Introduction to Machine Learning
for Molecules and Materials

Matthias Rupp

Fritz Haber Institute of the Max Planck Society, Berlin, Germany

2017 NYU Shanghai Summer School on
Machine Learning in the Molecular Sciences

June 12–16, Shanghai, China

NYU

華東師範大學
EAST CHINA NORMAL UNIVERSITY

NYU 上海 SHANGHAI 纽约大学

Outline

1. Introduction to machine learning
   overview, basic concepts, some applications

2. Learning with kernels
   kernel ridge regression, Gaussian process regression

3. Property prediction
   applications, examples, representation

4. Practical session
   hands-on exercises using `qmmlpack`

Machine learning

Machine learning (ML) studies algorithms whose performance **improves with data** ("learning from experience"). Mitchell, McGraw Hill, 1997

Data $\boldsymbol{X}$ $\rightarrow$

$\rightarrow$ Model $\hat{f}$

- widely applied, many problem types and algorithms
- systematic identification of regularity in data for prediction & analysis
- interpolation in high-dimensional spaces
- inductive, data-driven; empirical in a principled way
- connections to statistics, mathematics, computer science, physics, ...
  example: information theory

Literature

Conferences:

- Annual Conference on Neural Information Processing Systems (NIPS)
- International Conference on Machine Learning (ICML)
- Conference on Learning Theory (COLT)

Textbooks:

- Vapnik: Nature of Statistical Learning Theory, Springer, 2001
- Duda, Hart, Stork: Pattern Classification, Wiley, 2001
- Bishop: Pattern Recognition and Machine Learning, Springer, 2006
- Hastie, Tibshirani, Friedman: Elements of Statistical Learning, Springer, 2003

https://nips.cc, https://icml.cc, https://www.learningtheory.org

Examples of machine learning applications

• brain-computer interfaces	flipper; dictation
• natural language processing	Google translate
• recommender systems, advertising	burglar example
• fraud detection, network security	
• robotics, autonomous vehicles	
• image processing, computer vision	paintings
• games	
• oil industry / geology	Gaussian processes
• . . .	

molecular and materials sciences, bioinformatics

Types of problems

Unsupervised learning: Data do not have labels
Given $\{x_i\}_{i=1}^{n}$, find structure

- dimensionality reduction

Burges, now Publishers, 2010

Supervised learning: Data have labels
Given $\{(x_i, y_i)\}_{i=1}^{n}$, predict $\tilde{y}$ for new $\tilde{x}$

- novelty detection
- classification
- regression

Semi-supervised learning: Some data have labels
Given $\{(x_i, y_i)\}_{i=1}^{n}$ and $\{x_i\}_{i=1}^{m}$, $m \gg n$, predict $\tilde{y}$ for new $\tilde{x}$

Types of problems

Matrix completion:
Given a partially occupied matrix, find missing elements
Example: ligands versus protein receptors

Active learning: Algorithm chooses data to label
Choose n data $\{x_i\}_{i=1}^{n}$ to predict $\tilde{y}$ for new $\tilde{x}$

Reinforcement learning: Algorithm acts based on rewards
Given a state space, algorithm learns to maximize rewards for its actions

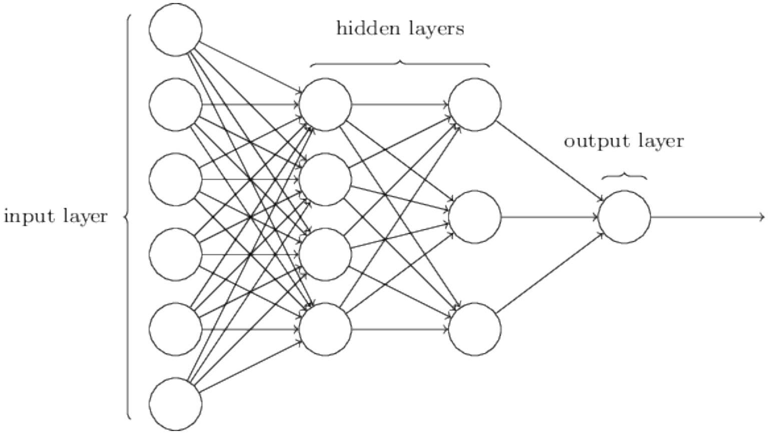
Online learning: Algorithm predicts data as they arrive
Stream of data to predict, minimize overall error

Covariate shift: Algorithm adapts to changing data
Predicted data come from a different distribution than training data

Algorithms

- artificial neural networks → Prof. Tuckerman (afternoon)
- random forests → Prof. Zhang (tomorrow)
- support vector machines Cristianini & Shawe-Taylor, 2000
- kernel ridge regression → second lecture
- Gaussian processes → second lecture
- principal component analysis Jolliffe, 2004
- symbolic regression Schmidt, Lipson, 2009
- many others. . .

