CISC 5800 HW1

Question 1. Information Gain

Calculate information gain and prove whether 'Is Soft' is the best first split.

H(S) = Entropy[2, 2, 1, 1, 1] = 2.2360

For **Outside Color**:

H(S | A) = 0.6793

Gain = H(S) - H(S | A) = 1.5567

For **Inside Color**:

H(S | A) = 0.7872

Gain = H(S) - H(S | A) = 1.4488

For **Size**:

H(S | A) = 0.3936

Gain = H(S) - H(S | A) = 1.8424

For **Is Soft**:

H(S | A) = 1.2507

Gain = H(S) - H(S | A) = 0.9853

The best first split should be the feature with the largest information gain. In this dataset, it's **'Size'**, not 'Is Soft'.

Question 2. Concepts review

1. Label

Feature and label are two important concept in machine learning.

Generally, feature is the input data; while label are the output.

Take the dataset from slides as the example, the **outside color**, **inside color**, **weight**, **size** and **is soft** are the features. And **type** is the label.

2. Classification vs. Regression

- *From the prediction angle:*

During most situations, classification is used to get a discrete output, like a label or category. Correspondingly, regression is used to get a continuous output, like a real value.

I think they are kind of have the similar nature, and sometimes you can easily transform those two types of predictions. For example, when you use **Linear Regression** to get a vector as the output, which may look like: w * x + b. This is definitely a continuous value and is regression. But when you take this result using **Logistic Regression**, you would get a binary classification.

- *From the training angle:*
  They have a different target function. For the classification, there's log loss, hinge loss. For the regression, there's square loss.

3. Supervised vs. Unsupervised vs. Semi-supervised

- For supervised learning models, they have the input (x1, x2, ……,xn) and the output (y1, y2, ……, yn). The output would be either a real number in regression or a class label in classification.

- For the unsupervised learning, they only have the input (x1, x2, ……, xn). There's no default output. Based on different methods, we may get completely different result.

- For the semi-supervised learning, they have function estimation on both labeled and unlabeled data. This approach is motivated by the fact that labeled data is often costly to generate, whereas unlabeled data is generally not.

4. Precision Recall Curve vs. AUC Curve

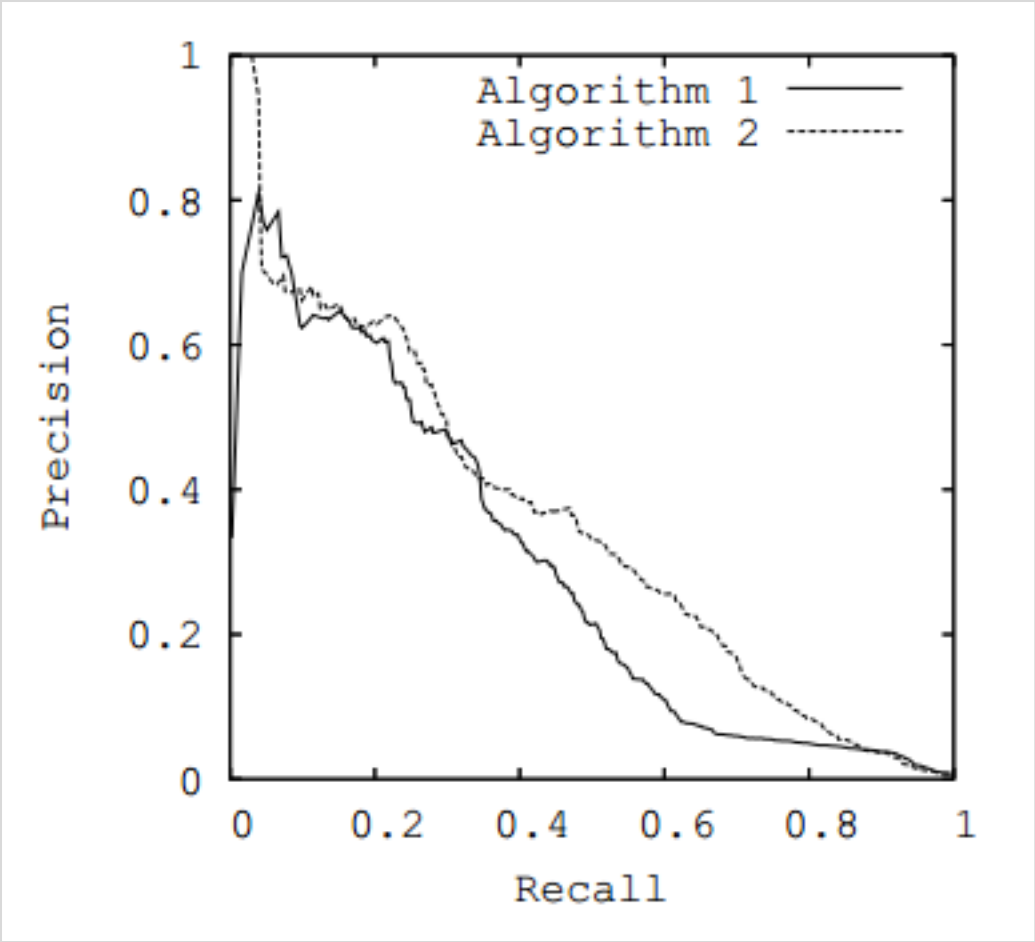
- Precision-Recall curves summarize the trade-off between the true positive rate and the positive predictive value for a predictive model using different probability thresholds.
  A PR curve is plotting Precision against Recall.
  Precision is defined as:

