    - 1 (term does occur)

$$d_{i,j} = \begin{cases} 0 & \text{if } t_j \notin \mathbf{d}_i \\ 1 & \text{if } t_j \in \mathbf{d}_i \end{cases}$$

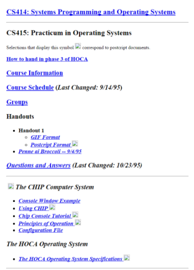
- Term Frequency (TF)
  - term is weighted with the frequency of its occurrence in the text

$$d_{i,j} = TF(\mathbf{d}_i, t_j)$$

- Term Frequency - Inverse Document Frequency (TF-IDF)
  - Idea: A term is characteristic for a document if
    - it occurs frequently in this document (TF)
    - occurs infrequently in other documents (IDF)
  - divides TF by DF
    (or multiplies TF with IDF)

$$d_{i,j} = \frac{TF(\mathbf{d}_i, t_j)}{DF(t_j)} = TF(\mathbf{d}_i, t_j) \cdot IDF(t_j)$$

Text Preprocessing Pipeline


CS414: Systems Programming and Operating Systems

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule (*Last Changed: 9/14/95*)

Groups

Handouts

- Handout 1
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers (*Last Changed: 10/23/95*)

The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

cs415, home, page, cs414, systems, programming, and, operating, systems, cs415, practicum, in, operating, systems, selections, that, display, this, symbol, correspond, to, postcript, documents, how, to, hand, in, phase, 3, of, hoca, course, information, course, schedule, last, changed, 9, 14, 95, groups, handouts, handout, 1, gif, format, postcript, format, penne, ai, broccoli, 9, 4, 95, questions, and, answers, last, changed, 10, 23, 95, the, chip, computer, system, console, window, example, using, chip, chip, console, tutorial, principles, of, operation, configuration, file, the, hoca, operating, system, the, hoca, operating, system, specifications, this, page, is, maintained, by, lorenzo, alvisi

cs415, home, page, cs414, system, program, oper, system, cs415, practicum, oper, system, select, displai, symbol, correspond, postcript, document, hand, phase, 3, hoca, cours, inform, cours, schedul, last, chang, 9, 14, 95, group, handout, handout, 1, gif, format, postcript, format, penn, ai, broccoli, 9, 4, 95, question, answer, last, chang, 10, 23, 95, chip, comput, system, consol, window, exampl, us, chip, chip, consol, tutori, principl, oper, configur, file, hoca, oper, system, hoca, oper, system, specif, page, maintain, lorenzo, alvisi

0.0, 0.0, 0.00357498983914, 0.00767062752856, 0.00548167648169, 0.0, 0.0, 0.0, 0.0, 0.0104392661432, 0.0, 0.068107989169, 0.0114700408283, 0.0, 0.0, 0.012853987485, 0.0, 0.0280722363513, 0.0, 0.014182639917, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0320005481006, 0.0, 0.0, 0.0, 0.0, 0.0171544554818, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0198307992493, 0.0, 0.0, 0.0, 0.0, 0.0202970510976, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0222141246101, 0.0, 0.0, 0.0, 0.112464211937, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0461382844637, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0244691331545, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …

Text Preprocessing Pipeline
Feature Subset Selection

- Using each word as a feature results in tens, hundreds, or thousands of thousands of features
- Many of them are
  - irrelevant
  - redundant
- Removing them can
  - increase efficiency
  - prevent overfitting
- Feature Subsect Selection techniques try to determine appropriate features automatically

Text Preprocessing Pipeline
Feature Subset Selection

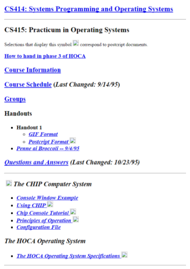
Unsupervised Feature Subset Selection

- Using domain knowledge
  - some features may be known to be irrelevant, uninteresting or redundant
- Frequency-based selection
  - select features based on statistical properties
  - e.g. IDF: hypothesis that terms with high document frequency are more important (except stop words)

Supervised Feature Subset Selection

- Filter approaches
  - compute some measure (e.g. statistical) for estimating the ability to discriminate between classes
- Wrapper approaches
  - each search subset is tried with the learning algorithm

Text Preprocessing Pipeline



CS414: Systems Programming and Operating Systems

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule (*Last Changed: 9/14/95*)

Groups

Handouts

- Handout 1
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers (*Last Changed: 10/23/95*)

The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

cs415, home, page, cs414, systems, programming, and, operating, systems, cs415, practicum, in, operating, systems, selections, that, display, this, symbol, correspond, to, postcript, documents, how, to, hand, in, phase, 3, of, hoca, course, information, course, schedule, last, changed, 9, 14, 95, groups, handouts, handout, 1, gif, format, postcript, format, penne, ai, broccoli, 9, 4, 95, questions, and, answers, last, changed, 10, 23, 95, the, chip, computer, system, console, window, example, using, chip, chip, console, tutorial, principles, of, operation, configuration, file, the, hoca, operating, system, the, hoca, operating, system, specifications, this, page, is, maintained, by, lorenzo, alvisi

cs415, home, page, cs414, system, program, oper, system, cs415, practicum, oper, system, select, displai, symbol, correspond, postcript, document, hand, phase, 3, hoca, cours, inform, cours, schedul, last, chang, 9, 14, 95, group, handout, handout, 1, gif, format, postcript, format, penn, ai, broccoli, 9, 4, 95, question, answer, last, chang, 10, 23, 95, chip, comput, system, consol, window, exampl, us, chip, chip, consol, tutori, principl, oper, configur, file, hoca, oper, system, hoca, oper, system, specif, page, maintain, lorenzo, alvisi

0.0, 0.0, 0.00357498983914, 0.00767062752856, 0.00548167648169, 0.0, 0.0, 0.0, 0.0, 0.0104392661432, 0.0, 0.068107989169, 0.0114700408283, 0.0, 0.0, 0.012853987485, 0.0, 0.0280722363513, 0.0, 0.014182639917, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0320005481006, 0.0, 0.0, 0.0, 0.0, 0.0171544554818, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0198307992493, 0.0, 0.0, 0.0, 0.0, 0.0202970510976, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0222141246101, 0.0, 0.0, 0.0, 0.112464211937, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0461382844637, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0244691331545, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …

0.0, 0.0, 0.00357498983914, 0.00767062752856, 0.00548167648169, 0.0, 0.0, 0.0, 0.0, 0.0104392661432

Text Preprocessing Pipeline

TECHNISCHE UNIVERSITÄT DARMSTADT

CS414: Systems Programming and Operating Systems

CS415: Practicum in Operating Systems

Selections that display this symbol correspond to postcript documents.

