Road map for tutorial 8: Big-data analytics
Compressed sensing and screening for properties

Emre Ahmetcik,
Angelo Ziletti, and Luca M. Ghiringhelli
Fritz-Haber-Institut der MPG, Berlin

Hands-On DFT and Beyond: Frontiers of Advanced Electronic Structure and Molecular Dynamics Methods
Peking University, Beijing, China, July 30th to August 10th, 2018
Today’s tutorial

Method(s):
Compressed sensing for feature extraction (see this morning’s lecture):
- construction of the feature space
- LASSO, LASSO+$\ell_0$
- SISSO (Sure Independence Screening combined with Sparsifying Operator

See also:
Today's tutorial

Method(s):
Compressed sensing for feature extraction (see this morning’s lecture):
- construction of the feature space
- LASSO, LASSO+$\ell_0$
- SISSO (Sure Independence Screening combined with Sparsifying Operator)

Important take-home message:
How to do cross validation with feature selection

See also:
Method(s):
Compressed sensing for feature extraction (see this morning’s lecture):
- construction of the feature space
- LASSO, LASSO+$\ell_0$
- SISSO (Sure Independence Screening combined with Sparsifying Operator

Important take-home message:
How to do cross validation with feature selection

Showcase application:
Predicting energy differences between crystal structures:
(Meta-)stability of octet-binary compounds

See also:
Method(s):
Compressed sensing for feature extraction (see this morning’s lecture):
- construction of the feature space
- LASSO, LASSO+$\ell_0$
- SISSO (Sure Independence Screening combined with Sparsifying Operator

Important take-home message:
How to do cross validation with feature selection

Showcase application:
Predicting energy differences between crystal structures:
(Meta-)stability of octet-binary compounds

Implementation:
Python (mainly) notebook

See also:
Today’s tutorial

Implementation:
Browser-based-notebook interface

A Browser-based notebook is a virtual environment for programming, working via the user’s browser. In our case, all functionalities are running on the server: meaning, no implementation needed by the user.

Why?
It allows for both text and pieces of code that can interactively run in the same environment. Ideal for tutorials.

What to do:
Copy $HandsOn/tutorial_8_compressed_sensing into your working directory and follow the instruction on “Instructions.pdf”
The Novel Materials Discovery (NOMAD) Laboratory maintains the largest Repository, for input and output files of all important computational materials science codes.

From its open-access data, it builds several Big-Data Services helping to advance materials science and engineering.

To learn more, click on the buttons above. You can also watch our 3-minute summary on the NOMAD Laboratory CoE at YouTube (or at YOUKU in China).

NOMAD Scope and Overview
CONVERSION LAYER

NOMAD REPOSITORY

https://nomad-coe.eu

ENCYCLOPEDIA

ADVANCED GRAPHICS

THE ARCHIVE

BIG-DATA ANALYTICS
Subgroup discovery, compressed-sensing, and more... tutorials @ https://analytics-toolkit.nomad-coe.eu/
Try it later (no registration needed, Nothing to install, you need only a web browser)
Querying and visualizing the content of the NOMAD Archive
Querying and visualizing the content of the NOMAD Archive
Querying and visualizing the content of the NOMAD Archive

DASHBOARD
Welcome luca!

Open a new default notebook

Resource Usage

<table>
<thead>
<tr>
<th>Private Storage</th>
<th>Shared Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.85 MB</td>
<td>75.45 KB</td>
</tr>
</tbody>
</table>

My notebooks
Analytics Toolkit

— Predicting energy differences between crystal structures: (Meta-)-stability of octet-binary compounds

created by: Angelo Ziletti\textsuperscript{1}, Emre Ahmetcik\textsuperscript{1}, Runhai Ouyang\textsuperscript{1}, Ankit Kariryaa\textsuperscript{1}, Fawzi R. Mohamed\textsuperscript{1}, Luca Ghiringhelli\textsuperscript{1}, and Matthias Scheffler\textsuperscript{1}

\textsuperscript{1} Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, D-14195 Berlin, Germany

[Last updated: May 15, 2017]
— Predicting energy differences

Relative stability (energy difference between two structures):

**Primary features** (hover the mouse pointer over the feature names to see a full description):

- ✓ IP
- ✓ $r_p$
- ❏ $n_{period}$
- ✓ $E_A$
- ✓ $r_d$
- ✓ $r_s$
- ❏ $Z$
- ❏ $d^{AA}$

**Allowed operations:**

Given features $x$ and $y$, apply these operations for feature space construction:

- ✓ $x + y$
- ❏ $x \cdot y$
- ✓ $x^2$
- ✓ $\exp(x)$
- ✓ $\exp(-x)$
- ❏ SCD
- ❏ $x - y$
- ✓ $x/y$
- ✗ $x^3$
- ❏ $x^{(-1)}$
- ❏ $\sqrt{x}$
- ❏ $\log(x)$
Predicting energy differences

Relative stability (energy difference between two structures):

**Primary features** (hover the mouse pointer over the feature names to see a full description):

- IP
- \( r_p \)
- \( n_{\text{period}} \)
- \( d^{\text{AA}} \)
- \( r_d \)
- \( r_s \)
- EA

Radius at which the radial probability density of the s valence orbital is maximum

**Allowed operations:**

Given features \( x \) and \( y \), apply these operations for feature space construction:

- \( x+y \)
- \( x \cdot y \)
- \( x^2 \)
- \( \exp(x) \)
- \( \exp(-x) \)
- \( x/y \)
- \( |x-y| \)
- \( x^{-1} \)
- \( \sqrt{x} \)
- \( \log(x) \)

Standard Cauchy distribution
Furthermore, other tutorials one the analytics home page:

- Application of SISSO to metal/insulator classification (ref. my talk in the morning)
- Application of SISSO to topological insulators (ref Matthias’ talk tomorrow)
- Application of subgroup discovery (ref. my talk in the morning) to RS/ZB classification and gold clusters’ structure-property relationship
- Application of ConvNets to crystal structure recognition (ref. my talk in the morning)
- Error bars in DFT calculations
- Cluster expansion hamiltonians for binary alloys (ref. Santiago’s tutorial on Monday)

... and more
Today’s tutorial

A couple of tips:
- expand tutorial sections

+ Introduction to the compressed sensing methods
+ Get the data from the NOMAD Archive
+ Determining low-dimensional descriptors with the $\ell_0$ method
+ Approximations to the $\ell_0$ method
+ Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool
+ Predicting properties of new materials
+ The SISSO for classification
Today’s tutorial

A couple of tips:
- expand tutorial sections

+ Approximations to the $\ell_0$ method

- Approximations to the $\ell_0$ method

- Perform a LASSO minimization and the LASSO+$\ell_0$ method.
- Compare the solutions with the ones from the $\ell_0$ method.

```javascript
1 // 'Import' relevant functions from chapters before
2 var functions_list = ["json_list", "nomad_structure_list", 'get_energies', 'get_energy_diffs', 'get_descriptors', 'get_data', 'L0'];
3 var n_functions = functions_list.length
4 var i;
5 for (i = 0; i < n_functions; i++) {
6   beaker.evaluate(functions_list[i]);
7   beaker.print('import ' + functions_list[i]);
8 }
```

Run
Today’s tutorial

A couple of tips:
- run cells

Approximations to the $\ell_0$ method

- Perform a LASSO minimization and the LASSO+$\ell_0$ method.
- Compare the solutions with the ones from the $\ell_0$ method.
Today's tutorial

A couple of tips:
- collapse code/output cells

- Increase/reduce font size with Ctrl + +/−
- Back to normal size with ctrl + 0
1. Introduction to the compressed sensing methods

2. Get the data from the NOMAD Archive

3. Determining low-dimensional descriptors with the $\ell_0$ method

4. Approximations to the $\ell_0$ method

5. Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool

6. Predicting properties of new materials

7. The SISSO (Sure Independence Screening combined with Sparsifying Operator) for classification
1. Introduction to the compressed sensing methods

2. Get the data from the NOMAD Archive

3. Determining low-dimensional descriptors with the $\ell_0$ method

4. Approximations to the $\ell_0$ method

5. Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool

6. Predicting properties of new materials

7. The SISSO (Sure Independence Screening combined with Sparsifying Operator) for classification
Tutorial example 2-4: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features
- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence s orbital
- Radius of valence p orbital
- Radius of valence d orbital
- ...


