Static Data Analysis and Mining with RAVEN

RAVEN workshop

www.inl.gov

Idaho National Laboratory

Outline

- Introduction
- Clustering Methods in RAVEN
- Dimensionality Reduction in RAVEN
- Clustering Example
- Dimensionality reduction example

Data Mining

Extraction of implicit, previously unknown and potentially useful information from data

Exploration and analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

- Alternative Names
  - Knowledge discovery (mining) in databases (KDD)
  - Knowledge extraction
  - Data/pattern analysis
  - Data archeology
  - Information harvesting

Why Data Mining?

Data mining is fairly new in the context considered here….

Opportunity

- During uncertainty quantification/sensitivity analysis lots of data is being collected and warehoused
- Computers and electronic storage are cheaper and faster

Needs

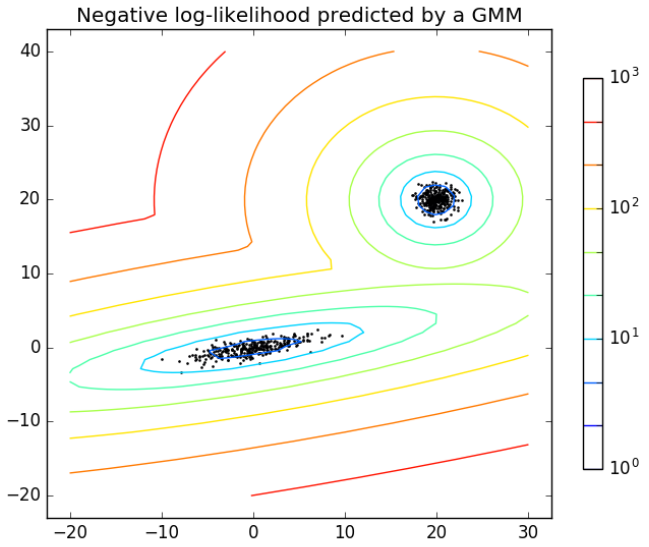
- Need to understand/gain knowledge on both input and output space
- Drowning in data but starving knowledge
- Extraction of interesting knowledge (rules, regularities, patterns, constraints) from data in large databases
- Answer to the question “Why, by whom?” uncertainty is generated

Clustering

- Automatically determining different groups in the data
- Useful for finding different regions in the output
- RAVEN implements a variety of methods:
  - Gaussian mixture models
  - K-Means
  - Affinity
  - Mean shift
  - Spectral clustering
  - DBSCAN

Gaussian Mixture Model

- Probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters
- It incorporate information about the covariance structure of the data as well as the centers of the latent Gaussians

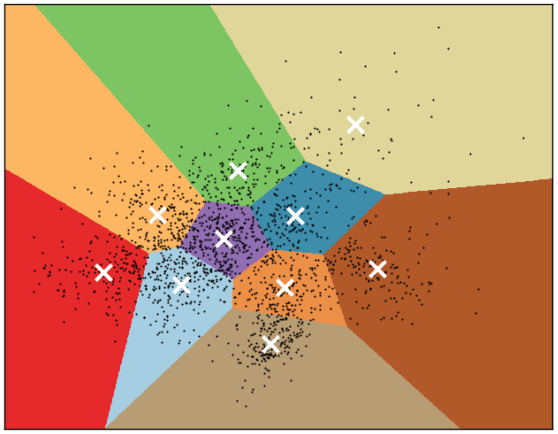


Two-component Gaussian mixture model: *data points, and equiprobability surfaces of the model [source sklearn]*

K-Means

- Method to cluster data by trying to separate samples in ***n*** groups of equal variance, minimizing a criterion known as the inertia
- The k-means algorithm divides a set of ***N*** samples ***X*** into ***K*** disjoint clusters ***C***, each described by the mean $\boldsymbol{\mu_i}$ of the samples in the cluster (centroids)
- The K-means algorithm chooses centroids that minimize the inertia:

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_j - \mu_i||^2)$$

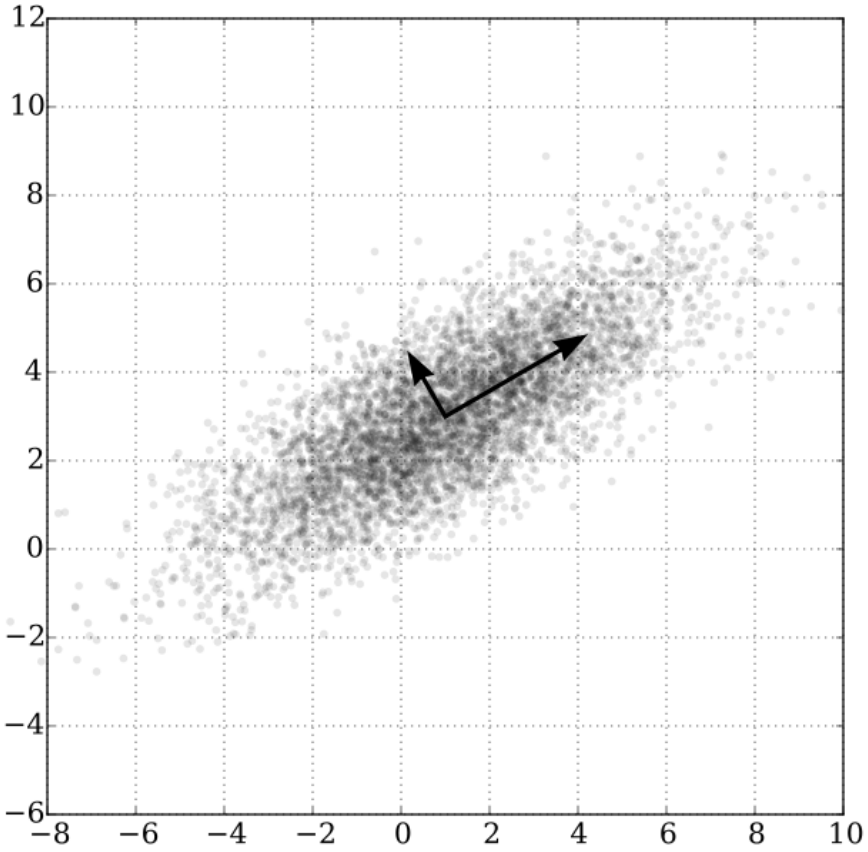


Dimensionality Reduction

- Used when datasets have many dimensions
- Used to avoid the “curse of dimensionality”
- Available methods in RAVEN
  - Principle Component Analysis
  - Truncated Singular Value Decomposition and Latent Semantic Analysis
  - Independent Component Analysis

Principal Component Analysis

- PCA is used to decompose a multivariate dataset in a set of successive orthogonal components that explain a maximum amount of the variance



RAVEN Example 1
Gaussian Mixture Clustering

RAVEN Example 1: Gaussian Mixture Clustering

- Steps
  1. Load data-set
  2. Post-Process the data
  3. Create a dataObject (PointSet) and plot the results

