

11/20/2024

- Quiz 3: Molecular Simulations
- Last analyses in Exercise 12
- Quantitative Structure-Property Relationships

Quantitative Structure-Property Relationships

- Today's lecture is intended to help you achieve the following learning objective: Develop a model for a quantitative structure-property relationship. Describe its domain of applicability.
- QSPR Topics
 - Machine learning
 - Features/descriptors
 - Algorithms
 - Validation
 - Limitations

QSPR is an application of machine learning

- Machine learning [1]
 - “gives computers the ability to learn without being explicitly programmed” - Arthur Samuel (1959)
 - “A computer program is said to learn from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.” -Tom Mitchell (1997)
- In Quantitative Structure-Property Relationships (QSPR)
 - Experience E: chemicals
 - Task T: predicting physical property, e.g.
 - boiling point, chromatography retention times
 - activity against a biological target = QSAR
 - signaling efficacy (#s for agonist, antagonist, partial agonist, inverse agonist)
 - ADMET (absorption, distribution, metabolism, and excretion - toxicity)
 - Performance P: correlation/error in validation set/real-world applications

Features/Descriptors

- Enables chemicals to be input into machine learning models
- How would you describe a chemical in numbers that can be put into a mathematical formula?
- Features can be
 - from enumerable properties, e.g. number of a certain element, number of H bond donors/acceptors, presence/absence of an element functional group
 - based on 3D structures
 - of a chemical, e.g. surface area, radius of gyration, moments of inertia
 - of a protein-ligand complex
- <http://www.rdkit.org/docs/GettingStartedInPython.html#list-of-available-descriptors>

Algorithms

- Linear regression
 - $y = mx + b$ for simple linear regression
 - $y = m_1x_1 + m_2x_2 + \dots + b$ for multiple linear regression
 - x is projected onto a new space in partial least squares regression
- Neural networks
 - sets of nodes that transform inputs into an output
 - allows for nonlinear relationships.
- Deep learning
 - neural networks with multiple layers
 - increasingly popular and powerful with faster computers and larger datasets
- Not an exhaustive list!

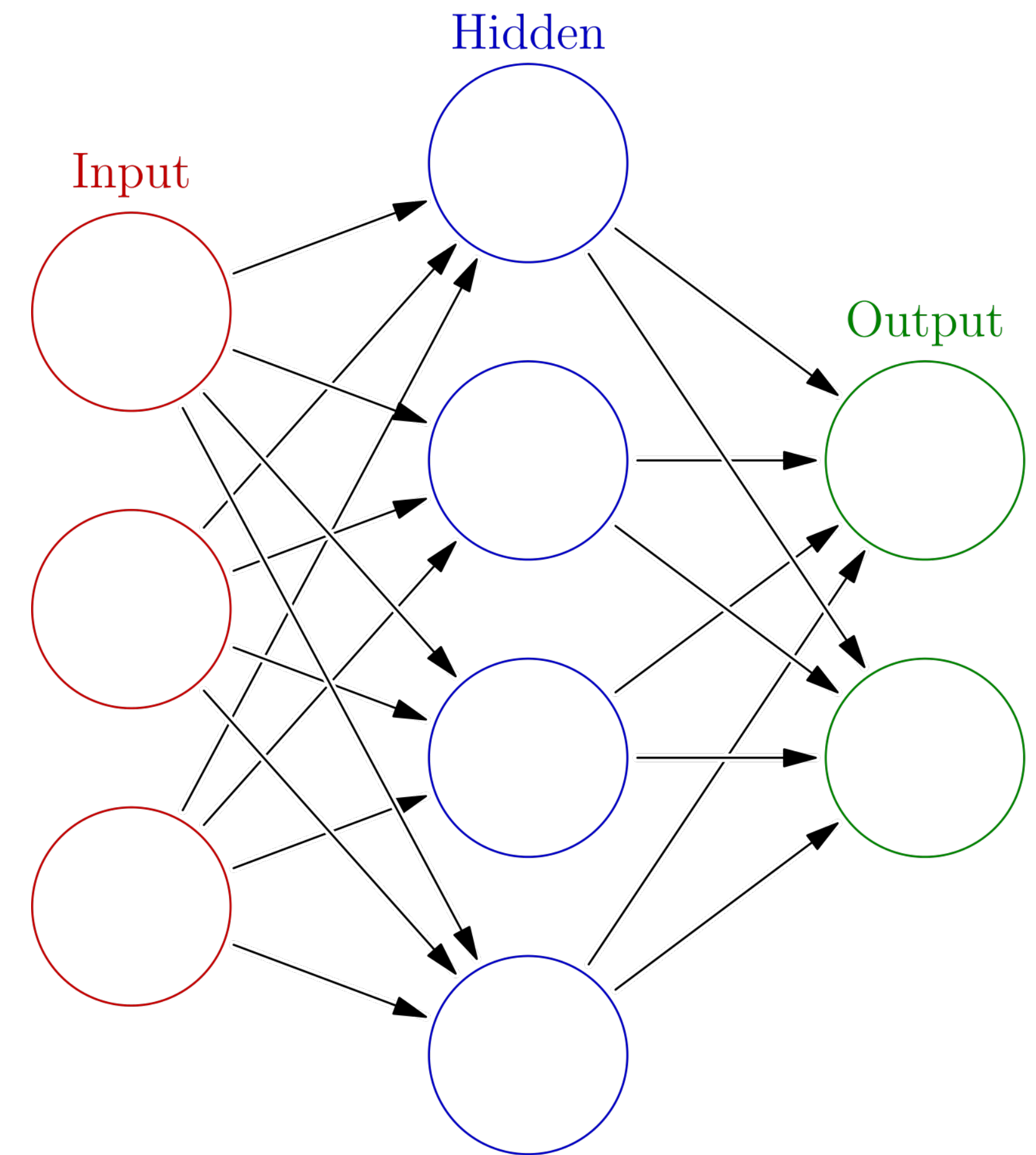


Diagram of an artificial neural network [2]

Validation

- How do you know if a model is any good?
- Data are split into training and test sets
 - Both sets should represent the same population
- Model is built using the training set
- Calculations on test set are compared to data

Limitations

- What do you think some limitations of QSPR might be?
- A high-quality training set is necessary
- Models can only be applied to molecules similar to the training set
- Sometimes small changes in structure can lead to large changes in a property
 - adding a methyl can make it so that a ligand no longer fits into a binding pocket
 - in QSAR, known as an “activity cliff”

References

- [1] <https://www.geeksforgeeks.org/introduction-machine-learning/>
- [2] Downloaded from https://commons.wikimedia.org/wiki/File:Colored_neural_network.svg and reused under the [CC BY-SA 3.0](https://creativecommons.org/licenses/by-sa/3.0/) license