DATA SCIENCE FOR NEUROSCIENCE

Pedro Larrañaga
Computational Intelligence Group
Universidad Politécnica de Madrid
Data science tools in neuroscience

1. Circular statistics
   • von Mises distribution

2. Spatial statistics
   • Complete spatial randomness
   • Random sequential adsorption

3. Clustering
   • Probabilistic

4. Classification
   • Paradigms
   • Evaluation

5. Bayesian networks
   • Conditional independence
   • Bayesian classifiers
   • Multi-dimensional classification

6. Optimization
   • Evolutionary computation
1. Circular statistics. Branching angles of dendritic trees

- Dendritic arbors of layer III pyramidal neurons
- Seven cortical areas: M1, M2, S1, S2, V1, V2, PrL/II

Bielza, Benavides-Piccione, López-Cruz, Larrañaga, DeFelipe (2014). Branching angles of pyramidal cell dendrites follow common geometrical design principles in different cortical areas. *Scientific Reports*, 4, article 5909
2. Spatial statistics. Distribution of synapses

- 1695 synaptic junctions from ten 3D samples
- Neuropil of layer III of the young rat somatosensory cortex

2. Spatial statistics. Distribution of synapses

Complete spatial randomness (CSR)

Model:

Homogeneous spatial Poisson point process, with intensity $\lambda > 0$:

- Number of points in any subregion $B$ follow a Poisson ($\lambda \cdot \text{vol}(B)$)
- Number of points in disjoint regions are independent random variables
- Locations in $B$ are i.i.d. and uniformly distributed

...but what if points occupy a volume and therefore cannot overlap?

Model:

Random sequential adsorption (RSA)
2. Spatial statistics. Distribution of synapses

3. Clustering

<table>
<thead>
<tr>
<th></th>
<th>Height</th>
<th>Angle</th>
<th>...</th>
<th>Eccentricity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.02</td>
<td>0.42</td>
<td></td>
<td>0.81</td>
</tr>
<tr>
<td>2</td>
<td>0.04</td>
<td>0.12</td>
<td></td>
<td>0.73</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>0.04</td>
<td>0.37</td>
<td></td>
<td>0.77</td>
</tr>
</tbody>
</table>

Similarity in multidimensional spaces

- Maximize homogeneity within clusters
- Maximize heterogeneity between clusters
3. Clustering of human pyramidal cell dendritic spines

- 3D reconstructed 2,000 spines from layer III pyramidal neuron human cingulate cortex (aged 40)
- Each spine is characterized with 54 morphological variables (some of them directional variables)

3. Clustering. Probabilistic

Finite mixture modeling

\[ f(x \mid \theta) = \sum_{k=1}^{K} \pi_k f_k(x \mid \theta_k) \]

\[ f_k(x \mid \theta_k) \sim \mathcal{N}(x \mid \mu_k, \Sigma_k) \]

3. Probabilistic clustering of human pyramidal cell dendritic spines

4. Classification. Interneuron classification

4. Classification. Interneuron classification

4. Classification. Interneuron classification

## 4. Classification. Interneuron classification

<table>
<thead>
<tr>
<th></th>
<th>(X_1=\text{Length})</th>
<th>(X_2=\text{Terminal degree})</th>
<th>(\ldots)</th>
<th>(X_{2,886}=\text{Partition asymmetry})</th>
<th>(C=\mathbf{F}_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2947.20</td>
<td>1648</td>
<td></td>
<td>9.67</td>
<td>Martinotti</td>
</tr>
<tr>
<td>2</td>
<td>3545.45</td>
<td>1942</td>
<td></td>
<td>5.33</td>
<td>Large basket</td>
</tr>
<tr>
<td></td>
<td>\vdots</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>240</td>
<td>1684.50</td>
<td>480</td>
<td></td>
<td>6.00</td>
<td>Chandelier</td>
</tr>
</tbody>
</table>

**Algorithm**

**Model**

**Question**

**Answer**

<table>
<thead>
<tr>
<th></th>
<th>(X_1=\text{Length})</th>
<th>(X_2=\text{Terminal degree})</th>
<th>(\ldots)</th>
<th>(X_{2,886}=\text{Partition asymmetry})</th>
<th>(C=\mathbf{F}_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>n+1</td>
<td>1619.28</td>
<td>1160</td>
<td></td>
<td>4.00</td>
<td>? Chandelier</td>
</tr>
</tbody>
</table>

4. Classification. Algorithms

K nearest neighbors

Classification trees

Naïve Bayes

Neural networks

Support vector machines

Random forest

4. Classification. Feature subset selection

**Filter approach**

All features → Search strategy → Feature subset → Feature subset assessment → Subset score → Best feature subset