RAVEN Example 1: Gaussian Mixture Clustering

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Models>
 <PostProcessor name="GuassianMixtureBlobs" subType="DataMining">
   <KDD lib="SciKitLearn">
     <Features>x1,x2</Features>
     <SKLtype>mixture|GMM</SKLtype>
     <covariance_type>full</covariance_type>
     <n_components>3</n_components>
     <n_iter>10000</n_iter>
     <init_params>wc</init_params>
   </KDD>
 </PostProcessor>
</Models>
```

Input file name: test_dataMiningGaussianMixture.xml

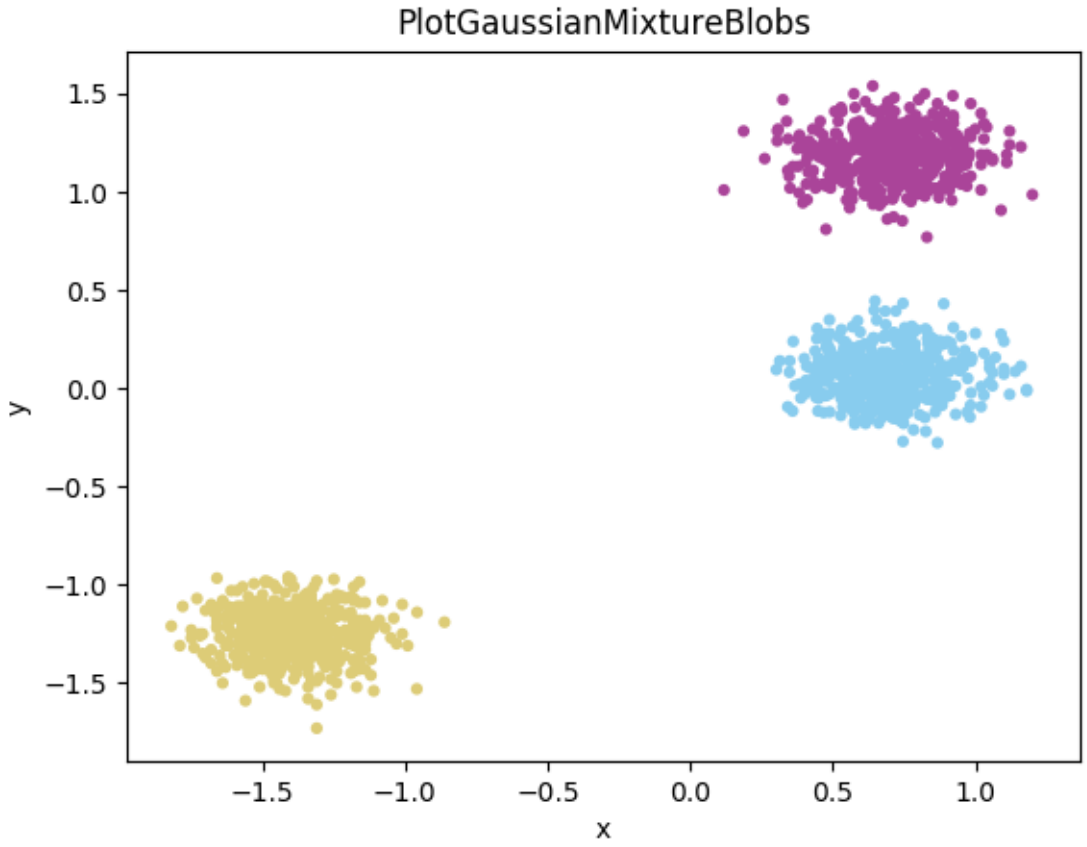
RAVEN Example 1: Gaussian Mixture Clustering

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Steps>
  <IOStep name="readIn">
    <Input     class="Files"          type=""          >DataSetsFile</Input>
    <Output    class="DataObjects"    type="PointSet" >DataSets</Output>
  </IOStep>
  <PostProcess name="GaussianMixtureBlobs">
    <Input     class="DataObjects"    type="PointSet"       >DataSets</Input>
    <Model     class="Models"         type="PostProcessor" >GaussianMixtureBlobs</Model>
    <Output    class="DataObjects"    type="PointSet"       >DataSets</Output>
    <Output    class="OutStreams"     type="Plot"           >Plotdata</Output>
  </PostProcess>
  <IOStep name="output">
    <Input     class="DataObjects"    type="PointSet"    >DataSets</Input>
    <Output    class="OutStreams"     type="Plot"        >PlotGaussianMixtureBlobs</Output>
  </IOStep>
</Steps>
```

Input file name: test_dataMiningGaussianMixture.xml

RAVEN Example 1: Gaussian Mixture Clustering



RAVEN Example 2
K-Means Clustering

RAVEN Example 2: K-Means Clustering

- Steps
  1. Load data-set (fuel performance)
  2. Post-Process the data
  3. Create a dataObject (PointSet) and plot the results

