Supervised vs. unsupervised learning

Protein sequence space

Prior knowledge - the same class despite low sequence similarity
Supervised classification problem

1. Training data (we need examples to learn from)
2. Learning (how to learn from examples)
3. Validation (how to find trade off between accuracy and generalization)
**Training data:**

- A collection of records (objects) \( \mathbf{x} \). Each record contains a set of features and the class \( C \) that it belongs to.

<table>
<thead>
<tr>
<th>Age</th>
<th>tumor-size</th>
<th>inv-nodes</th>
<th>irradiat</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>23</td>
<td>12</td>
<td>yes</td>
<td>recurrence-events</td>
</tr>
<tr>
<td>32</td>
<td>12</td>
<td>3</td>
<td>yes</td>
<td>no-recurrence-events</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2</td>
<td>no</td>
<td>no-recurrence-events</td>
</tr>
<tr>
<td>45</td>
<td>3</td>
<td>6</td>
<td>yes</td>
<td>recurrence-events</td>
</tr>
</tbody>
</table>

\[ \{ \mathbf{x}_i, C_i \} \quad i = 1 \ldots N \]
How to choose feature space?

weights vs. heights

weights vs. estrogen

weights vs. testosteron

- adults
- kids
Learning:

- Find a model $y(x;w)$ that describes the objects of each class as a function of the features and adaptive parameters (weights) $w$.

$$Er(C_i,C_j) = \begin{cases} 
0, & \text{if } i = j \\
1, & \text{if } i \neq j
\end{cases}$$
What is the best model: accuracy vs. generalization

- Find a model $y(x;w)$ that avoids overfitting – too high accuracy on the training set may result in poor generalization (classification accuracy on new instances of the data).
Algorithms for supervised learning

- **LDA/FDA** (Linear/Fisher Discriminate Analysis) (simple linear cuts, kernel non-linear generalizations)
- **SVM** (Support Vector Machines) (optimal, wide margin linear cuts, kernel non-linear generalizations)
- Decision trees (logical rules)
- **k-NN** (k-Nearest Neighbors) (simple non-parametric)
- Neural networks (general non-linear models, adaptivity, “artificial brain”)

K-means clustering for unsupervised pattern discovery

- Choose the number of clusters (k), choose randomly their centers.
- Compute the mean (or median) vector for all items in each cluster.
- Reassign items to the cluster whose center is closest to the item, iterate the above two steps.
- Problems: spherical clusters, low “noise” tolerance, local minima:
Error function to be minimized.
**k-NN vs. k-means**

or supervised vs. unsupervised pattern discovery.

**Critical decisions to be made:**

1. Do I want to utilize a prior knowledge e.g. about functionally related genes? *Yes -> k-NN, No -> k-means*
2. What similarity measure (metric) to choose?
3. Which $k$ is best? Try different $k$'s and compare the results!
4. When to stop optimization (*k-means*)? Try re-running $k$-means several times!
5. Are the results significant? Use cross validation and biological knowledge!