$$\text{Precision} = \frac{tp}{tp + fp}$$

  Recall is defined as:

$$\text{Recall} = \frac{tp}{tp + fn}$$

  A typical PR curve looks like this, which shows two PR curves for Algorithm 1 and Algorithm 2.



The goal is to have a model be at the upper right corner,which is basically getting only the true positives with no false positives and no false negatives – a perfect classifier.
The precision recall area under curve (PR AUC) is just the area under the PR curve. The higher it is, the better the model is.

- ROC Curves summarize the trade-off between the true positive rate and false positive rate for a predictive model using different probability thresholds.

5. Overfit vs Underfit

- Overfitting occurs when the model or the algorithm fits the data too well. Specifically, overfitting occurs if the model or algorithm shows low bias but high variance.

- Underfitting occurs when the model or the algorithm does not fit the data well enough. Specifically, underfitting occurs if the model or algorithm shows low variance but high bias. Underfitting is often a result of an excessively simple model.

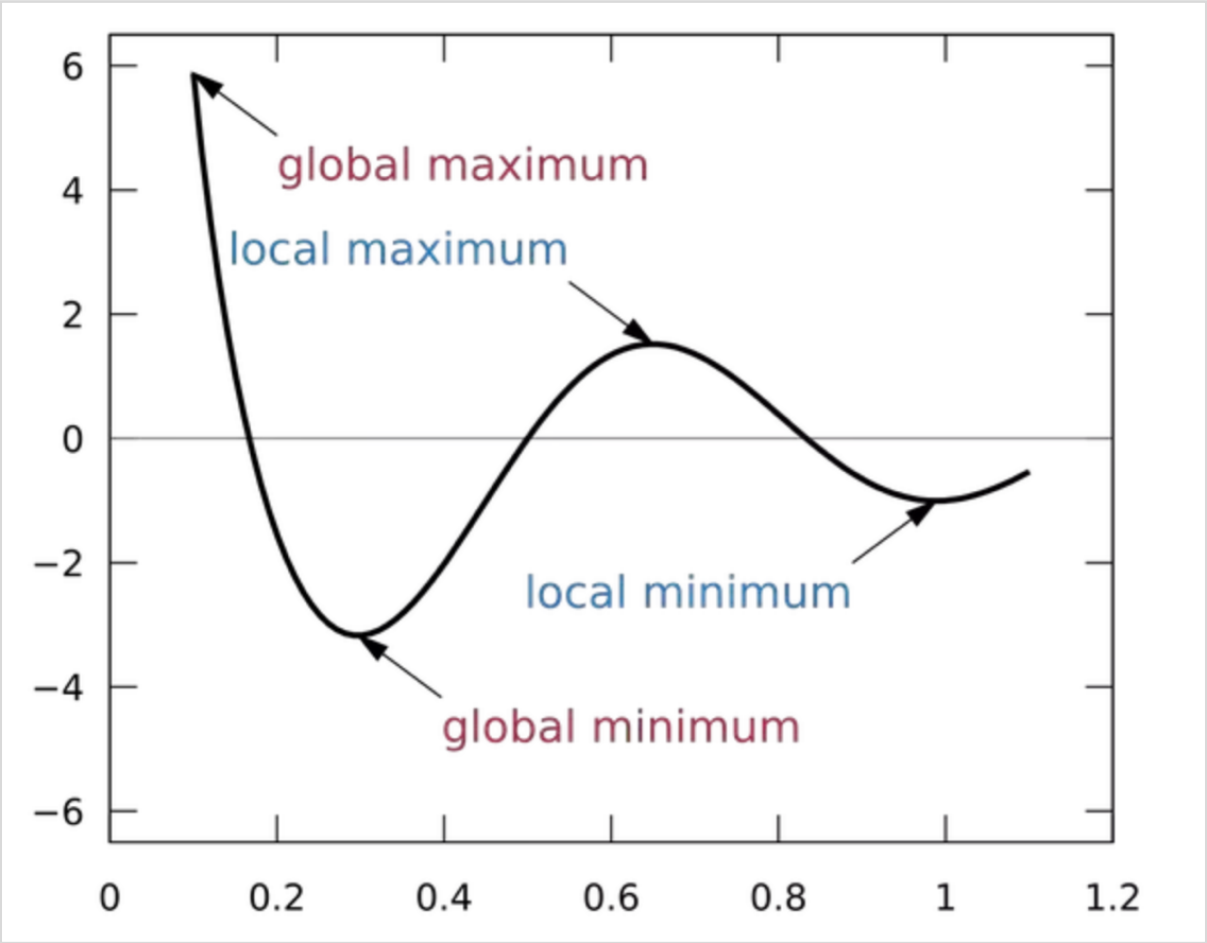
6. Converge

Convergence refers to a state reached during training in which training loss and validation loss change very little or not at all with each iteration after a certain number of iterations. In other words, a model reaches convergence when additional training on the current data

will not improve the model. In deep learning, loss values sometimes stay constant or nearly so for many iterations before finally descending, temporarily producing a false sense of convergence.

7. Local Minimum

A local minimum of a function (typically a cost function in machine learning, which is something we want to minimize based on empirical data) is a point in the domain of a function that has the following property: the function evaluates to a greater value at every other point in a neighborhood around the local minimum (a neighborhood in this case can correspond to a "ball" around the minimum) than the local minimum itself.

