Introduction to Data Science

Data Science Essentials

NASHVILLE
SOFTWARE
SCHOOL

Goals for today

- Review last session coding tasks
- Intro to Machine Learning

Review last session coding tasks

week4_review notebook

Machine Learning

A “set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data”.

Rather than using designed rules, have the model automatically determine how to predict by showing past examples.

Can be applied to problems for which classical methods are not well-suited (eg. large, high-dimensional data sets such as images).

Often the focus is on **prediction** instead of hypothesis-driven **inference**.

Types of Machine Learning

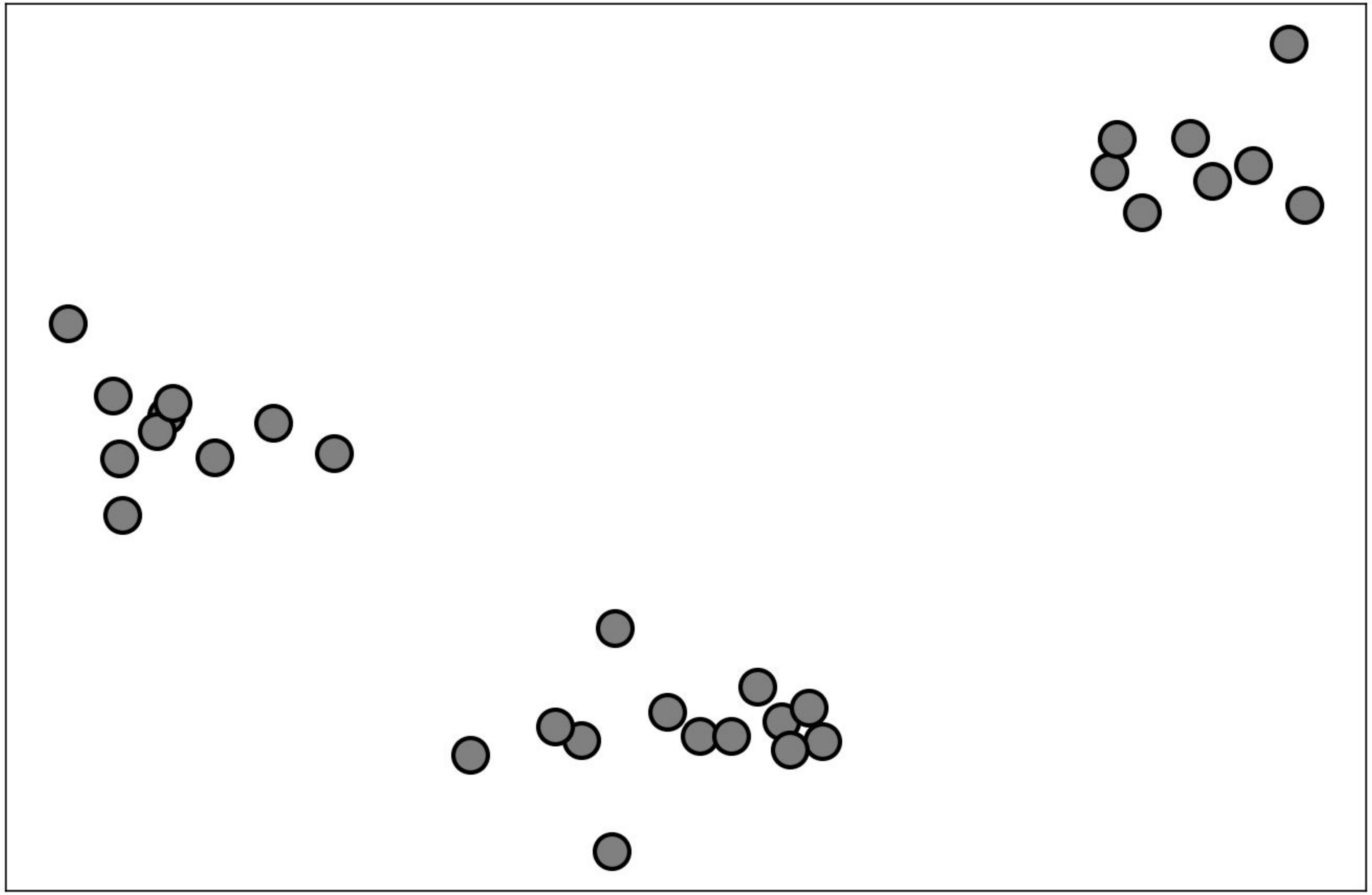
Supervised Learning:

- Goal: Predict a target variable, given a set of predictors
- Requires a labeled set of data in order to fit a model
- Example: Learning to classify email as spam or not spam based on a set of labeled training examples

Unsupervised Learning:

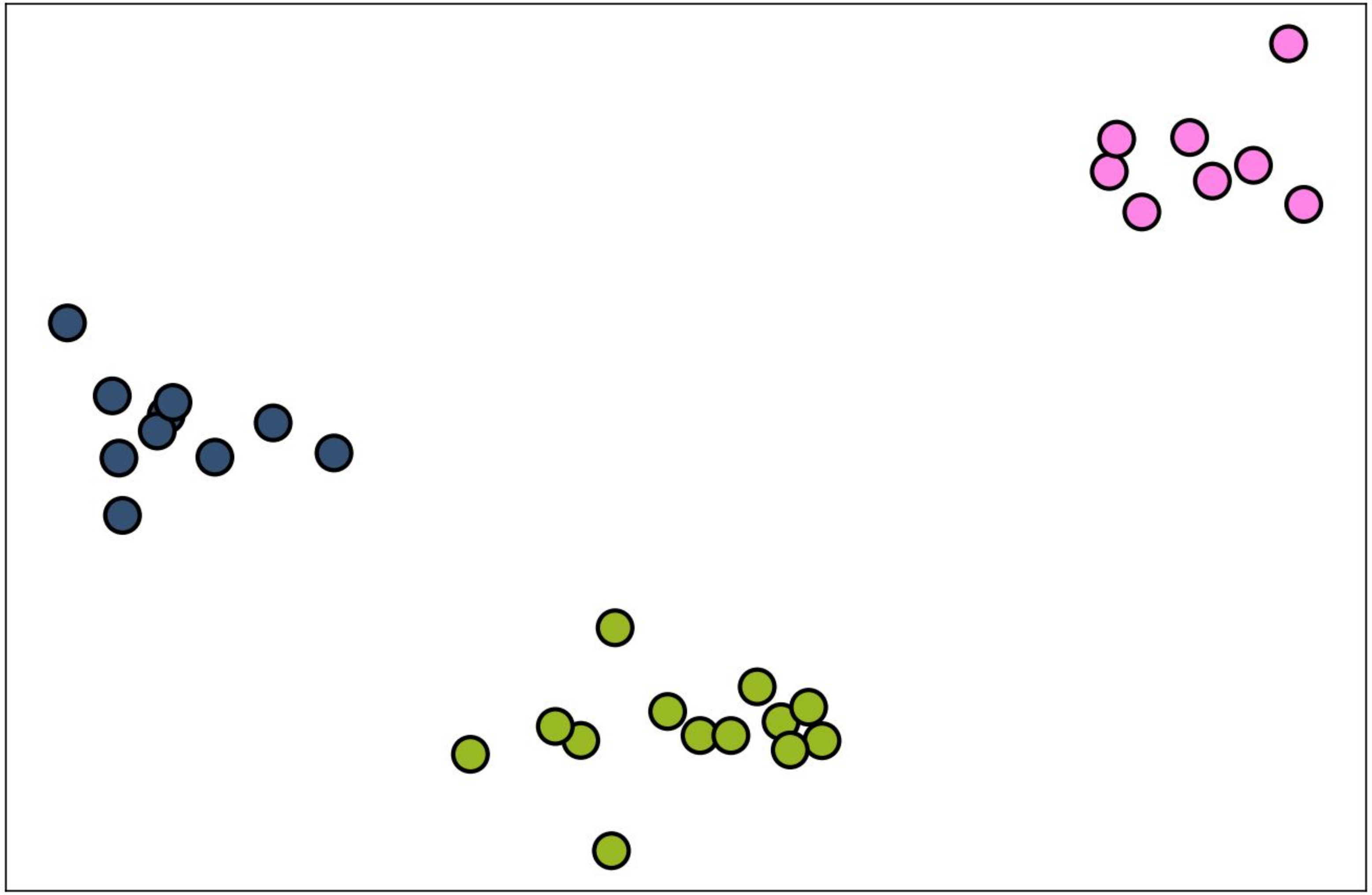
- Goal: Uncover natural relationships/groupings within the data
- Examples: Clustering, Mixture Models

Unsupervised Learning



Can be used to uncover groupings or clusters in a dataset.