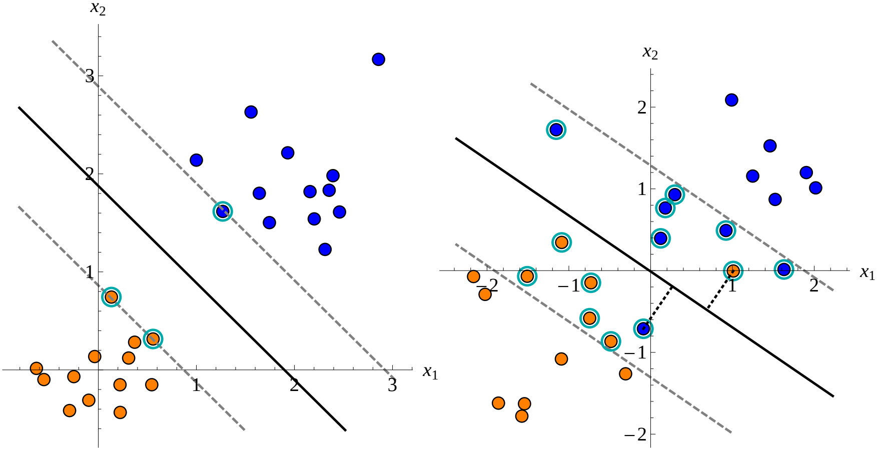
Artificial neural networks



$$f(x_{i,j}) = h\Big(\sum_{k=1}^{n_i} w_{i-1,k} f(x_{i-1,k})\Big)$$

- parametric model
- universal function approximator
- training via non-convex optimization
- → Prof. Tuckerman

Support vector machines



linear separable problem

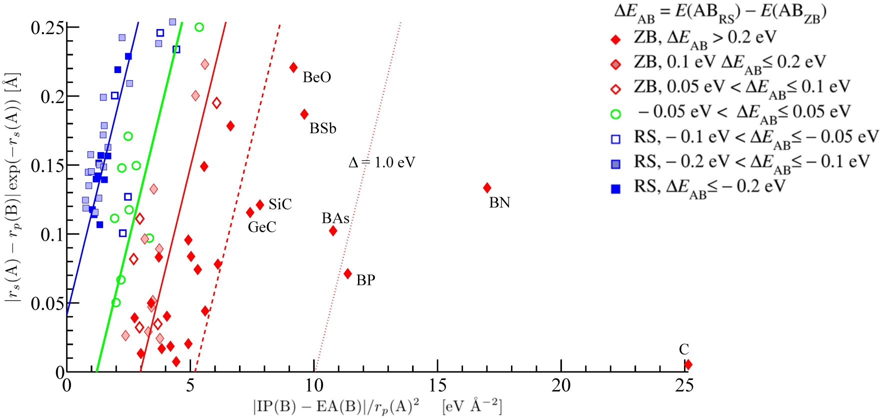
linear inseparable problem

maximal margin plane bisects (reduced) convex hull closest points

Ivanciuc: *J. Chem. Inf. Model.* 40, 1412, 2000; Bennett, Campbell: *SIGKDD Explor.* 2, 1, 2000

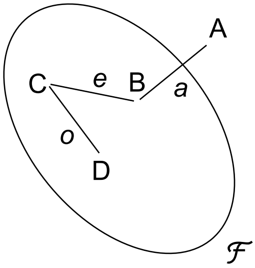
Symbolic regression

- stochastic search in the space of analytic functions
- fast, interpretable models



Schmidt, Lipson, *Science*, 5923, 2009; Ghiringhelli et al, *Phys. Rev. Lett.*, 2015

Learning theory



prediction error =

approximation error a

+ estimation error e

+ optimization error o

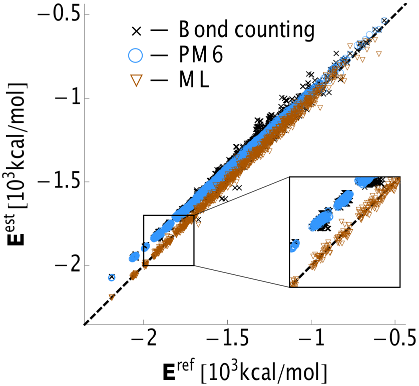
$\mathcal{F}$ = model class, A = true model, B = best model in class, C = best identifiable model (data), D = best identifiable model (optimization)

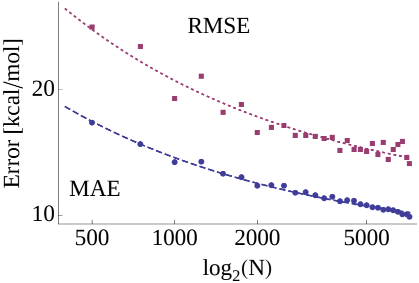
Changes in size of $\mathcal{F} \Leftrightarrow a$ vs. $e \Leftrightarrow$ **bias-variance trade-off**

Bottou & Bousquet, *NIPS* 2007

Example: predicting atomization energies

- 7 165 small organic molecules (H,C,N,O,S; 1–7 non-H atoms)
- DFT PBE0 atomization energies
- today, errors are ~ 0.5 kcal/mol for this dataset





Rupp, Tkatchenko, Müller, von Lilienfeld: *Phys. Rev. Lett.*, 2012

Design: the inverse problem

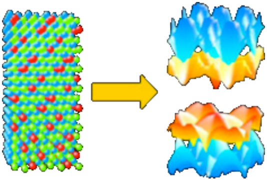
Find a molecule or material with given properties

Example:

Maximize binding constant pK_i to PPARγ,
with aqueous solubility $\log S < 0$, at T = 37°C and pH = 7.2.