How to hand in phase 3 of HOCA

Course Information

Course Schedule (*Last Changed: 9/14/95*)

Groups

Handouts

- **Handout 1**
  - *GIF Format*
  - *Postcript Format*
- *Penne ai Broccoli -- 9/4/95*

Questions and Answers (*Last Changed: 10/23/95*)

The CHIP Computer System

- *Console Window Example*
- *Using CHIP*
- *Chip Console Tutorial*
- *Principles of Operation*
- *Configuration File*

The HOCA Operating System

- *The HOCA Operating System Specifications*

cs415, home, page, cs414, systems, programming, and, operating, systems, cs415, practicum, in, operating, systems, selections, that, display, this, symbol, correspond, to, postcript, documents, how, to, hand, in, phase, 3, of, hoca, course, information, course, schedule, last, changed, 9, 14, 95, groups, handouts, handout, 1, gif, format, postcript, format, penne, ai, broccoli, 9, 4, 95, questions, and, answers, last, changed, 10, 23, 95, the, chip, computer, system, console, window, example, using, chip, chip, console, tutorial, principles, of, operation, configuration, file, the, hoca, operating, system, the, hoca, operating, system, specifications, this, page, is, maintained, by, lorenzo, alvisi

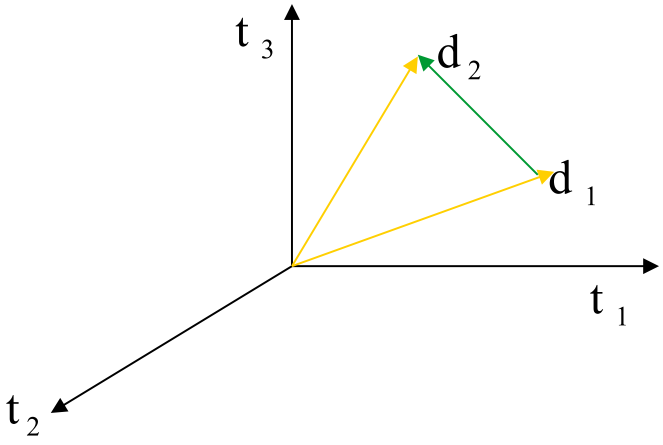
cs415, home, page, cs414, system, program, oper, system, cs415, practicum, oper, system, select, displai, symbol, correspond, postcript, document, hand, phase, 3, hoca, cours, inform, cours, schedul, last, chang, 9, 14, 95, group, handout, handout, 1, gif, format, postcript, format, penn, ai, broccoli, 9, 4, 95, question, answer, last, chang, 10, 23, 95, chip, comput, system, consol, window, exampl, us, chip, chip, consol, tutori, principl, oper, configur, file, hoca, oper, system, hoca, oper, system, specif, page, maintain, lorenzo, alvisi

0.0, 0.0, 0.00357498983914, 0.00767062752856, 0.00548167648169, 0.0, 0.0, 0.0, 0.0, 0.0104392661432, 0.0, 0.068107989169, 0.0114700408283, 0.0, 0.0, 0.012853987485, 0.0, 0.0280722363513, 0.0, 0.014182639917, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0320005481006, 0.0, 0.0, 0.0, 0.0, 0.0171544554818, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0198307992493, 0.0, 0.0, 0.0, 0.0, 0.0202970510976, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0222141246101, 0.0, 0.0, 0.0, 0.112464211937, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0461382844637, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0244691331545, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, …

0.0, 0.0, 0.00357498983914, 0.00767062752856, 0.00548167648169, 0.0, 0.0, 0.0, 0.0, 0.0104392661432

Similarity of Document Vectors

- First Idea:
  - Distance between $\mathbf{d}_1$ and $\mathbf{d}_2$ is the length of the vector $|\mathbf{d}_1 - \mathbf{d}_2|$ (measured with Euclidean distance)



- Why is this not a great idea?
  - Short documents would be more similar to each other by virtue of length, not topic

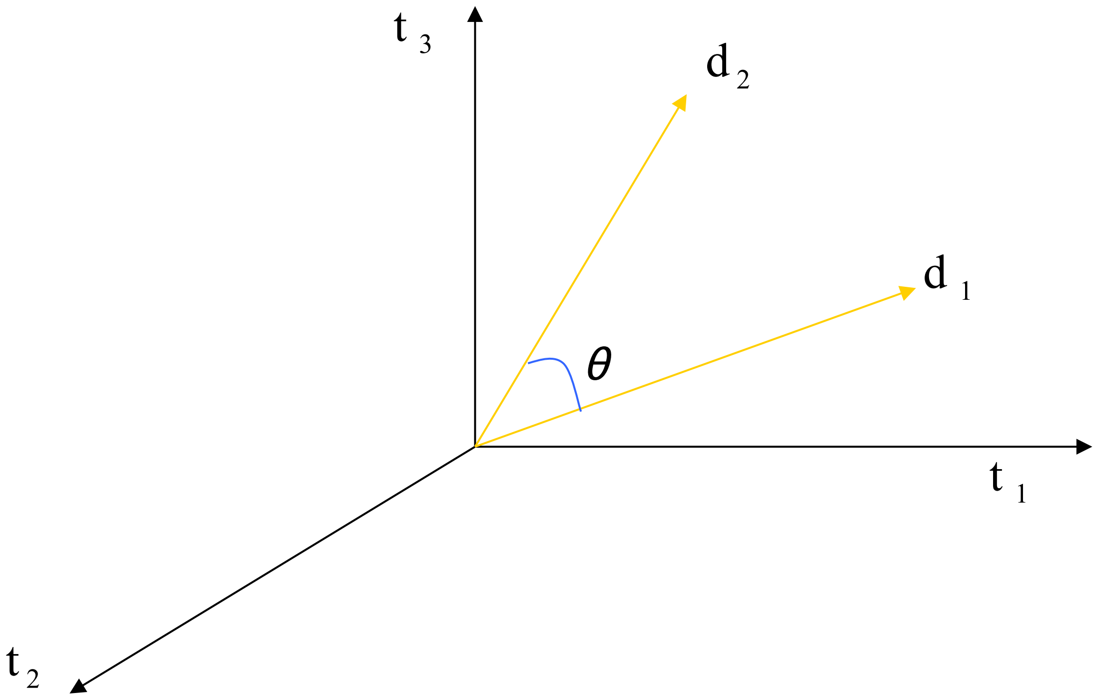
  → We have to deal with the issue of length normalization
  - explicit normalization (as, e.g., through normalized TF)

- Alternative approaches?
  - We can also implicitly normalize by looking at *angles* between document vectors instead

Manning and Raghavan

Cosine similarity

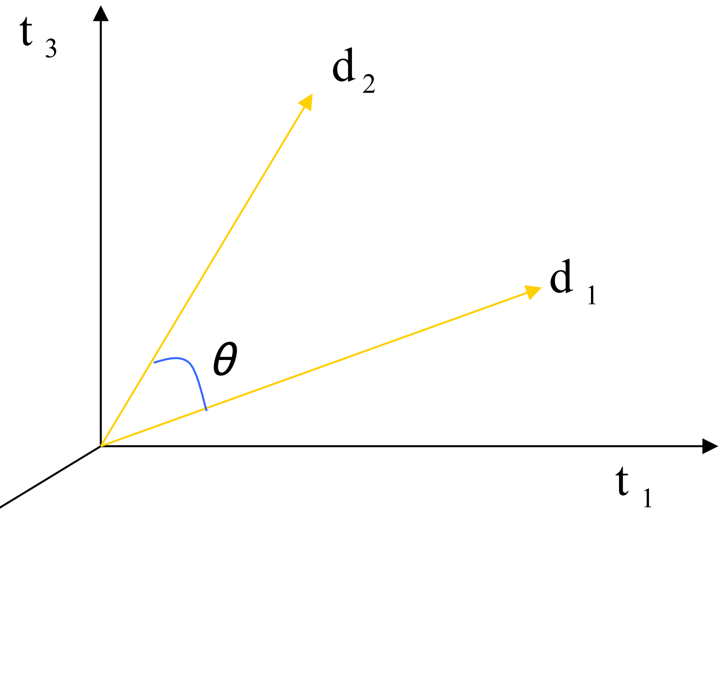
- Distance between vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ *captured* by the cosine of the angle θ between them.