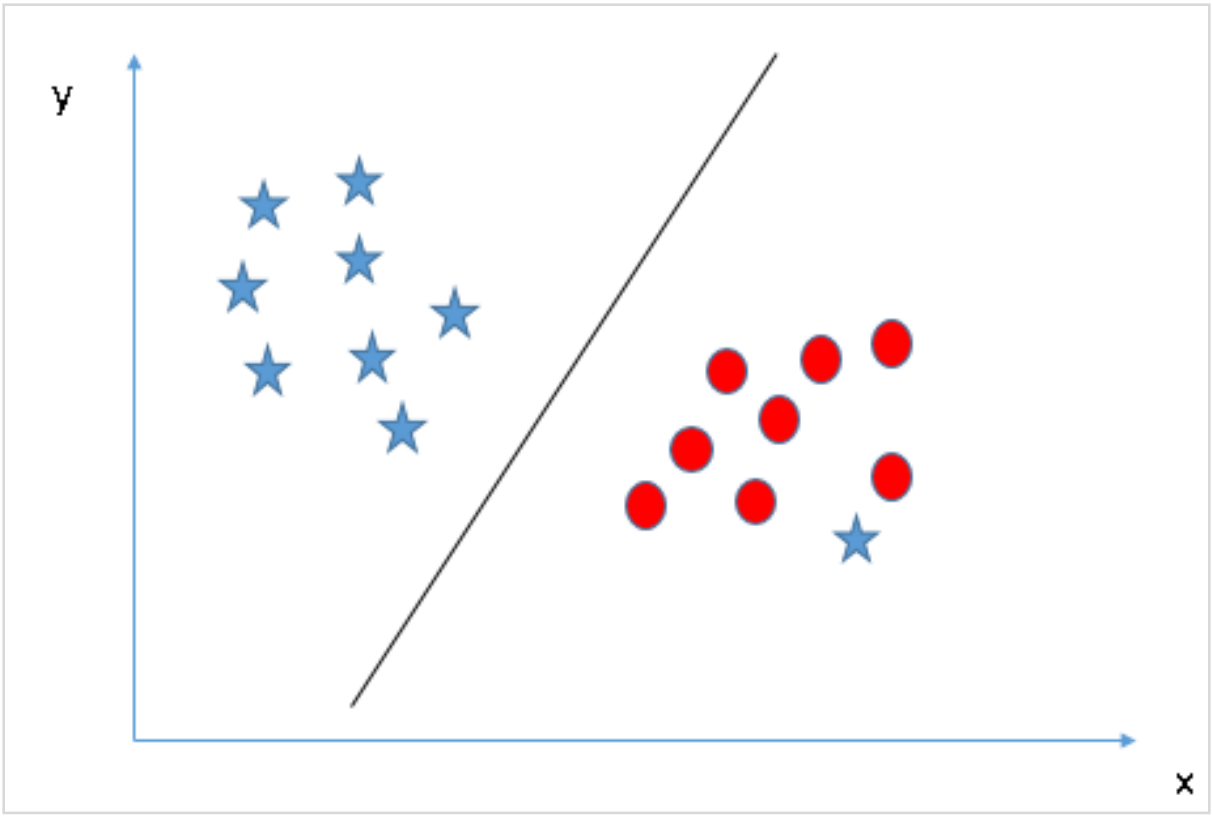


8. Linear vs Non-linear

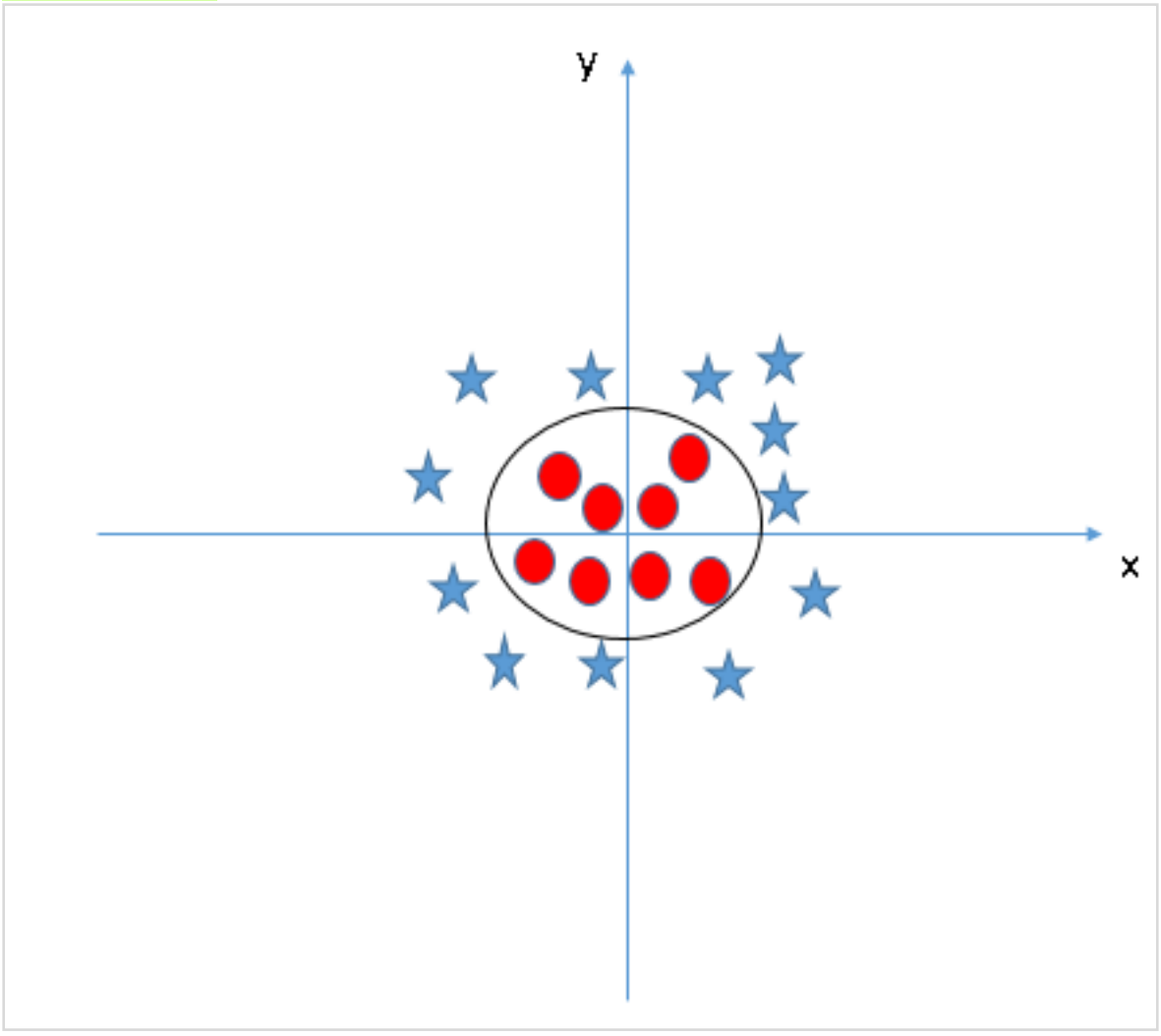
In the machine learning context, there are two focus point when we talk about linear and non-linear, one is the question, another one is the model.

