ETH zürich

Machine Learning Methods for Biomedical Data

Felipe Llinares-López and Damian Roqueiro
Machine Learning & Computational Biology Lab
D-BSSE, ETH Zürich

Tutorial AM2: Machine learning methods in the analysis of genomic and clinical data. July 6, 2018

Machine learning

A definition by Tom Mitchell Mitchell [1997]

...“[A] machine learns with respect to a particular task $\mathcal{T}$, performance metric $\mathcal{P}$, and type of experience $\mathcal{E}$,

if the system reliably improves its performance $\mathcal{P}$ at task $\mathcal{T}$, following experience $\mathcal{E}$...”

- Recommender systems, e.g. Netflix®
- Tailoring of ads and newsfeeds in social networks, e.g. Facebook®
- Web searches and ranking of pages, e.g. Google® search
- Spam filtering of e-mails, and many others

Machine learning

A definition by Tom Mitchell Mitchell [1997]

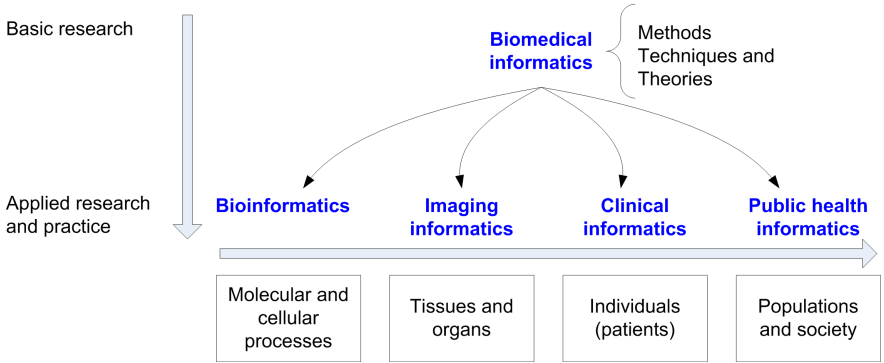
...“[A] machine learns with respect to a particular task $\mathcal{T}$, performance metric $\mathcal{P}$, and type of experience $\mathcal{E}$,

if the system reliably improves its performance $\mathcal{P}$ at task $\mathcal{T}$, following experience $\mathcal{E}$...”

- Recommender systems, e.g. Netflix®
- Tailoring of ads and newsfeeds in social networks, e.g. Facebook®
- Web searches and ranking of pages, e.g. Google® search
- Spam filtering of e-mails, and many others

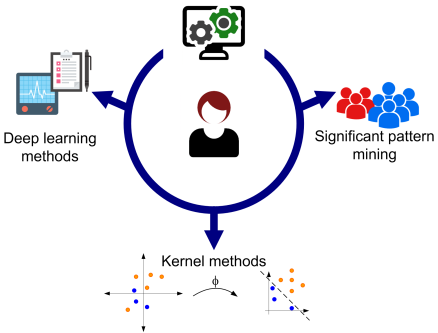
Biomedical data

Spectrum of data types



Source: Extracted and modified from Shortliffe and Cimino [2014]

Tutorial overview



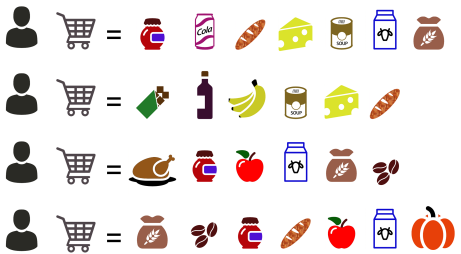
Source: All icons in figures were downloaded & modified from: flaticon.com (designed by Freepik)

Module I. Significant pattern mining

Feature interaction $\rightarrow$ biomarker discovery
Multiple hypothesis testing problem
Contributions and applications to genomic studies

Significant pattern mining: motivation

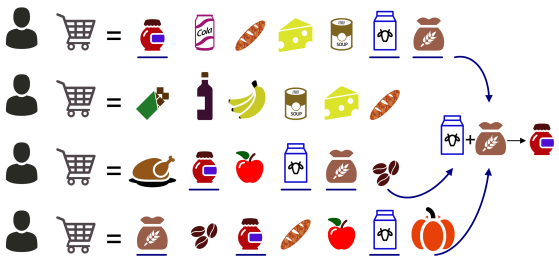
Affinity analysis



- **Task 1**: identify sets of products that are jointly bought by most customers
- **Task 2**: construct association rules of frequent itemsets
- Apriori property: No superset of an infrequent itemset can be frequent

Significant pattern mining: motivation

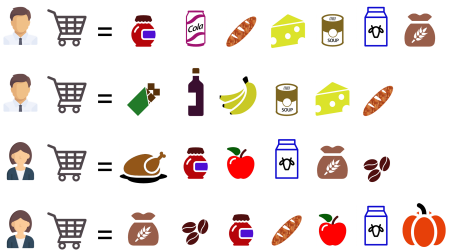
Affinity analysis



- **Task 1**: identify sets of products that are jointly bought by most customers
- **Task 2**: construct association rules of frequent itemsets
- Apriori property: No superset of an infrequent itemset can be frequent