Cosine similarity

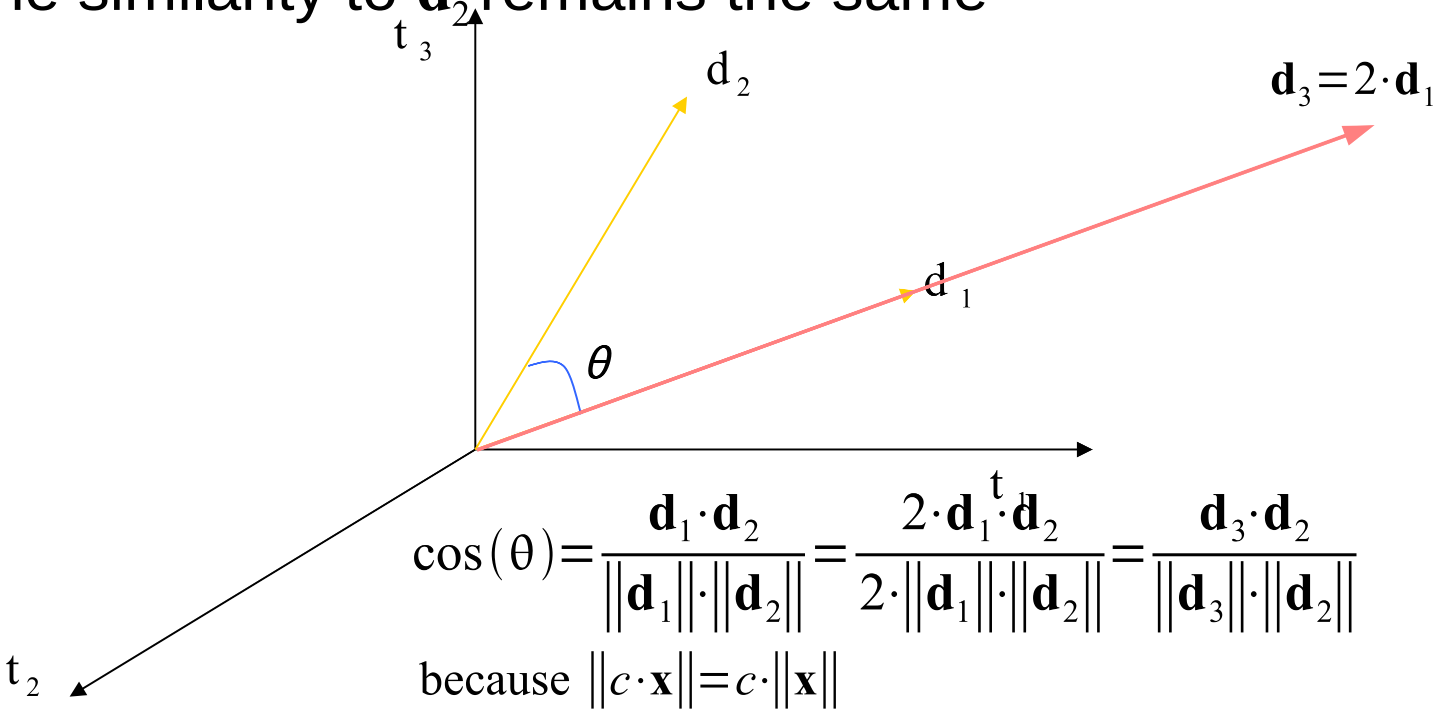
- Distance between vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ *captured* by the cosine of the angle θ between them.

$$\cos(\theta) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|} = \frac{\sum_{i=1}^{n} d_{1,i} d_{2,i}}{\sqrt{\sum_{i=1}^{n} d_{1,i}^2} \cdot \sqrt{\sum_{i=1}^{n} d_{2,i}^2}}$$



Cosine similarity

- Distance between vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ *captured* by the cosine of the angle θ between them.
- the distance is invariant to re-scaling the vector
  - e.g., if two copies of document $\mathbf{d}_1$ are concatenated to a new document $\mathbf{d}_3$, the similarity to $\mathbf{d}_2$ remains the same



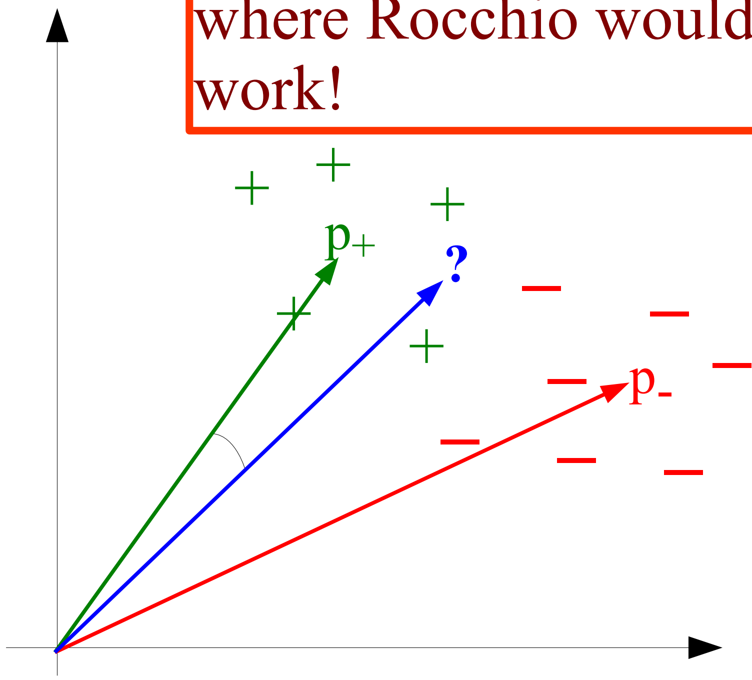
$$\cos(\theta) = \frac{\mathbf{d}_1 \cdot \mathbf{d}_2}{\|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|} = \frac{2 \cdot \mathbf{d}_1 \cdot \mathbf{d}_2}{2 \cdot \|\mathbf{d}_1\| \cdot \|\mathbf{d}_2\|} = \frac{\mathbf{d}_3 \cdot \mathbf{d}_2}{\|\mathbf{d}_3\| \cdot \|\mathbf{d}_2\|}$$

because $\|c \cdot \mathbf{x}\| = c \cdot \|\mathbf{x}\|$

Rocchio Classifier (Nearest Centroid Classifier)

- based on ideas for Rocchio Relevance Feedback
- compute a prototype vector $\mathbf{p}_c$ for each class c
  - average the document vectors for each class
- classify a new document according to distance to prototype vectors instead of documents
- assumption:
  - documents that belong to the same class are close to each other (form one cluster)

Q: Imagine simple scenarios where Rocchio would not work!