- The linear and non-linear about the question refers to the distribution of sample data points. It's mainly about whether we can use a linear hyperplane to distinguish the input or not.

Linear Input



Non-linear Input



Examples on model angle below:

- Linear regression using L1 and L2 norm, which are also called Lasso and Ridge regression.
- SVM using linear and non-linear kernel function, like the RBF.
- Neural Networks. Each node in the network is a logistic model. But when we combine them, the combination is a non-linear model.

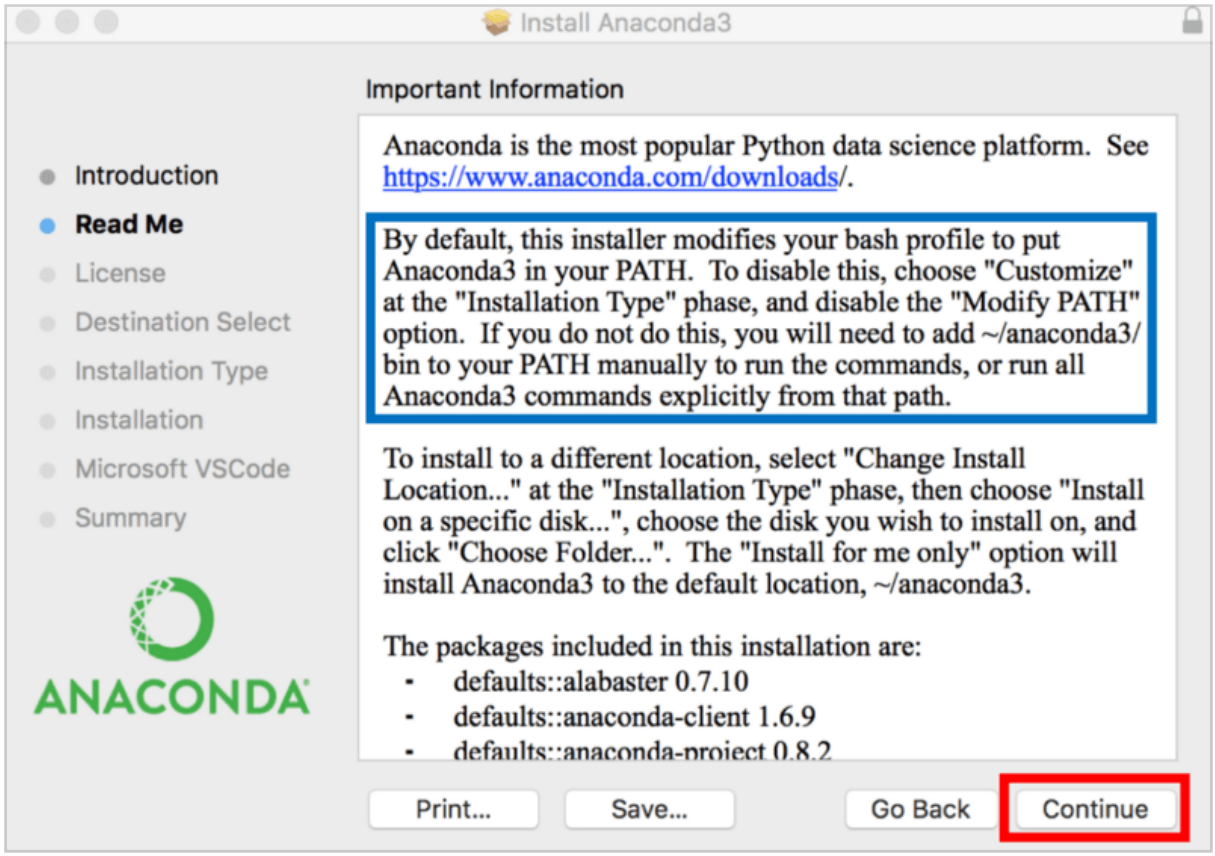
Extra Credit - Anaconda

OS and Hardware

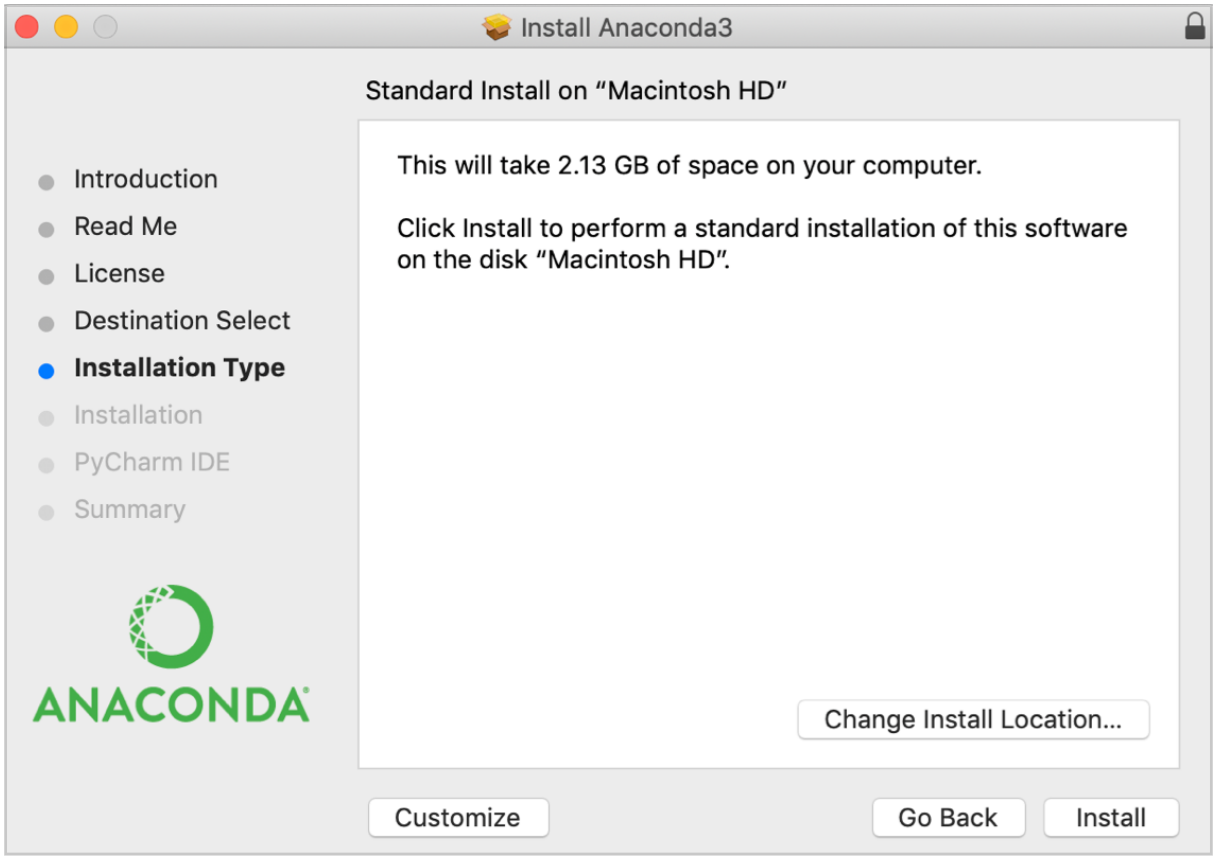
- Mac OS 10.14.5
- MacBook Pro 15 inch 2016 Fall