RAVEN Example 2: K-Means Clustering

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Models>
 <PostProcessor name="KMeans1" subType="DataMining">
   <KDD lib="SciKitLearn">
     <Features>output</Features>
     <SKLtype>cluster|KMeans</SKLtype>
     <n_clusters>5</n_clusters>
     <precompute_distances>True</precompute_distances>
     <init>k-means++</init>
   </KDD>
 </PostProcessor>
</Models>
```

Input file name: test_dataMiningKMeans.xml

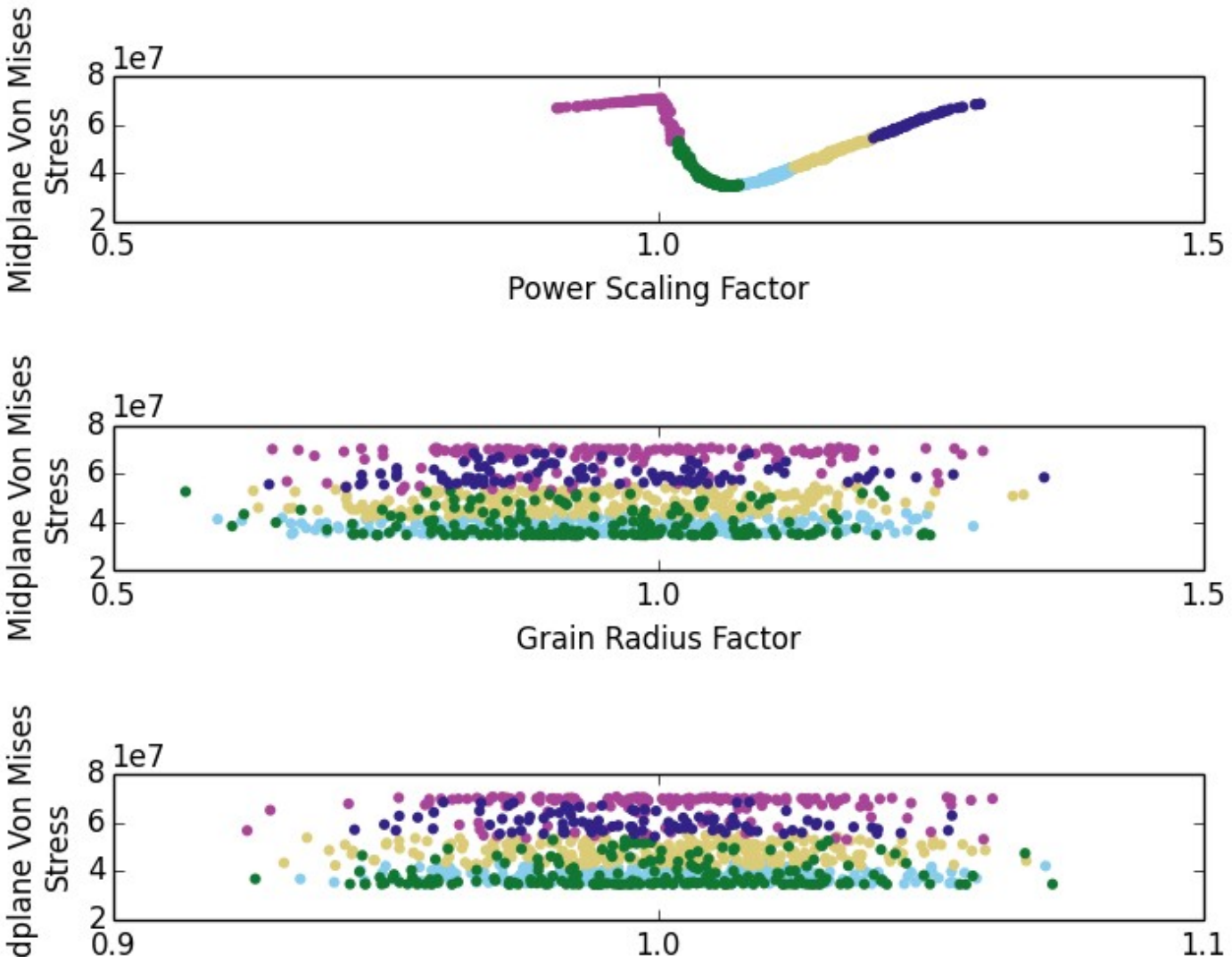
RAVEN Example 2: K-Means Clustering

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Steps>
  <IOStep name="readIn">
    <Input     class="Files"          type=""          >bisonDBCSV</Input>
    <Output    class="DataObjects"    type="PointSet" >bisonData</Output>
  </IOStep>
  <PostProcess name="GaussianMixtureBlobs">
    <Input     class="DataObjects"    type="PointSet"       >bisonData</Input>
    <Model     class="Models"         type="PostProcessor" >KMeans1</Model>
    <Output    class="DataObjects"    type="PointSet"       >bisonData</Output>
    <Output    class="OutStreams"     type="Plot"           >PlotKMeans1</Output>
    <Output    class="OutStreams"     type="Plot"           >PlotAll</Output>
    <Output    class="OutStreams"     type="Print"          >dump_data</Output>
  </PostProcess>
</Steps>
```

Input file name: test_dataMiningKMeans.xml

RAVEN Example 2: K-Means Clustering



RAVEN Example 3
PCA Dimensionality Reduction

RAVEN Example 3: Dimensionality Reduction

- Steps
  1. Load data-set (iris database)
  2. Post-Process the data
  3. Create a dataObject (PointSet) and plot the results