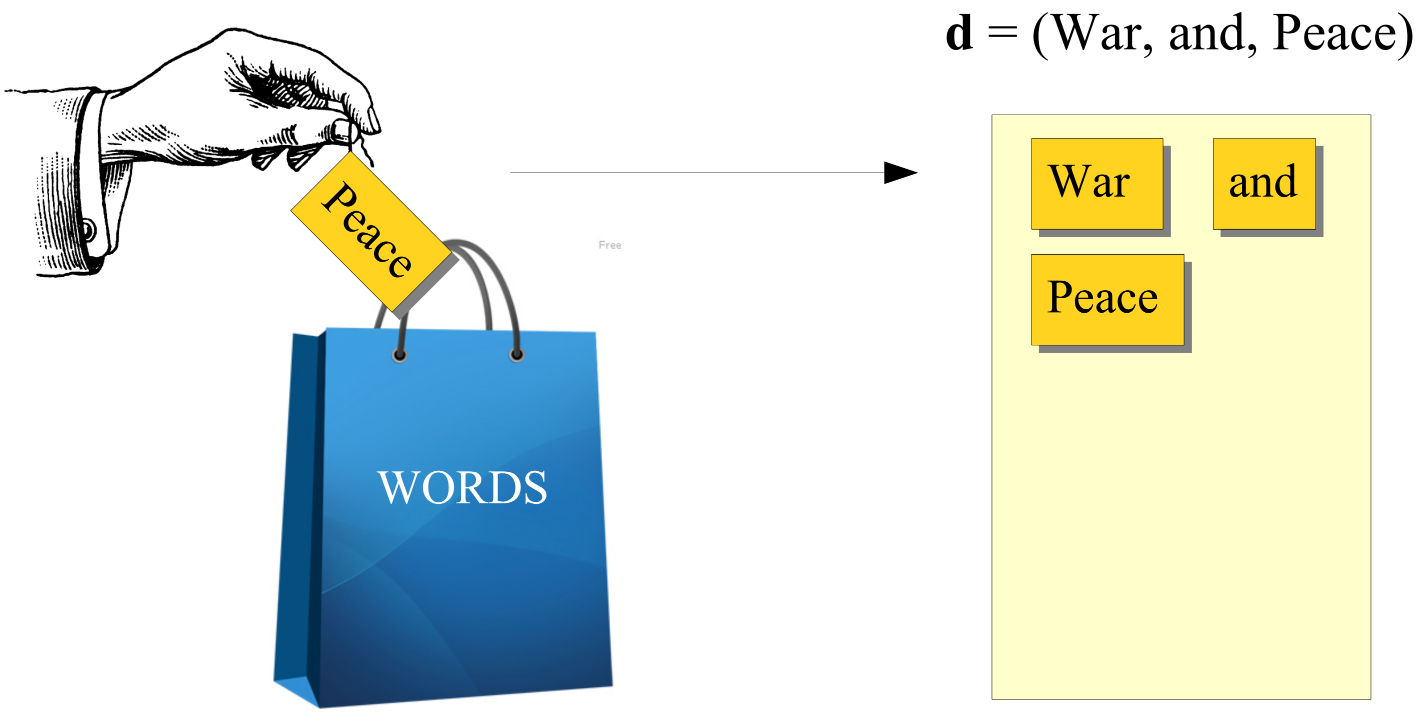


Bag of Words Model

- assumes that the document has been generated by repeatedly drawing one word out of a bag of words
  - like drawing letters out of a Scrabble-bag, but with replacement
- words in the bag may occur multiple times, some more frequently than others
  - like letters in a Scrabble-bag
  - each word w is drawn with a different probability $p(w)$

Probabilistic Document Model

- Repeatedly drawing from the bag of words results in a sequence of randomly drawn words → a document
  - $\mathbf{d} = (t_1, t_2, \ldots, t_{|\mathbf{d}|})$ where $t_j = w_{k_j} \in W$



Class-conditional Probabilities

- Different classes have different bags of words

Sports

Business

Politics

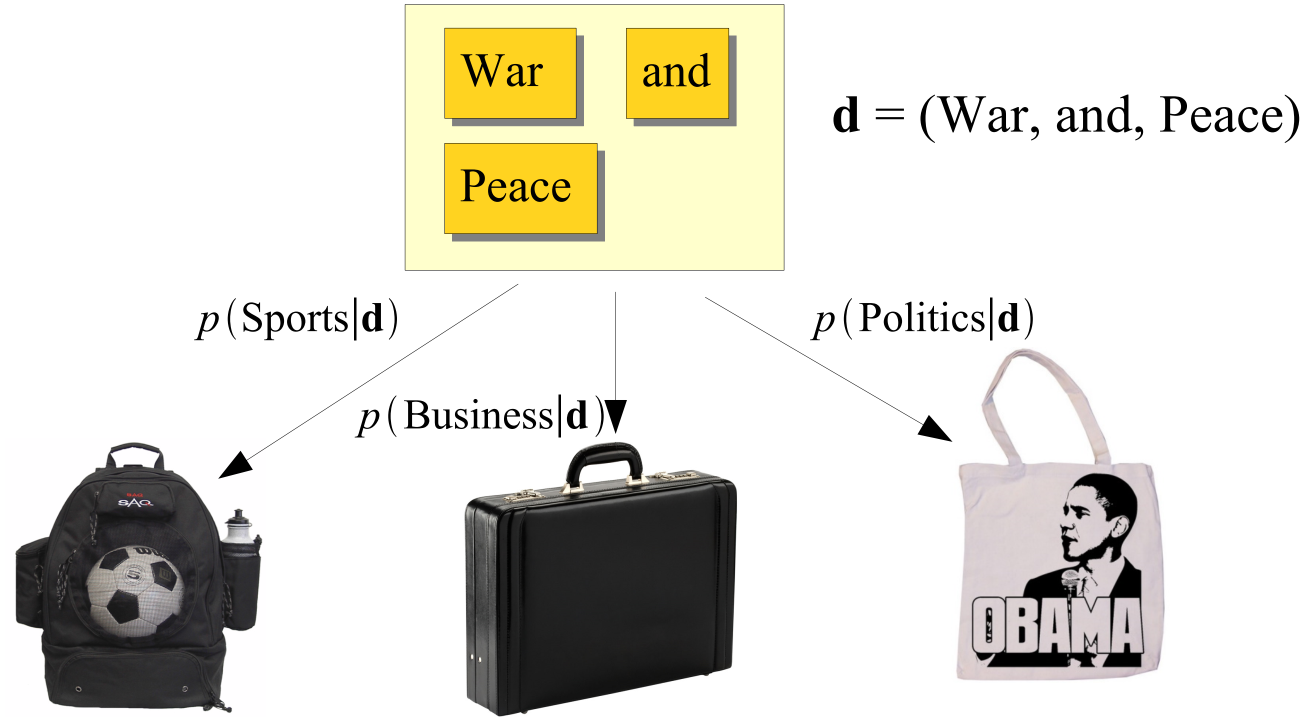
- probabilities of words in different classes are different
  - the sports bag contains more sports words, etc.
  - Formally: $p(w|c_i) \neq p(w|c_j) \neq p(w)$

Independence Assumption

- the probability that a word occurs does not depend on the context (the occurrence or not-occurrence of other words)
  - it only depends on the class of the document
- In other words:
  - Knowing the previous word in the document (or any other word) does not change the probability that a word occurs in position t_i
    $$p(t_i = w_{k_i} | t_j = w_{k_j}, c) = p(t_i = w_{k_i} | c)$$
    we will write this shorter as
    $$p(t_i | t_j, c) = p(t_i | c)$$
- Important:
  - the independence assumption does not hold in real texts!
  - but it turns out that it can still be used in practice

Probabilistic Text Classification

- Answer the question:
  - From which bag was a given document $\mathbf{d}$ generated?



