Machine Learning

Application of AI/Machine Learning to Biological Problems

What is machine learning?

- We want to create predictive models for properties
- If we have known, annotated training data: supervised learning
- If not, we have to perform unsupervised learning
- Models generally are parametric systems of equations
- Parameters are tuned during training

What are challenges in machine learning?

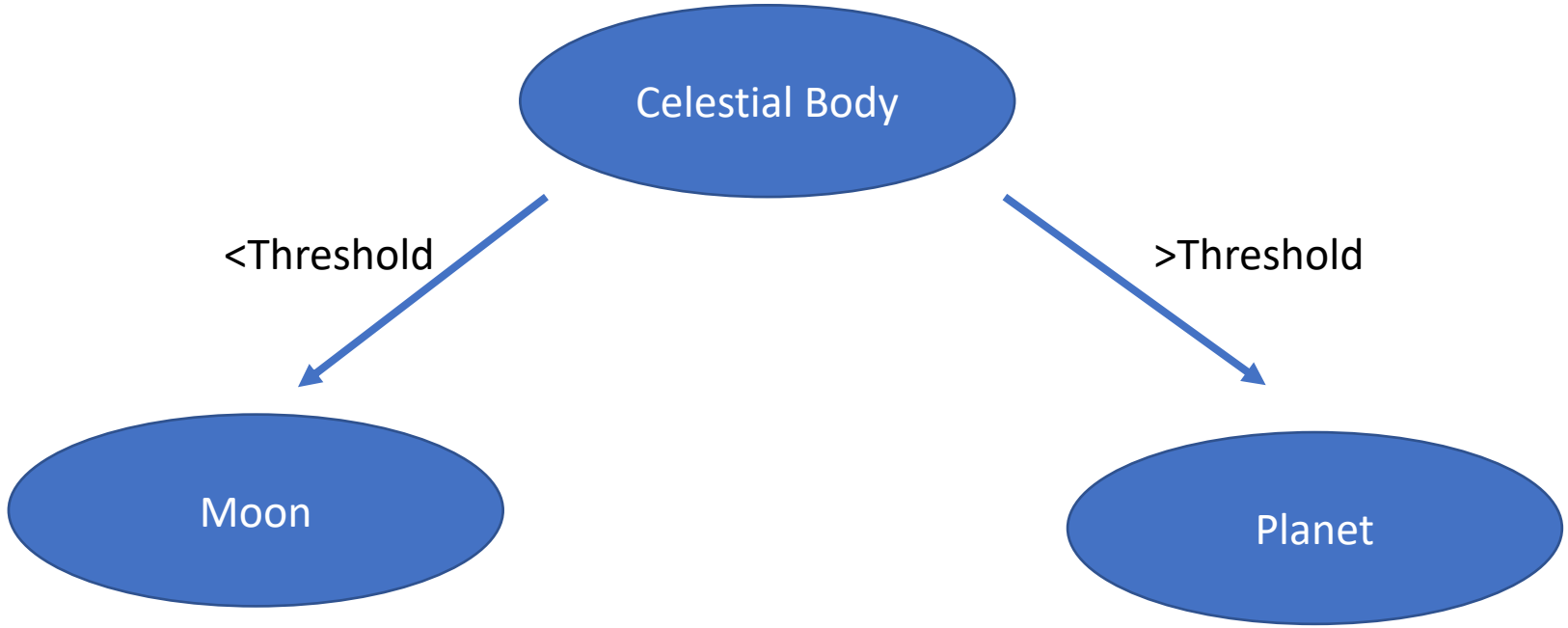
- Availability of training data (e.g. only a few data points are known)
- Input data is often messy: missing data, incorrect data (e.g. wrong order of magnitude), irrelevant data
- Models have many parameters, often more parameters than data points (this is particularly true for CNNs)
- Overfitting to the training set is a serious concern

How does (supervised) machine learning work?