RAVEN Example 3: Dimensionality Reduction

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Models>
 <PostProcessor name="PCA" subType="DataMining">
   <KDD lib="SciKitLearn">
     <Features>x1,x2,x3,x4</Features>
     <SKLtype>decomposition|PCA</SKLtype>
     <n_components>2</n_components>
   </KDD>
 </PostProcessor>
</Models>
```

Input file name: test_dataMiningExactPCA.xml

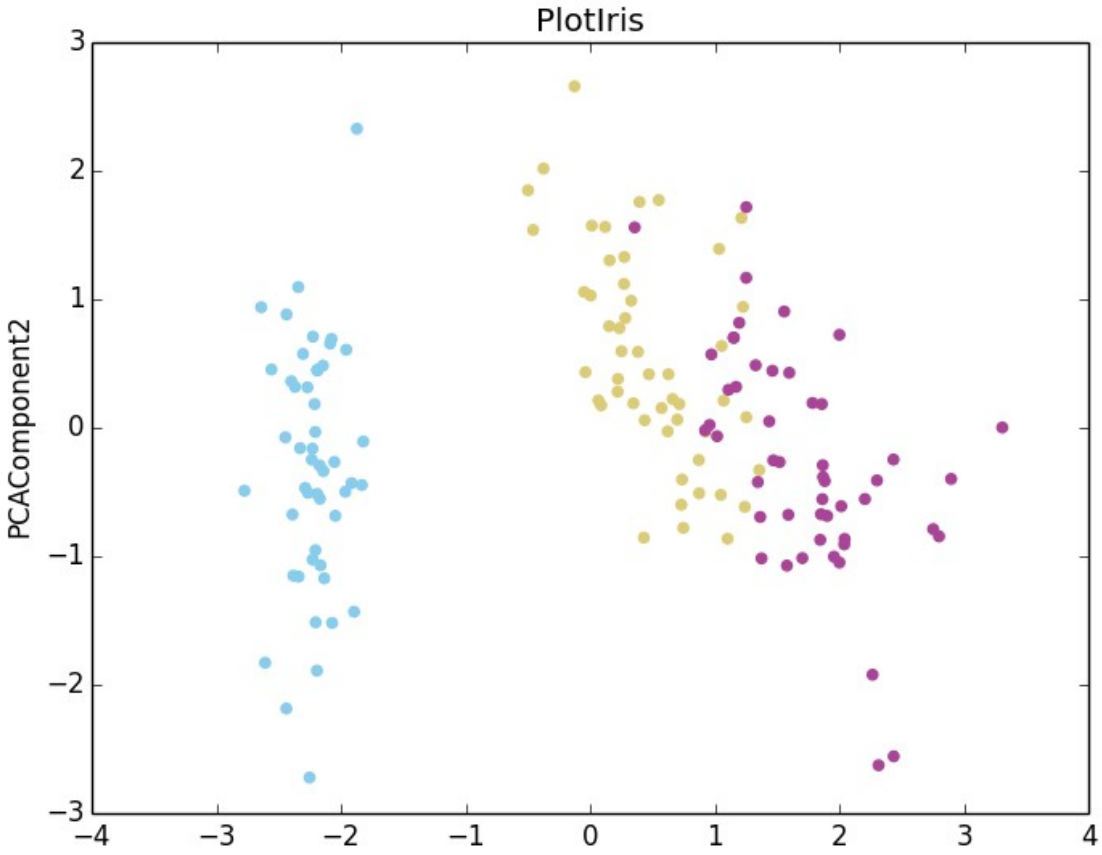
RAVEN Example 3: Dimensionality Reduction

Distributions	Models	Samplers	Databases	DataObjects	Steps

```
<Steps>
  <IOStep name="readIn">
    <Input     class="Files"          type=""          >DataSetsFile</Input>
    <Output    class="DataObjects"    type="PointSet"  >DataSets</Output>
  </IOStep>
  <PostProcess name="PCAIris">
    <Input     class="DataObjects"    type="PointSet"       >DataSets</Input>
    <Model     class="Models"         type="PostProcessor"  >PCA</Model>
    <Output    class="DataObjects"    type="PointSet"       >DataSets</Output>
    <Output    class="OutStreams"     type="Plot"           >Plotdata</Output>
  </PostProcess>
  <IOStep name="output">
    <Input     class="DataObjects"    type="PointSet"    >DataSets</Input>
    <Output    class="OutStreams"     type="Plot"        >PlotGaussianMixtureBlobs</Output>
  </IOStep>
</Steps>
```

Input file name: test_dataMiningExactPCA.xml

Exact PCA Dimensionality Reduction Example Output



Questions