- Answer is found by estimating the probabilities $p(c|\mathbf{d})$

Simple Naïve Bayes Classifier for Text (Mitchell 1997)

TECHNISCHE UNIVERSITÄT DARMSTADT

- a document is a sequence of n terms

$$p(\mathbf{d}|c) = p(t_1, t_2, \ldots. t_n|c)$$

- Apply Independence Assumption:
  - $p(t_i|c)$ is the probability with which the word $t_i = w_{i_j}$ occurs in documents of class c

$$p(\mathbf{d}|c) = \prod_{i=1}^{|\mathbf{d}|} p(t_i|c)$$

- Naïve Bayes Classifier
  - putting things together:

$$\boxed{c = \arg\max_c \prod_{i=1}^{|\mathbf{d}|} p(t_i|c)\, p(c)}$$

Estimating Probabilities

TECHNISCHE UNIVERSITÄT DARMSTADT

- Estimate for prior class probability $p(c)$
  - fraction of documents that are of class c

- Word probabilities can be estimated from data
  - estimated from fraction of document positions in each class on which the term occurs
    - put all documents of class c into a single (virtual) document
    - compute the frequencies of the words in this document

- Straight-forward approach:
  - estimate probabilities from the frequencies in the training set
  - word w occurs $n(\mathbf{d},w)$ times in document $\mathbf{d}$
  - What happens if there is a new word in a test document? Solutions?

$$p(t_i = w | c) = \frac{n_{w,c}}{\sum_{w \in W} n_{w,c}}$$

$$n_{w,c} = \sum_{\mathbf{d} \in c} n(\mathbf{d}, w)$$

Estimating Probabilities
Laplace Correction

- Straight-forward approach:
  - estimate probabilities from the frequencies in the training set
  - word w occurs $n(\mathbf{d},w)$ times in document $\mathbf{d}$

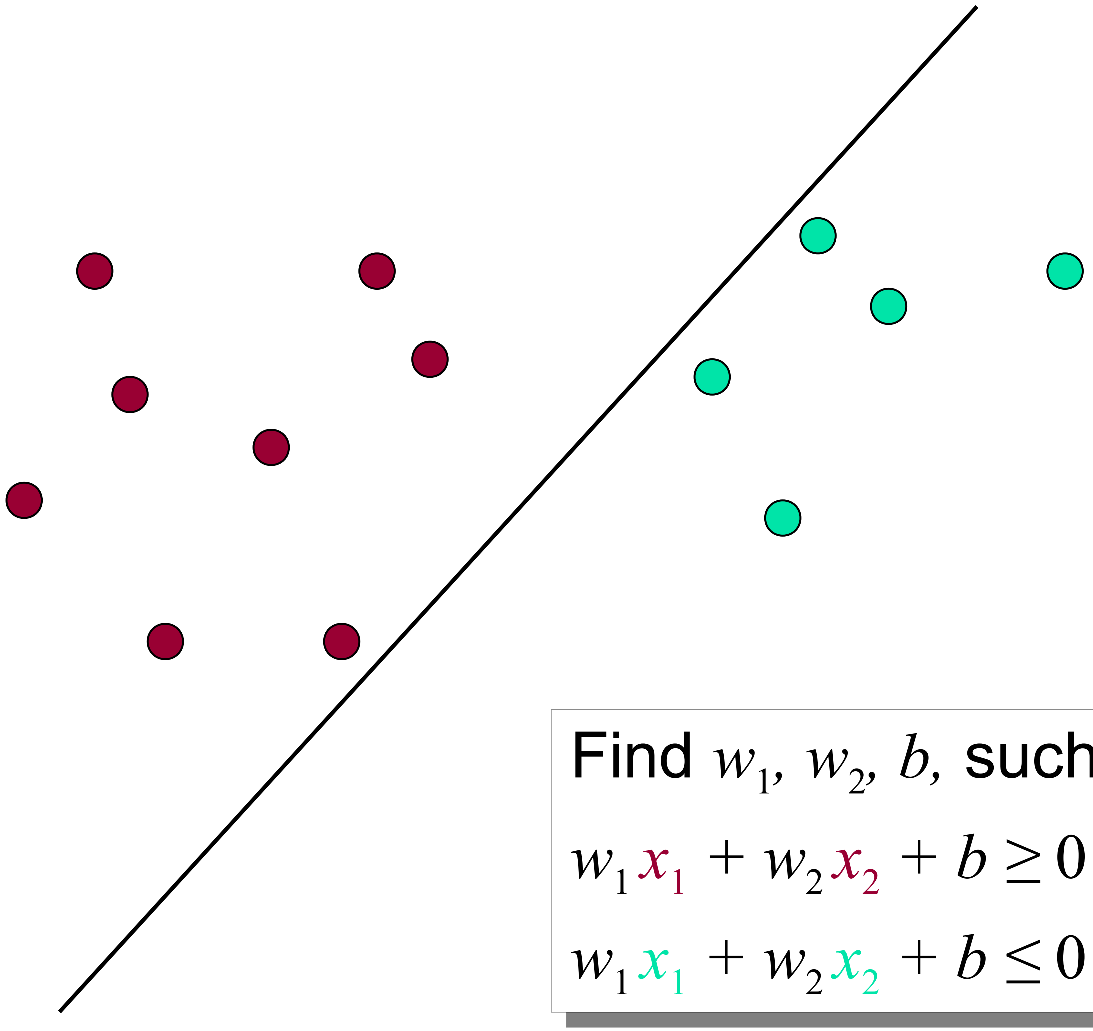
$$p(t_i = w|c) = \frac{n_{w,c}}{\sum_{w \in W} n_{w,c}}$$

$$n_{w,c} = \sum_{\mathbf{d} \in c} n(\mathbf{d}, w)$$

- Problem:
  - test documents may contain new words
  - those will be have estimated probabilities 0
  - assigned probability 0 for all classes
- Smoothing of probabilities:
  - basic idea: assume a prior distribution on word probabilities
  - e.g., Laplace correction assumes each word occurs at least once in a document

$$p(t_i = w|c) = \frac{n_{w,c} + 1}{\sum_{w \in W} (n_{w,c} + 1)} = \frac{n_{w,c} + 1}{\sum_{w \in W} n_{w,c} + |W|}$$