Significant pattern mining: motivation

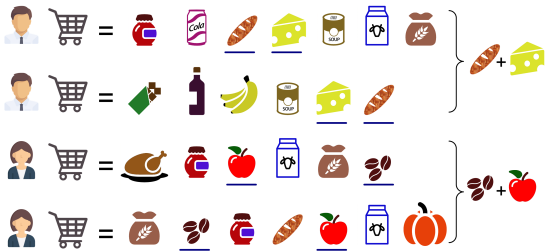
Discriminative pattern mining



- From **unlabeled** to **labeled** data
- Task: identify sets of products frequently bought by one group and not the other

Significant pattern mining: motivation

Discriminative pattern mining



- From **unlabeled** to **labeled** data
- **Task**: identify sets of products frequently bought by one group and not the other

Significant pattern mining: motivation

Discriminative pattern mining

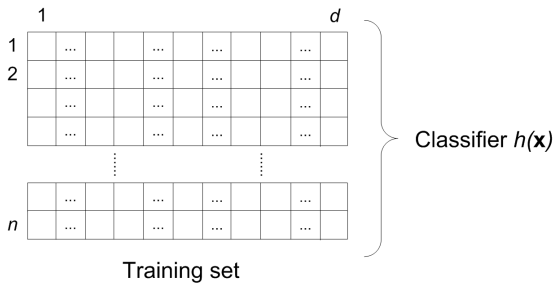
- How to measure the statistical significance of these findings?
- **Challenges**: Traverse vast search space while attaining reasonable statistical power

Module II. Comparison of structured data

(Implicitly) Mapping data to different space
Kernel methods
Applications in graphs

Comparison of structured data: motivation

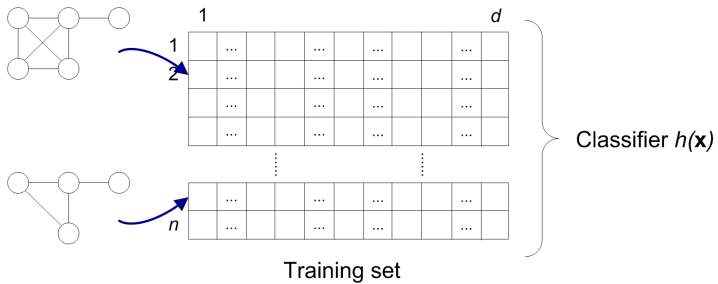
Construction of a classifier



- Machine learning toolboxes expect data in matrix form
- How to best "transform" the original data for maximum classification performance?

Comparison of structured data: motivation

Construction of a classifier

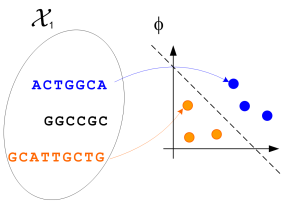


- Machine learning toolboxes expect data in matrix form
- How to best "transform" the original data for maximum classification performance?

Comparison of structured data: motivation

Mapping structured data to a vectorial space

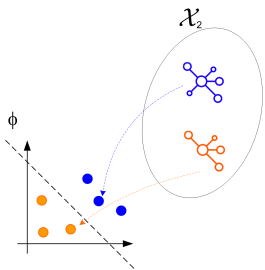
- **Task**: Compare two objects via a mapping to a vector space
- Kernel methods have shown great versatility



Comparison of structured data: motivation

Mapping structured data to a vectorial space

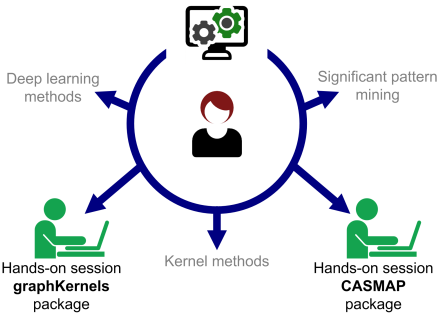
- **Task**: Compare two objects via a mapping to a vector space
- Kernel methods have shown great versatility



Module III. Deep learning and applications to biomedical data

Text mining
Skip-gram model
Link between graph kernels and text mining

Game plan for this morning



References I

T. M. Mitchell. *Machine Learning*. McGraw-Hill, Inc., New York, NY, USA, 1 edition, 1997. ISBN 0070428077, 9780070428072.

E. H. Shortliffe and J. J. Cimino. *Biomedical Informatics – Computer Applications in Health Care and Biomedicine*. Springer-Verlag London, 4th edition, 2014. ISBN 978-1-4471-4474-8.

Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomical? *PLOS Biology*, 13(7):1–11, 07 2015. doi: 10.1371/journal.pbio.1002195. URL `https://doi.org/10.1371/journal.pbio.1002195`.