Unsupervised Learning



Can be used to uncover groupings or clusters in a dataset.

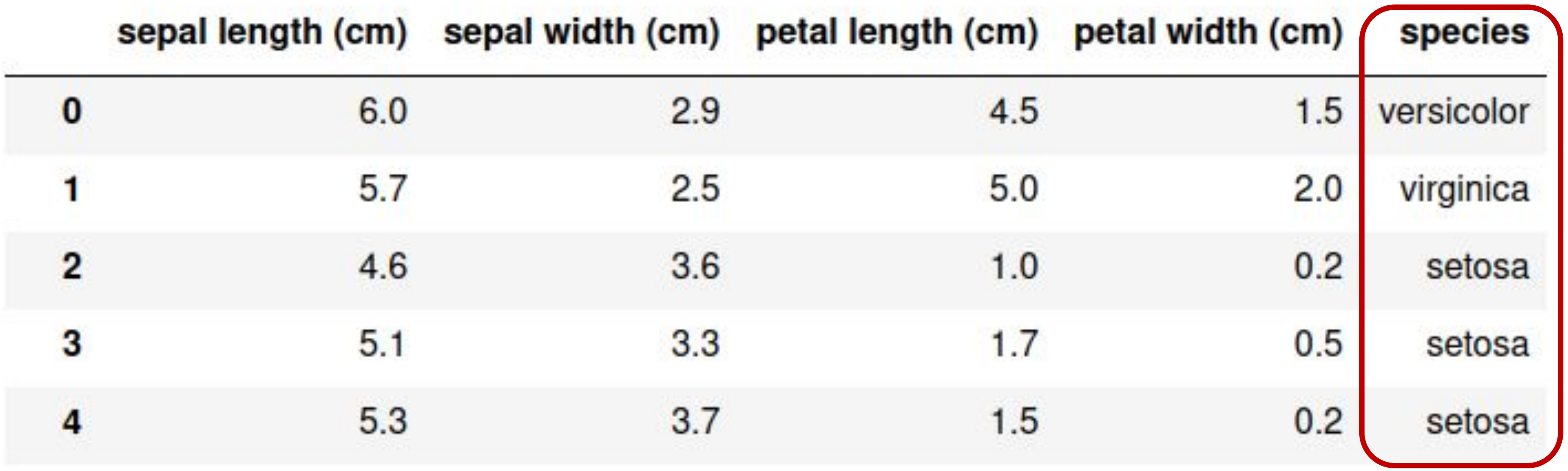
Supervised Learning

You ***supervise*** your computer as it learns by giving it known outcomes for a sample of explanatory variables. This involves creating **labeled training data**.

Some common supervised learning algorithms are **linear regression**, **logistic regression**, **classification**, **support vector machines**, and **decision trees**.

