

Modeling data

Model objectives

- **Regression:** Relationship between variables.
- **Classification:** Segregate data into discrete classes.
- **Clustering:** Segregate data into discrete clusters.
- **Dimensionality reduction:** Simplify data (often for visualization).
- **Encode/Decode:** Input/output transformation.
- **Simulation:** Representation of real system.

Model optimization

The values of the parameters in a model are typically optimized by minimizing some measure of error between the model's prediction and the data being modeled. Familiarizing yourself with the topics below will provide you a conceptual foundation for understanding what is involved in optimizing a model's parameters.

- Gradient descent
- Local vs global optimization

Model selection and generalization

One of the most important concepts in machine learning is that the model that most closely predicts a dataset is not necessarily the best choice model (in fact, it is often not the best). The reason is that you don't want your model to overfit the noise in the data. A model that overfits noise will have an excellent prediction for the dataset it was optimized on, but it will not generalize well to new data with new random noise. This concept is referred to as the *bias vs. variance tradeoff*. Familiarizing yourself with the topics below will provide you a solid foundation for selecting a model that generalizes to new data or that is only as complex as needed.

- Cross validation
- Bayesian information criterion (BIC)
- Silhouette score