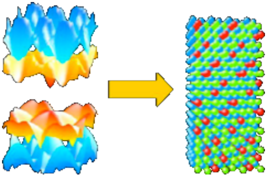
Assumption: Determining properties is possible, but expensive ("oracle")

Direct approach



screen compound collections
map structure to property

Inverse approach



design compounds
map property to structures

Figure: Franceschetti, Zunger: Nature, 60, 1999

Confidence estimates

> "It is not the estimate [...] that matters so much
> as the degree of confidence with the opinion"
>
> Taleb, Random House, 2004

Known in cheminformatics as **domain of applicability**:

How to determine whether new data $\boldsymbol{x}$ are in the interpolation region?

Quantitatively: How far away is $\boldsymbol{x}$ from the training data?

- naive approaches can help as filters variable ranges → blackboard
- distance alone is insufficient in high-dimensional spaces → blackboard
- with the usual i.i.d. assumption in ML, this problem does not exist

Sushko et al.: *J. Chem. Inf. Model.*, 2094, 2010; Sushko et al.: *J. Chemometr.*, 202, 2010

Tendencies in ML for experimental versus computed data

experimental

- fewer data
- strong noise
- "integrated" properties
  descriptors
- enrichment in screening
- limited by synthesis
- cheminformatics
  quantitative structure-activity/
  property relationships

computed

- more data
- no noise
- dependence on atom coordinates
  unique representations
- fast and accurate predictions
- limited by approximations
- interpolation of ab initio data
  quantum mechanics / machine
  learning models

Selassie, Verma; in: Abraham, Rotella (eds.), Burger's Med. Chem., 7th ed, vol. 1, Wiley, 2010

Nomenclature

The words *descriptor* and *fingerprint* originate from cheminformatics.

Descriptor: ("descriptive parameter")
Any numerical encoding of a (structural) property of a molecule

> "The molecular descriptor is the final result of a [...] mathematical procedure..." (Todeschini & Consonni, Wiley, 2009)

Often a vector of **heterogeneous** properties, selected *ad hoc*

Fingerprint: (subclass of descriptors)
Fixed-length (bit) pattern characterising a molecule

Usually *homogeneous* and *topology*-based (substructure fingerprints)

Representation: (subclass of descriptors)
Fulfills theoretical requirements for accurate predictions

Introduced to distinguish from *ad hoc* descriptors

Descriptors

- computable properties in vector form; graph kernels
- used for experimental properties in cheminformatics
- use chemical abstractions, typically not unique and discontinuous
  ⇒ best for “integrated” properties

1-pentyl acetate

🟩 Bonds in longest chain: 7

🟦 Rotatable bonds: 4

🟥 Negative partial charge surface fraction: 0.13

🟨 Hydrogen bond acceptors: 1

. . .

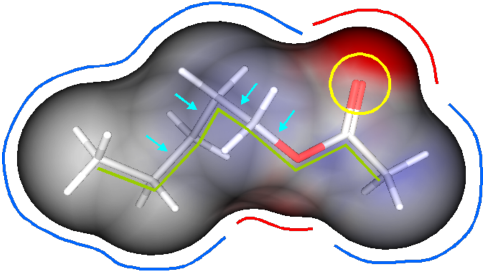


Figure: Michael Schmuker

Todeschini, Consonni: Handbook of Molecular Descriptors, Wiley, 2009;
Rupp, Schneider, Schneider: J. Comput. Chem., 108, 2008.

Representations

- numerical encoding of atomistic system for accurate interpolation
- together with kernel, defines space / basis functions

Requirements

- **invariant**: against transformations preserving the property
  in particular translation, rotation, homonuclear permutations
- **unique**: different in property ⇒ different in representation
  allows reconstruction of system
- **smooth**: continuous, ideally differentiable
  works together with ML; needed for forces
- **general**: encode any system, including molecules and crystals
- **fast**: cheaper to compute than reference method
- **efficient**: supports learning by requiring few reference data

Huo, Rupp: arXiv 1704.06439, 2017; → Profs. von Lilienfeld, Tuckerman

Sources of data

experimental data

Literature

Databases:

- PubChem (pubchem.ncbi.nlm.nih.gov, >90 M compounds)
- Online Chemical Database (ochem.eu, >1.3 M records)
- Springer Materials
- Cambridge Crystallographic Database

computed data

Literature

Databases:

- Materials Project (materialsproject.org)
- Novel Materials Discovery (nomad-coe.eu)
- Open Quantum Materials Database
- AFLOWLib

Summary

- machine learning finds regularity in data for analysis or prediction, improving with more data
- there are many problem types and algorithms
- it can predict experimental and computational outcomes