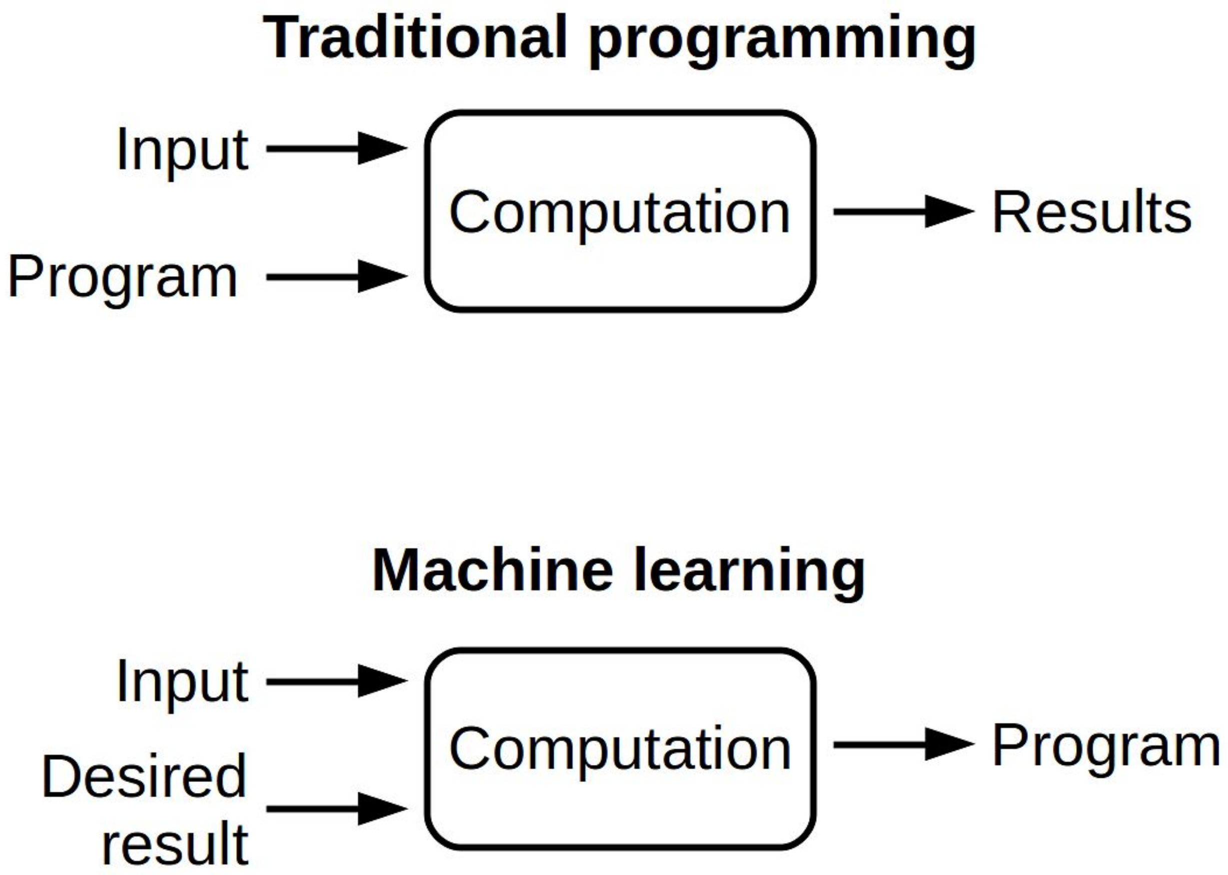
target variable

----------------predictor variables----------------



	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)	species
0	6.0	2.9	4.5	1.5	versicolor
1	5.7	2.5	5.0	2.0	virginica
2	4.6	3.6	1.0	0.2	setosa
3	5.1	3.3	1.7	0.5	setosa
4	5.3	3.7	1.5	0.2	setosa

Supervised Learning



Traditional programming

Input

Program

Computation

Results

Machine learning

Input

Desired result

Computation

Program

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Goal: Predict the probability (*p*) of the target variable being True.

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Goal: Predict the probability (*p*) of the target variable being True.

How: Predict the log-odds of *p* as a linear function of your predictor variables:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Goal: Predict the probability (*p*) of the target variable being True.

How: Predict the log-odds of *p* as a linear function of your predictor variables:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

For example, we might predict whether an iris is of the setosa species using this formula:

$$\log \frac{p}{1-p} = 9.7 - 0.37 \cdot (\text{sepal length}) - 2.9 \cdot (\text{petal length})$$

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Goal: Predict the probability (*p*) of the target variable being True.

How: Predict the log-odds of *p* as a linear function of your predictor variables:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

⇕

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}}$$

Logistic Regression

Can be used when we have a binary (True/False) target variable.

Goal: Predict the probability (*p*) of the target variable being True.

How: Predict the log-odds of *p* as a linear function of your predictor variables:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k$$

⇕

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_k \cdot x_k)}}$$

Fitting this model means that we use an algorithm to determine acceptable values for the coefficients (the beta values).

Supervised Learning Steps

1. Preprocess data
   a. Dealing with missing values
   b. Handling categorical variables
   c. Separating into training and test sets
   d. Scaling (sometimes)
2. Fit model
   a. Selecting model type
   b. Choosing hyperparameters
3. Evaluate model on test set
4. Iterate the model building process to improve performance

Questions?