**Wrapper approach**

All features → Search strategy → Feature subset → Classification algorithm → Estimated accuracy → Best feature subset

---
4. Classification. Figures of merit

Misclassification matrix (for binary classifiers)

\[
\begin{array}{c|c|c}
\text{Predicted class} & \text{True} & \text{False} \\
\hline
\text{True positive} & TP & \text{False positive} \\
\text{False negative} & FN & \text{True negative} \\
\end{array}
\]

Accuracy: \[
\frac{TP + TN}{TP + TN + FP + FN}
\]

Sensitivity: \[
\frac{TP}{TP + FN}
\]

Specificity: \[
\frac{TN}{TN + FP}
\]

F1 - Score: \[
\frac{2TP}{2TP + FP + FN}
\]

Area under the ROC curve

4. Classification. Validation: No honest

\[ \hat{p}_M = \frac{1}{n} \sum_{i=1}^{n} \delta(c^{(i)} = c^{(i)}_M) \]
4. Classification. Validation: k-fold cross-validation

\[ \hat{\mathbf{p}}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{\mathbf{p}}_i \]

4. Classification. Interneuron classification

2,886 morphological predictors, 240 neurons

<table>
<thead>
<tr>
<th>Feature 1: Intralaminar vs. Translaminar</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>57.68</td>
<td>58.51</td>
<td>77.59</td>
<td>82.16*</td>
<td>72.2</td>
<td>73.44</td>
<td>82.57*</td>
<td>85.48*</td>
<td>82.16*</td>
<td>75.93</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>64.73</td>
<td>54.36</td>
<td>79.67</td>
<td>82.99*</td>
<td>69.71</td>
<td>75.93</td>
<td>83.82*</td>
<td>85.48*</td>
<td>84.23*</td>
<td>79.67</td>
</tr>
<tr>
<td>CfsSubset</td>
<td>75.93</td>
<td>75.1</td>
<td>81.33</td>
<td>84.23*</td>
<td>73.86</td>
<td>80.08</td>
<td>84.65*</td>
<td>80.08</td>
<td>82.16*</td>
<td>80.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature 2: Intracolumnar vs. Transcolumnar</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>59.75</td>
<td>62.66*</td>
<td>52.28</td>
<td>75.52*</td>
<td>57.68</td>
<td>65.56*</td>
<td>74.27*</td>
<td>68.46*</td>
<td>66.39*</td>
<td>58.09*</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>66.39</td>
<td>63.07*</td>
<td>53.11</td>
<td>76.35*</td>
<td>64.32*</td>
<td>65.98*</td>
<td>75.52*</td>
<td>68.88*</td>
<td>70.12*</td>
<td>65.98*</td>
</tr>
<tr>
<td>CfsSubset</td>
<td>72.61</td>
<td>65.56*</td>
<td>76.76*</td>
<td><strong>81.33</strong>*</td>
<td>73.86*</td>
<td>73.03*</td>
<td>74.69*</td>
<td>70.54*</td>
<td>76.35*</td>
<td>69.29*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature 3: Centered vs. Displaced</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>62.24</td>
<td>53.94</td>
<td>54.77</td>
<td><strong>68.88</strong>*</td>
<td>64.73*</td>
<td>68.05*</td>
<td>66.8*</td>
<td>67.63*</td>
<td>68.46*</td>
<td>62.24</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>64.73</td>
<td><strong>73.03</strong>*</td>
<td>65.98*</td>
<td>70.54*</td>
<td>65.56*</td>
<td>71.37*</td>
<td>70.54*</td>
<td>66.39*</td>
<td>72.2*</td>
<td>68.46*</td>
</tr>
<tr>
<td>CfsSubset</td>
<td>68.88*</td>
<td><strong>73.86</strong>*</td>
<td>70.54*</td>
<td>73.03*</td>
<td>65.15*</td>
<td>68.05*</td>
<td>63.9*</td>
<td>71.78*</td>
<td>68.46*</td>
<td>65.15*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature 4: Ascending vs. Descending vs. Both</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>34.44</td>
<td>27.8</td>
<td>44.4*</td>
<td>49.38*</td>
<td>41.91</td>
<td>38.59</td>
<td>33.61</td>
<td><strong>54.36</strong>*</td>
<td>40.25*</td>
<td>37.76</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>43.57*</td>
<td>33.2</td>
<td>43.98*</td>
<td><strong>49.79</strong>*</td>
<td>41.91</td>
<td>42.32</td>
<td>43.57*</td>
<td>46.89*</td>
<td>45.64*</td>
<td>42.74</td>
</tr>
<tr>
<td>CfsSubset</td>
<td>47.3*</td>
<td><strong>51.87</strong>*</td>
<td>47.3*</td>
<td>58.51*</td>
<td>47.3*</td>
<td>52.28*</td>
<td>48.13*</td>
<td>42.32</td>
<td><strong>60.17</strong>*</td>
<td>47.3*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature 5: Interneuron type (10 classes)</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>56.02*</td>
<td>19.09</td>
<td>45.23*</td>
<td><strong>58.51</strong>*</td>
<td>50.62*</td>
<td>53.94*</td>
<td>50.62*</td>
<td>47.72*</td>
<td>52.28*</td>
<td>40.25*</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td>60.17*</td>
<td><strong>26.14</strong>*</td>
<td>58.92*</td>
<td><strong>62.24</strong>*</td>
<td>49.79*</td>
<td>51.87*</td>
<td>48.55*</td>
<td>43.15*</td>
<td>58.09*</td>
<td>43.98*</td>
</tr>
<tr>
<td>CfsSubset</td>
<td>61*</td>
<td>43.57*</td>
<td><strong>61.41</strong>*</td>
<td>60.58*</td>
<td>58.09*</td>
<td>56.85*</td>
<td>53.94*</td>
<td>49.38*</td>
<td>56.85*</td>
<td>51.45*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature 6: Characterized vs. Uncharacterized</th>
<th>NB</th>
<th>NBdisc</th>
<th>RBFN</th>
<th>SMO</th>
<th>IB1</th>
<th>IB3</th>
<th>JRip</th>
<th>J48</th>
<th>RForest</th>
<th>RTree</th>
</tr>
</thead>
<tbody>
<tr>
<td>NoFSS</td>
<td>77.18</td>
<td>88.38</td>
<td>95.85</td>
<td><strong>97.93</strong>*</td>
<td>97.51</td>
<td>97.51</td>
<td>97.51</td>
<td>97.51</td>
<td>96.27*</td>
<td>95.85</td>
</tr>
<tr>
<td>Gain Ratio</td>
<td><strong>98.34</strong>*</td>
<td>73.86</td>
<td>97.51</td>
<td>96.68</td>
<td>97.1</td>
<td>97.51</td>
<td>97.93*</td>
<td>97.51</td>
<td><strong>98.34</strong>*</td>
<td></td>
</tr>
<tr>
<td>CfsSubset</td>
<td>97.51</td>
<td>89.63</td>
<td>96.27</td>
<td>97.1</td>
<td>95.44</td>
<td>95.02</td>
<td>97.93*</td>
<td>96.27</td>
<td>97.51</td>
<td><strong>99.17</strong>*</td>
</tr>
</tbody>
</table>