Finding a Linear Decision Boundary



Find w_1, w_2, b, such that

$w_1 x_1 + w_2 x_2 + b \geq 0$ for red points

$w_1 x_1 + w_2 x_2 + b \leq 0$ for green points

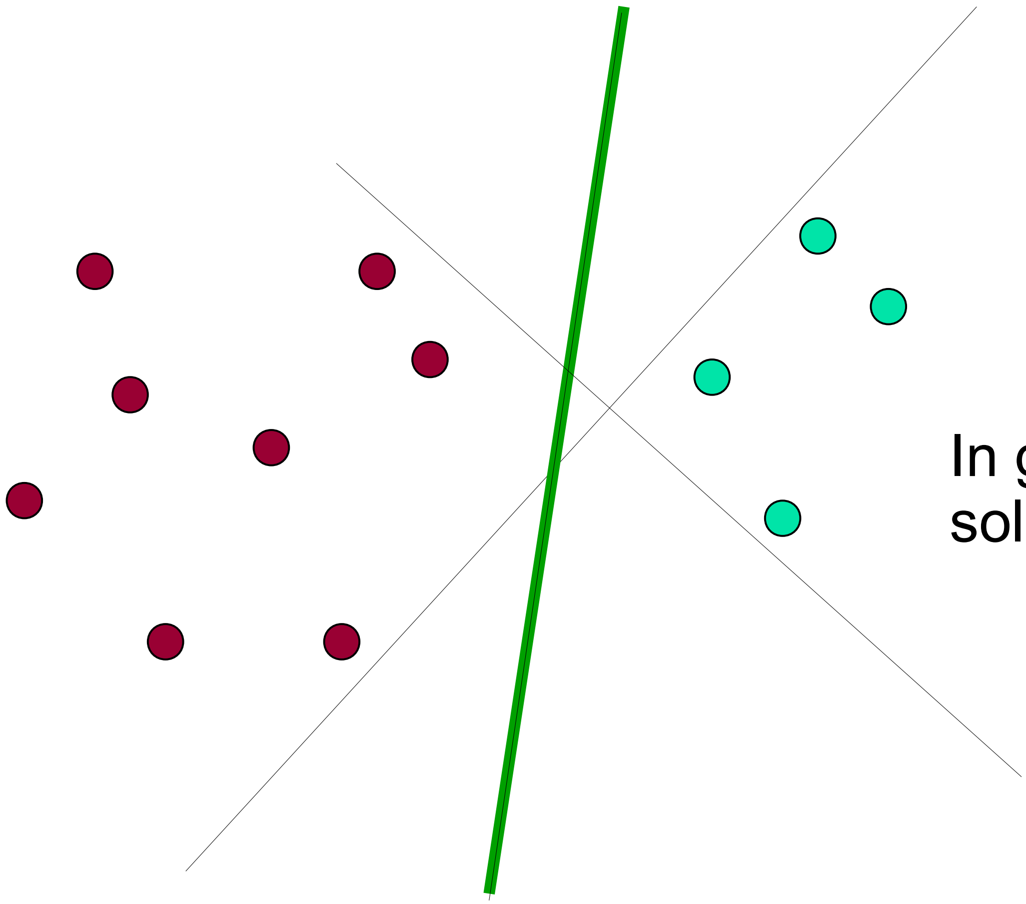
Manning and Raghavan

Fitting a linear decision boundary

- Discriminative approach
  - try to find a weight vector $\mathbf{w}$ so that the discrimination between the two classes is optimal
  - statistical approaches:
    - perceptrons (neural networks with a single layer)
    - logistic regression
  - most common approach in text categorization
    $\rightarrow$ support vector machines

Chakrabarti & Ramakrishnan

Which Hyperplane?



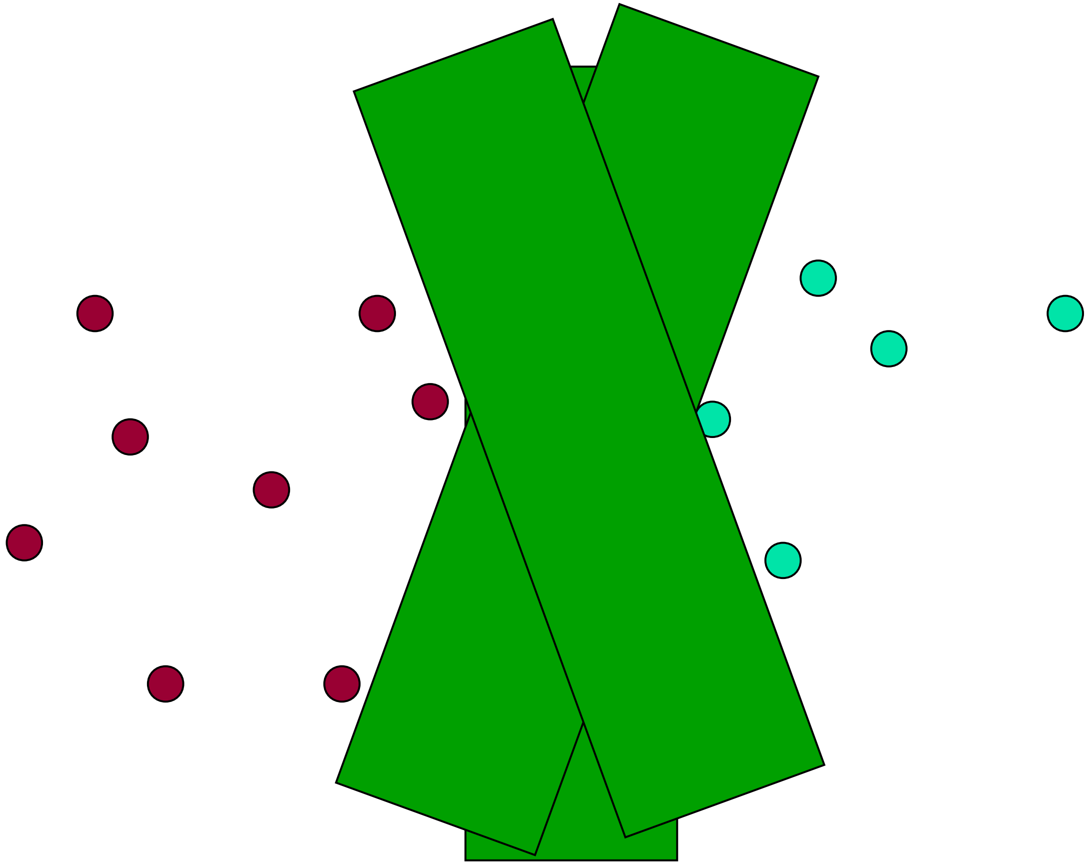
In general, many possible solutions for $\mathbf{w} = (w_1, w_2), b$

- Intuition 1: If there are no points near the decision surface, then there are no very uncertain classifications → better generalization

Manning and Raghavan

Support Vector Machines: Intuition

- Intuition 2: If you have to place a fat separator between classes, you have less choices, and so overfitting is not so easy



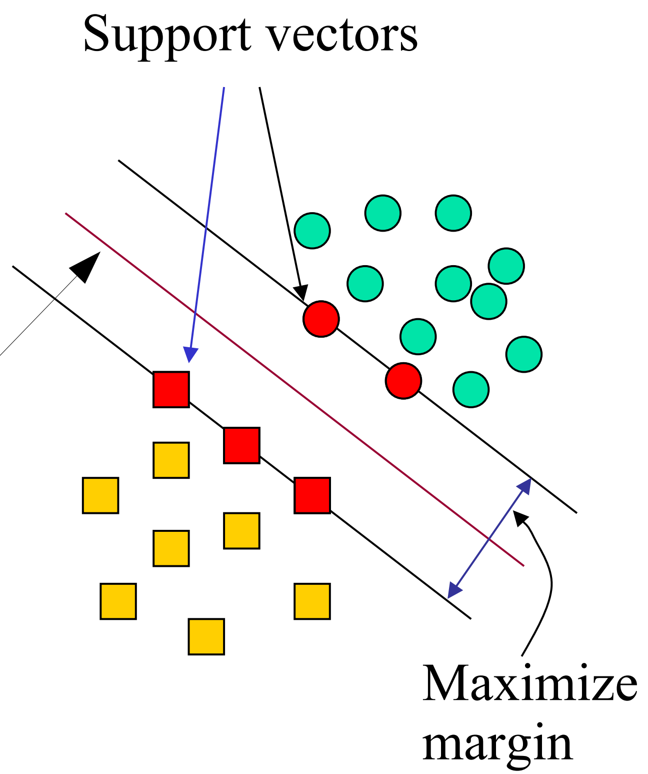
Manning and Raghavan

Support Vector Machine (SVM)