- Classification: Yes?/No?, What category?, Multiple Categories?
- Regression: Predict a target value

- We iteratively change our parameters and accept better (more predictive) models
- We need a measure for the quality of our model
- This measure is usually called a loss function

A simple ML model



Loss Functions

- Classification
  - Balanced Accuracy
  - True positive rate
  - Sensitivity
  - etc…
- Regression
  - Mean unsigned error (MUE)
  - Mean squared error
  - etc…

Optimization

- Generally, we have a high-dimensional plane (dimension equal to number of parameters + 1) that represents the value loss function at each set of parameters
- We want to find the parameters that minimize the loss function
- Usual numerical optimization strategies, e.g. steepest descent, conjugate gradient, etc…

Hyperparameters

- Hyperparameters describe your model, whereas parameters live within a model
- E.g. different kernel functions, different cutoffs, etc…

- Can also be systematically optimized, but each attempt needs a full training session of the underlying model, so computationally expensive
- The choice of model and its hyperparameters are key to get best performance
- Some models are better for specific conditions, e.g. sparse data, binary vs multi-class classification etc…

Machine Learning

Today’s Problems

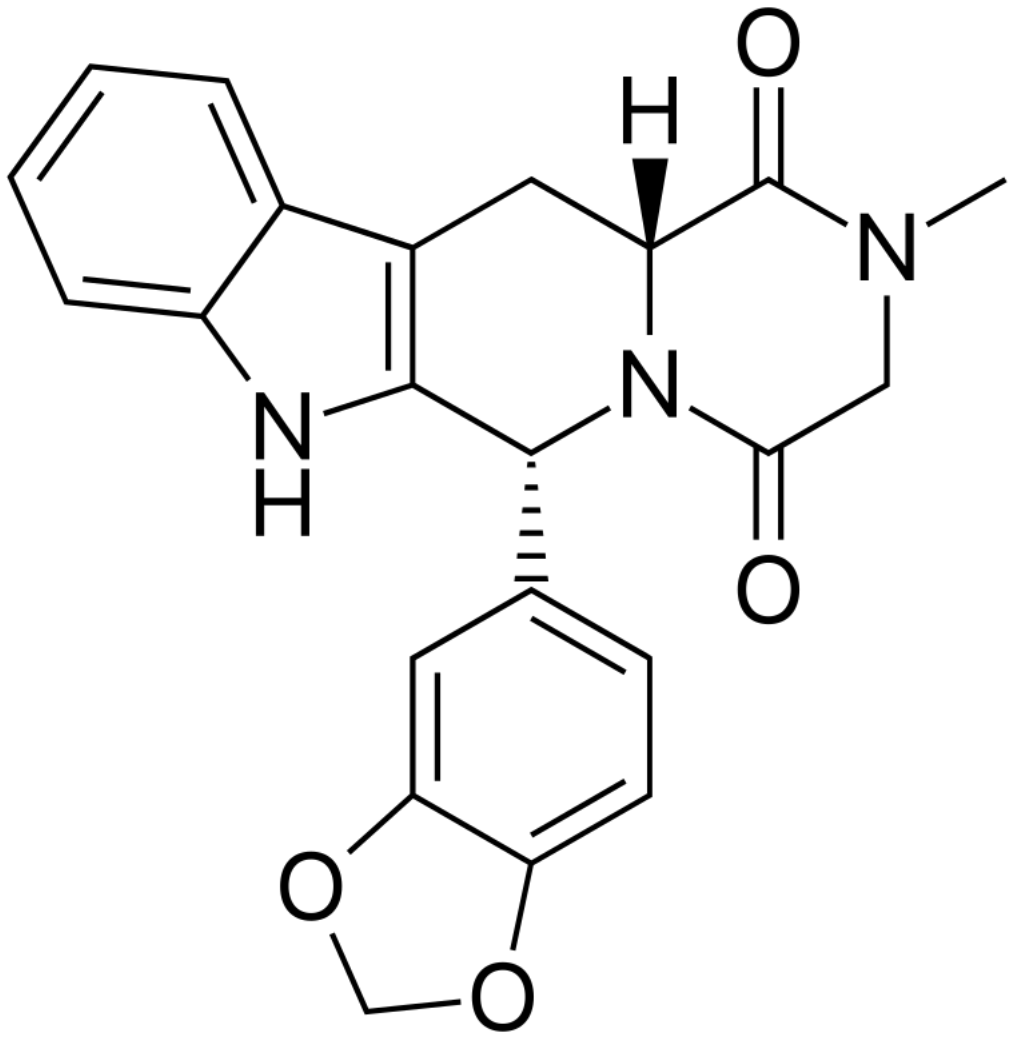
Scikit-Learn

- Scikit-Learn (sklearn) implements many machine learning models
  - Supervised Learning
  - Unsupervised Learning
  - Classification
  - Regression

Predicting Molecular Properties

- Useful for lead optimization, identification of drug compounds, toxicity etc.
- We are going to try to predict CNS permeability and logP for molecules
- We will use molecular descriptors as an input

What are molecular descriptors?



Machine Learning

Code at: https://github.com/rghuber/ml_class