5. Bayesian networks. Factorization of the JPD

\[
p(A, N, S, D, P) = p(A)p(N|A)p(S|A)p(D|N, S)p(P|S)
\]

5. Bayesian networks. Reasoning via inference

5. Bayesian networks. Reasoning via inference

Evidence: “Stroke = yes”

5. Bayesian networks. Reasoning via inference

Evidence: “Stroke = yes, Neuronal Atrophy = yes”

Evidence: “Stroke = yes, Neuronal Atrophy=yes, Age= young”

5. Bayesian networks. Modeling and simulation of pyramidal cells

- 3D reconstructions of 90 pyramidal cells
- Mouse neocortex layer III

5. Bayesian networks. Bayesian classifiers

Example for $k$-DB with $k = 2$

\[
P(c|x_1, x_2, x_3, x_4, x_5) \propto P(c)P(x_1|x_3, c)P(x_2|x_1, x_5, c)P(x_3|c)P(x_4|x_1, x_3, c)P(x_5|x_1, x_4, c)
\]
5. Bayesian networks. Knowledge discovery in Alzheimer’s disease with \(k\)-DBs

- 12 tissue samples (6 AD brains + 6 control brains) from the entorhinal cortex
- 7610 variables measuring gene expression level
- Induce many \(k\)-DBs by a resampling method (bootstrap)
- Output a Bayesian network structure with those arcs above a reliability threshold \(t\): only arcs occurring more than \(t\) times are retained

5. Bayesian networks. Knowledge discovery in Alzheimer’s disease with k-DBs

5. Bayesian networks. Knowledge discovery in Alzheimer’s disease with $k$-DBs

5. Bayesian networks. Knowledge discovery in Alzheimer’s disease with $k$-DBs

5. Bayesian networks. Knowledge discovery in Alzheimer’s disease with $k$-DBs


PDQ-39 and EQ-5D: quality of life instruments to measure the degree of disability in PD

39-item Parkinson’s Disease Questionnaire: a specific instrument

PDQ-39 captures patient’s perception of his illness covering 8 dimensions:

1. Mobility
2. Activities of daily living
3. Emotional well-being
4. Stigma
5. Social support
6. Cognitions
7. Communication
8. Bodily discomfort