$E(\text{Rocksalt}) - E(\text{Zinkblende})$
1. Introduction to the compressed sensing methods

2. Get the data from the NOMAD Archive

3. Determining low-dimensional descriptors with the $\ell_0$ method

4. Approximations to the $\ell_0$ method

5. Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool

6. Predicting properties of new materials

7. The SISSO (Sure Independence Screening combined with Sparsifying Operator) for classification
Tutorial example 5-6: predicting crystal structures from the composition

82 octet AB binary compounds

Ansatz: atomic features

- HOMO
- LUMO
- Ionization Potential
- Electron Affinity
- Radius of valence s orbital
- Radius of valence p orbital
- Radius of valence d orbital
- …?


\[ E(\text{Rocksalt}) - E(\text{Zinkblende}) \]
1. Introduction to the compressed sensing methods

2. Get the data from the NOMAD Archive

3. Determining low-dimensional descriptors with the $\ell_0$ method

4. Approximations to the $\ell_0$ method

5. Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool

6. Predicting properties of new materials

7. The SISSO (Sure Independence Screening combined with Sparsifying Operator) for classification
Tutorial example 7: predicting crystal structures from the composition
Today’s tutorial

Cross Validation, or, thou shalt not overfit

Graphical memes:

- Underfitting: 0.123 / 0.443
- Fitting: 0.044 / 0.068
- Overfitting: 0.036 / 0.939
Today’s tutorial

Cross Validation, or, thou shalt not overfit

Graphical memes:

Underfitting

Fitting

Overfitting

0.123 / 0.443

0.044 / 0.068

0.036 / 0.939

Error

Test set

Training set

Training/complexity
Cross Validation, or, thou shalt not overfit

Learn by example: Take leave-one-out cross validation (LOOCV):

From a dataset of $n$ elements
- label the elements from 1 to $n$ (arbitrary) and select element 1
- train the model for the remaining $n-1$
- test the prediction error on the left out element
- iterate $n$ times (every and each element is the left out)
- average over all $n$ iterations → prediction figure of merit
Learn by example: Take leave-one-out cross validation (LOOCV):

From a dataset of \( n \) elements
- label the elements from 1 to \( n \) (arbitrary) and select element 1
- train the model for the remaining \( n-1 \)
- test the prediction error on the left out element
- iterate \( n \) times (every and each element is the left out)
- average over all \( n \) iterations \( \rightarrow \) prediction figure of merit

Thou shalt not use any information about the left-out data point(s) for training the model
Learn by example: Take leave-one-out cross validation (LOOCV):

From a dataset of $n$ elements
- label the elements from 1 to $n$ (arbitrary) and select element 1
- train the model for the remaining $n-1$
- test the prediction error on the left out element
- iterate $n$ times (every and each element is the left out)
- average over all $n$ iterations $\rightarrow$ prediction figure of merit
Learn by example: Take leave-one-out cross validation (LOOCV):

From a dataset of $n$ elements
- label the elements from 1 to $n$ (arbitrary) and select element 1
- train the model for the remaining $n-1$
- test the prediction error on the left out element
- iterate $n$ times (every and each element is the left out)
- average over all $n$ iterations $\rightarrow$ prediction figure of merit

There are other, more stringent schemes (more elements are left out for test at each iteration), but LOOCV:
- enumerates all possible choice of training vs test. When data points are not so many, it allows for inspection, one by one.
- max errors are statistically meaningful
Today’s tutorial: the road map

1. **Introduction to the compressed sensing methods**
2. **Get the data from the NOMAD Archive**
   - See that the data from the repository can be accessed easily with simple python scripts.
   - Consider the specific case of 82 octet binary compounds in the RS and ZB structure. Process the data to build the target property $P$ of RS vs ZB energy differences.
   - Construct the descriptor matrix $D$ using primary features from the NOMAD atomic data collection.
3. **Determining low-dimensional descriptors with the $\ell_0$ method**
   - Perform an $\ell_0$-regularization to identify the best low-dimensional descriptors using the primary features.
   - Show that non-linear functions of the primary features improve the models significantly.
   - See that the $\ell_0$-regularization can rapidly become computational infeasible.
4. **Approximations to the $\ell_0$ method**
   - Perform a LASSO minimization and the LASSO+$\ell_0$ method.
   - Compare the solutions with the ones from the $\ell_0$ method.
5. **Visualization of the descriptors and experimenting with the LASSO+$\ell_0$ tool**
   - Reproduce the results from the reference publication by including further features.
   - Visualize the 2D descriptors in an interactive structure map.
   - Experiment with different settings and investigate the influence of the input parameters on the results. (OPTIONAL)
6. **Predicting properties of new materials**
   - Perform a leave-one-out cross-validation using the LASSO+$\ell_0$ method.
   - Analyze the prediction accuracy and how often the same descriptor is selected.
7. **The SISSO (Sure Independence Screening combined with Sparsifying Operator) for classification**
   - Run the SISSO code for classification.
   - Visualize the descriptor in an interactive 2D structure map.