TECHNISCHE UNIVERSITÄT DARMSTADT

- SVMs maximize the *margin* around the separating hyperplane.
  - a.k.a. large margin classifiers
- The decision function is fully specified by a subset of training samples, the *support vectors*.

$$\mathbf{w}^T \cdot \mathbf{x}_i + b = 0$$



- Formalization
  - $\mathbf{w}$: normal vector to decision hyperplane
  - $\mathbf{x}_i$: i-th data point
  - y_i: class of data point i ($+1$ or -1) NB: Not 1/0
  - Classifier is:

$$f(\mathbf{x}_i) = \text{sign}(\mathbf{w}^T \mathbf{x}_i + b)$$

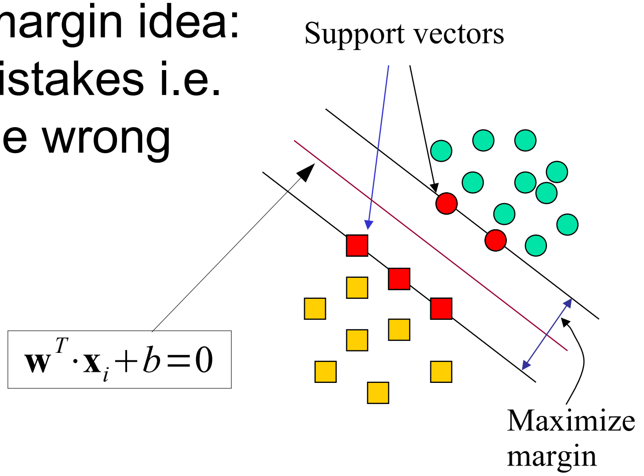
Manning and Raghavan

Support Vector Machine (SVM) Mathematics

TECHNISCHE UNIVERSITÄT DARMSTADT

Regularization 1: try to minimize the complexity of the model
→ better generalization

Regularization 2, Soft margin idea: slack variables allow mistakes i.e. training examples on the wrong side of the hyperplane



Find **w** and *b* such that

$$\Phi(\mathbf{w}) = \tfrac{1}{2}\,|\mathbf{w}|^2 + C\Sigma\xi_i \text{ is minimized}$$

Regularization control parameter: trade-off between underfitting and overfitting to training data

and for all $\{(\mathbf{x_i}, y_i)\}$:

$$y_i\,(\mathbf{w}^{\mathbf{T}}\mathbf{x_i} + b) \geq 1 - \xi_i \quad \text{and} \quad \xi_i \geq 0$$

Capacity Control
Occam's Razor

> Entities should not be multiplied beyond necessity.
>
> *William of Ockham (1285 - 1349)*

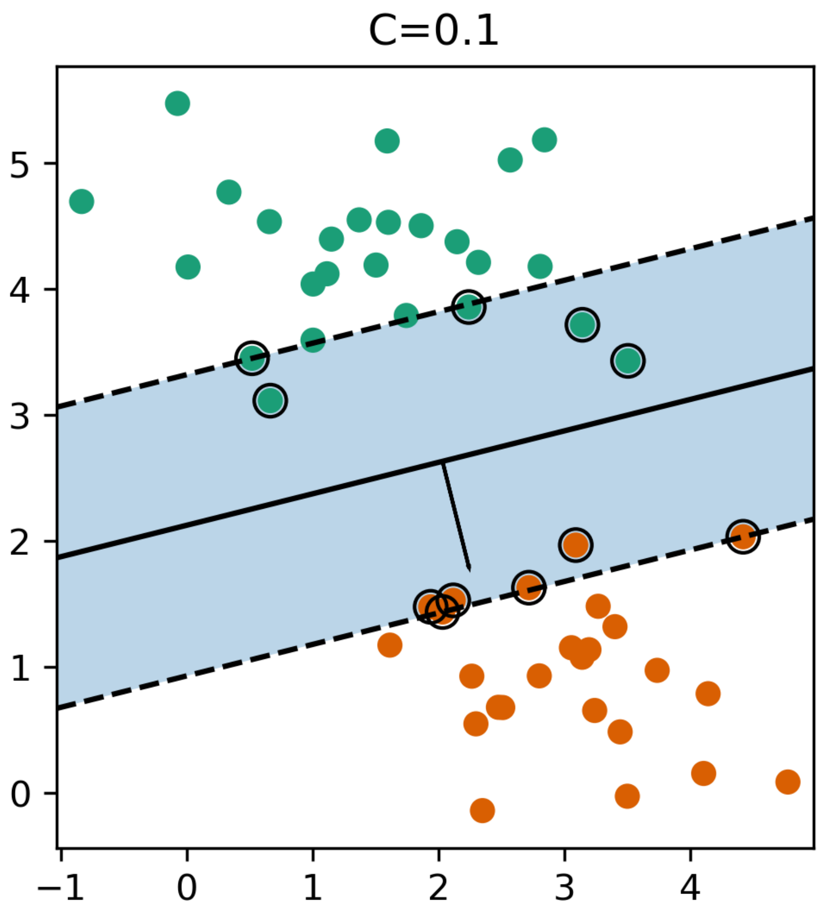
- Machine Learning Interpretation:
  - Among theories of (approximately) equal quality on the *training* data, simpler theories have a better chance to be more accurate on the *test* data
  - It is desirable to find a trade-off between *accuracy* and *complexity* of a model
- (Debatable) Probabilistic Justification:
  - There are more complex theories than simple theories. Thus a simple theory is less likely to explain the observed phenomena by chance.