<table>
<thead>
<tr>
<th>European Quality of Life - 5 Dimensions: a generic instrument</th>
</tr>
</thead>
<tbody>
<tr>
<td>EQ-5D is a generic measure of health for clinical and economic appraisal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems in walking about</td>
</tr>
<tr>
<td>I have some problems in walking about</td>
</tr>
<tr>
<td>I am confined to bed</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Self-care</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems with self-care</td>
</tr>
<tr>
<td>I have some problems washing and dressing myself</td>
</tr>
<tr>
<td>I am unable to wash and dress myself</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Usual activities (eg. work, study, housework, family or leisure activities)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no problems with performing my usual activities</td>
</tr>
<tr>
<td>I have some problems with performing my usual activities</td>
</tr>
<tr>
<td>I am unable to perform my usual activities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pain/discomfort</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have no pain or discomfort</td>
</tr>
<tr>
<td>I have moderate pain or discomfort</td>
</tr>
<tr>
<td>I have extreme pain or discomfort</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Anxiety/depression</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am not anxious or depressed</td>
</tr>
<tr>
<td>I am moderately anxious or depressed</td>
</tr>
<tr>
<td>I am extremely anxious or depressed</td>
</tr>
</tbody>
</table>


![Mapping PDQ-39 to EQ-5D](image)

\[ h : (PDQ_1, ..., PDQ_{39}) \rightarrow (EQ_1, ..., EQ_5) \]


<table>
<thead>
<tr>
<th>Method</th>
<th>Mean accuracy</th>
<th>Global accuracy</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MB–MBC</td>
<td>0.7119 ± 0.0338</td>
<td>0.2030 ± 0.0718</td>
<td>0.0650 ± 0.0156</td>
</tr>
<tr>
<td>CB–MBC</td>
<td>0.6807 ± 0.0285</td>
<td>0.1865 ± 0.0429</td>
<td>0.0905 ± 0.0167</td>
</tr>
<tr>
<td>Indep–MB–HITON</td>
<td>0.7009 ± 0.0427</td>
<td><strong>0.2051 ± 0.0835</strong></td>
<td>0.0699 ± 0.0188</td>
</tr>
<tr>
<td>Indep–MB–PC</td>
<td>0.6587 ± 0.0636</td>
<td>0.1867 ± 0.0937</td>
<td>0.0909 ± 0.0909</td>
</tr>
<tr>
<td>MNL</td>
<td>0.6926 ± 0.0430</td>
<td>0.1802 ± 0.0713</td>
<td>0.0759 ± 0.0152</td>
</tr>
<tr>
<td>OLS</td>
<td>0.4201 ± 0.0252</td>
<td>0.0123 ± 0.0046</td>
<td>0.1832 ± 0.0373</td>
</tr>
<tr>
<td>CLAD</td>
<td>0.4254 ± 0.0488</td>
<td>0.0143 ± 0.0171</td>
<td>0.1962 ± 0.0360</td>
</tr>
</tbody>
</table>

\[ d = 5, N = 488 \]

- **Mean accuracy** over the \( d \) class variables: 
  \[
  \text{Acc}_m = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{N} \sum_{l=1}^{N} \delta(\hat{c}_l, c_l)
  \]

- **Global accuracy** over the \( d \)-dimensional class variable: 
  \[
  \text{Acc}_g = \frac{1}{N} \sum_{l=1}^{N} \delta(\hat{c}_l, c_l)
  \]

- **MSE** between the true and predicted EQ-5D utility scores
6. Optimization. Optimal wiring of pyramidal cells

- Analyze the neuronal wiring of 288 3D reconstructed complete basal arborizations of pyramidal cells
- Layers II, III, IV, Va, Vb and VI of the hindlimb somatosensory cortical región of P14 rats

6. Optimization. Optimal wiring of pyramidal cells

6. Optimization. Optimal wiring of pyramidal cells

**Fig 2. Mean wiring length (real vs. optimized).** Mean wiring length (µm) of the 48 analyzed cells in each cortical layer (red) versus mean wiring length of the shortest arborizations found by our optimization algorithm for each layer (green). The optimization algorithm found an equal or slightly better (shorter) wiring for all the neurons in all the layers. We found the biggest difference with respect to the real wiring in layer Va, where the synthetic wiring was, on average, 2.06% shorter than the real wiring. The smallest difference occurred in layer IV, where the optimized wiring was, on average, 1.01% shorter than the real wiring.
Computational Intelligence Group

http://cig.fi.upm.es/
DATA SCIENCE FOR NEUROSCIENCE

Pedro Larrañaga
Computational Intelligence Group
Universidad Politécnica de Madrid