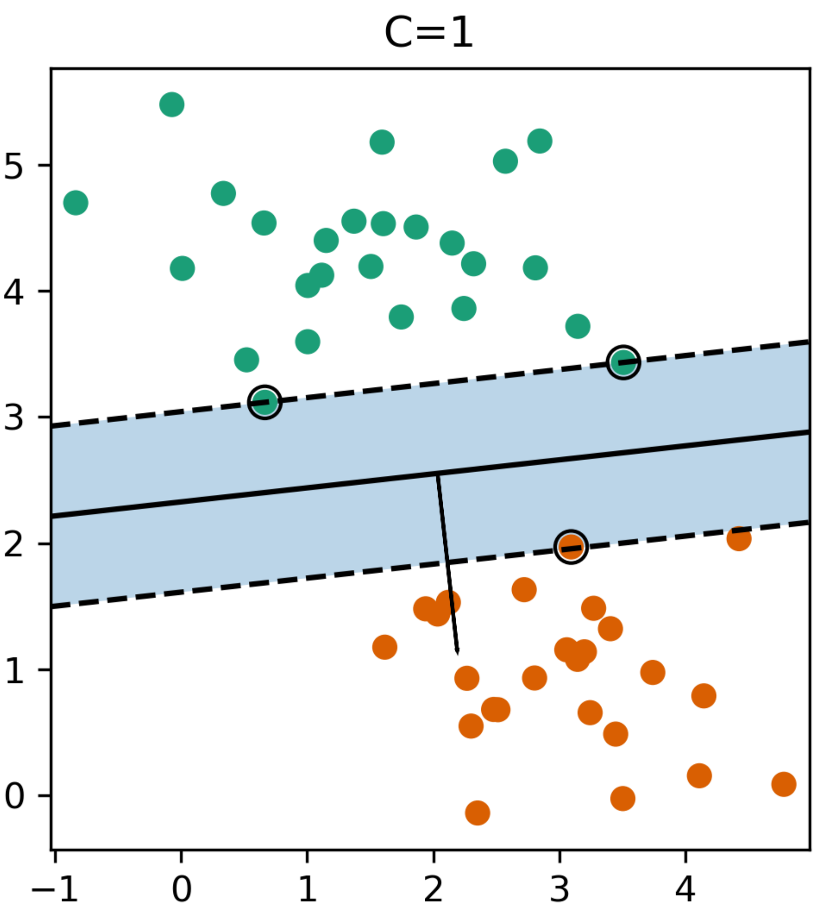
Capacity Control
Overfitting

- Given
  - a fairly general model class (e.g., rules)
  - enough degrees of freedom (e.g., no length restriction)
- you can always find a model that explains the data
  → Overfitting
- Such concepts do not generalize well!
- Particularly bad for noisy data
  - Data often contain errors due to
    - inconsistent classification
    - measurement or annotation errors
    - missing values
    - some other kinds of noise

→ Capacity control

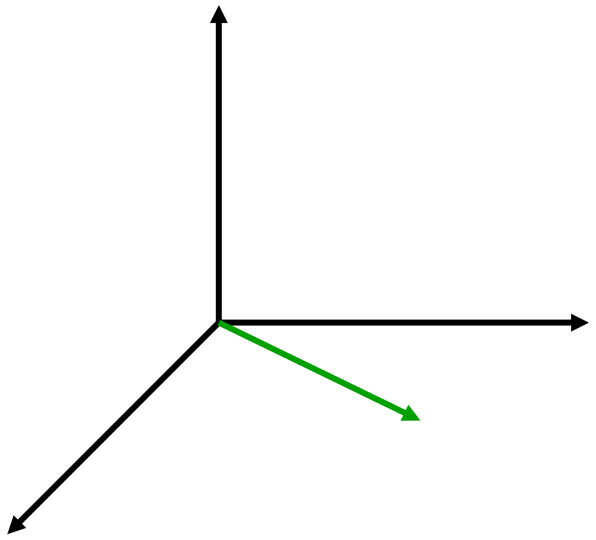
Regularization
Example for linear SVMs





High Dimensional Data

- Pictures like the one at right are misleading!
  - Documents are zero along almost all axes
  - Most document pairs are very far apart
    - (i.e., not strictly orthogonal, but only share very common words and a few scattered others)



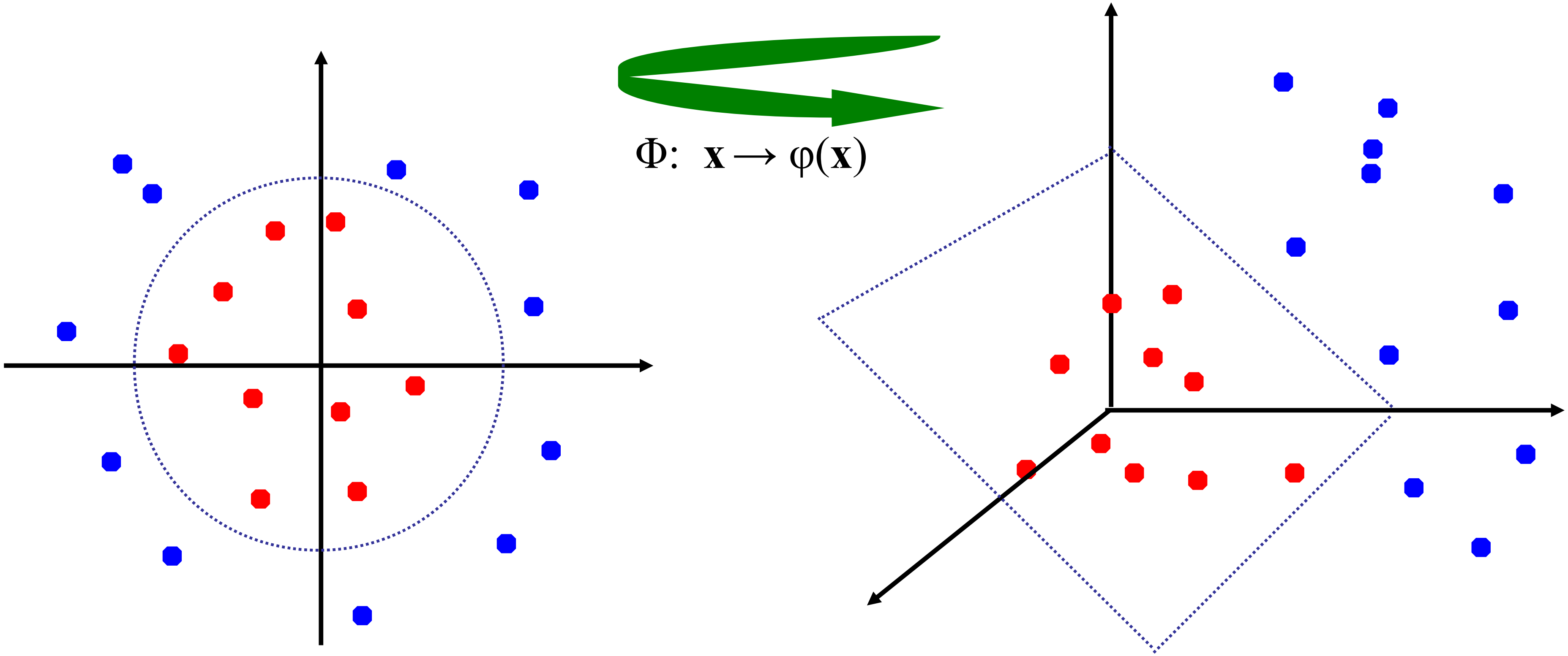
- In classification terms:
  - virtually all document sets are separable, for almost any classification

- This is part of why linear classifiers are quite successful in text classification
  → SVMs with linear Kernels are usually sufficient!

Manning and Raghavan

Non-linear SVMs
Feature spaces

- **General idea:** the original feature space can always be mapped to some higher-dimensional feature space where the training set is separable:



Manning and Raghavan

Non-linear SVMs
Kernel-Trick

TECHNISCHE UNIVERSITÄT DARMSTADT

- Replace inner product operation by Kernel function

  $K(\mathbf{x_i},\mathbf{x_j}) = \varphi(\mathbf{x_i})^{\mathrm{T}}\varphi(\mathbf{x_j})$

  → make data separable

  → map data into better representational space
- Common kernels
  - Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \cdot \mathbf{x}_j$
  - Polynomial: $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \cdot \mathbf{x}_j)^d$
  - Radial basis function (infinite dimensional space)

$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}}$$

Manning and Raghavan

Regularization
Example for non-linear SVMs