macOS Graphical Install

1. Download the graphical macOS installer for your version of Python.
2. OPTIONAL: Verify data integrity with MD5 or SHA-256 . For more information on hashes, see What about cryptographic hash verification? .
3. Double-click the downloaded file and click continue to start the installation.
4. Answer the prompts on the Introduction, Read Me, and License screens.

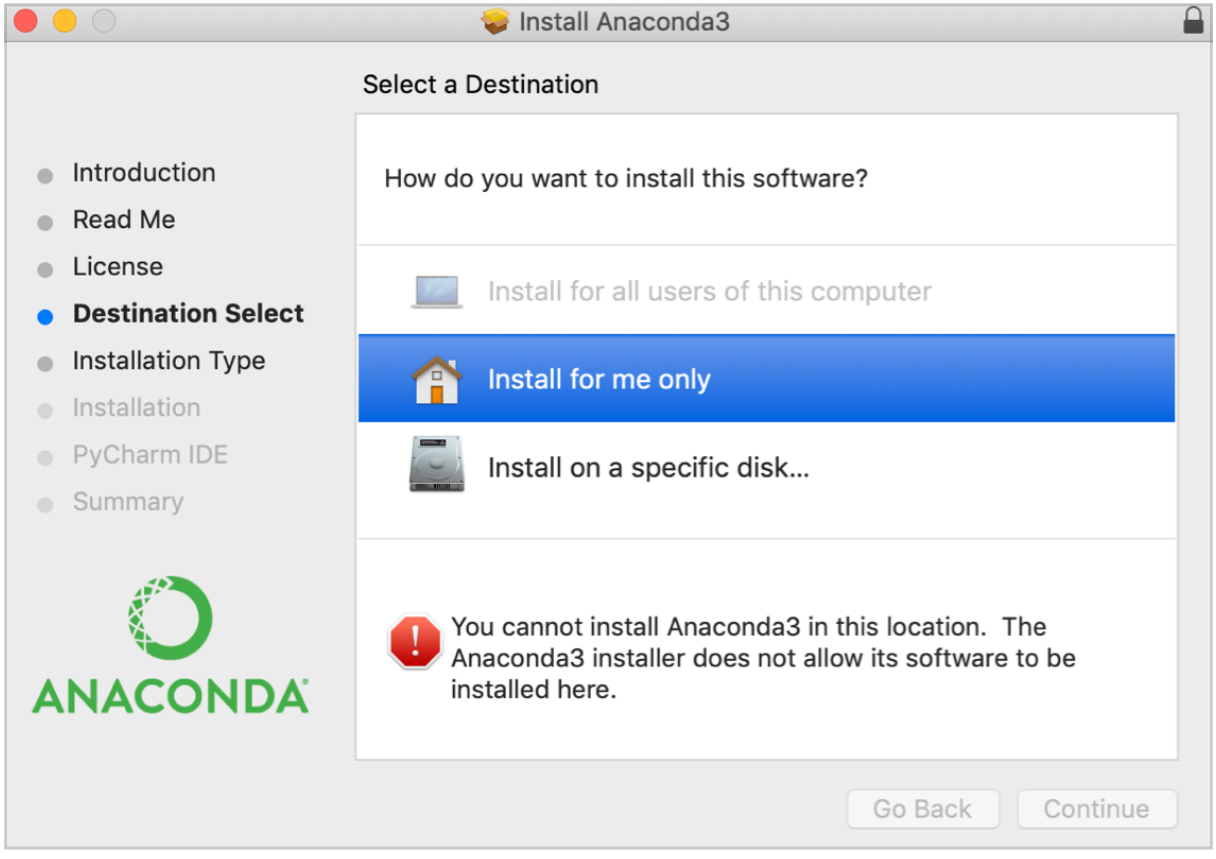


5. Click the Install button to install Anaconda in your home user directory (recommended):



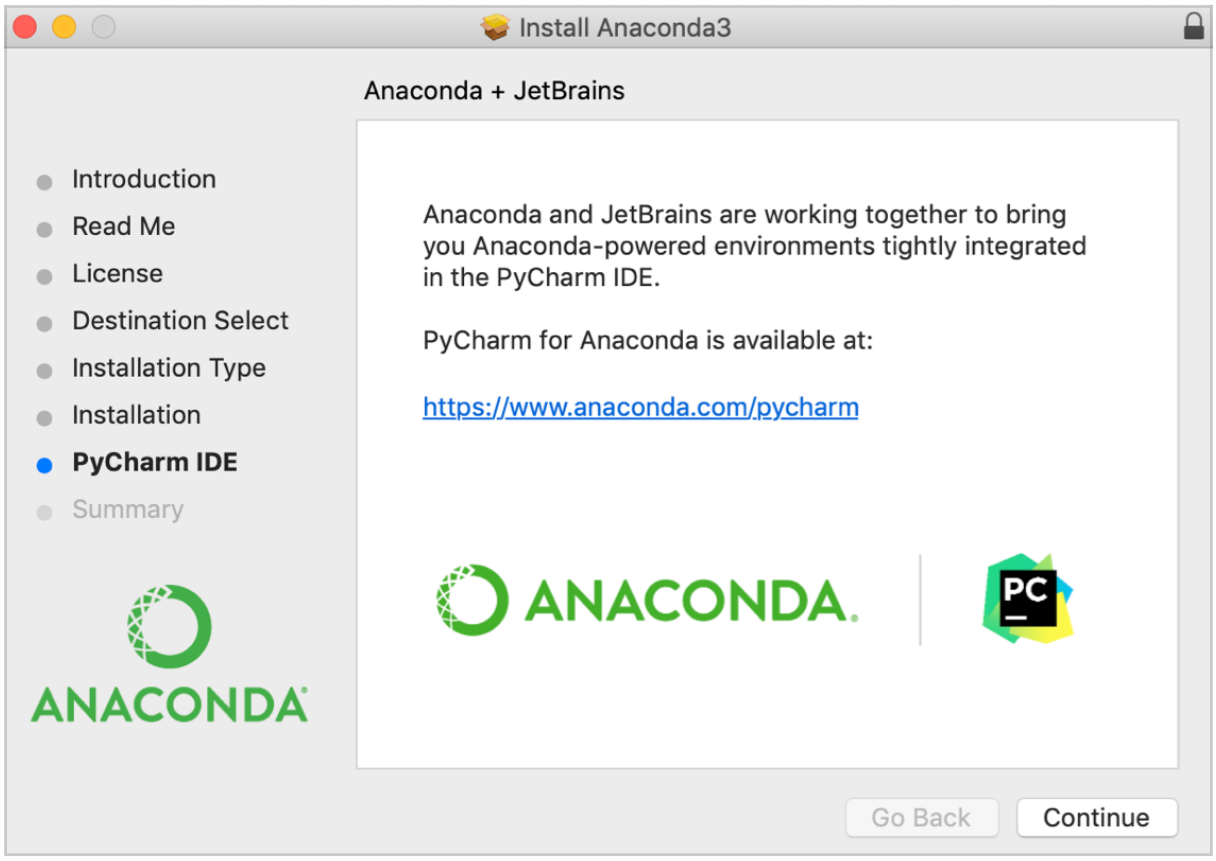
6. OR, click the Change Install Location button to install in another location (not recommended).

   On the Destination Select screen, select Install for me only.



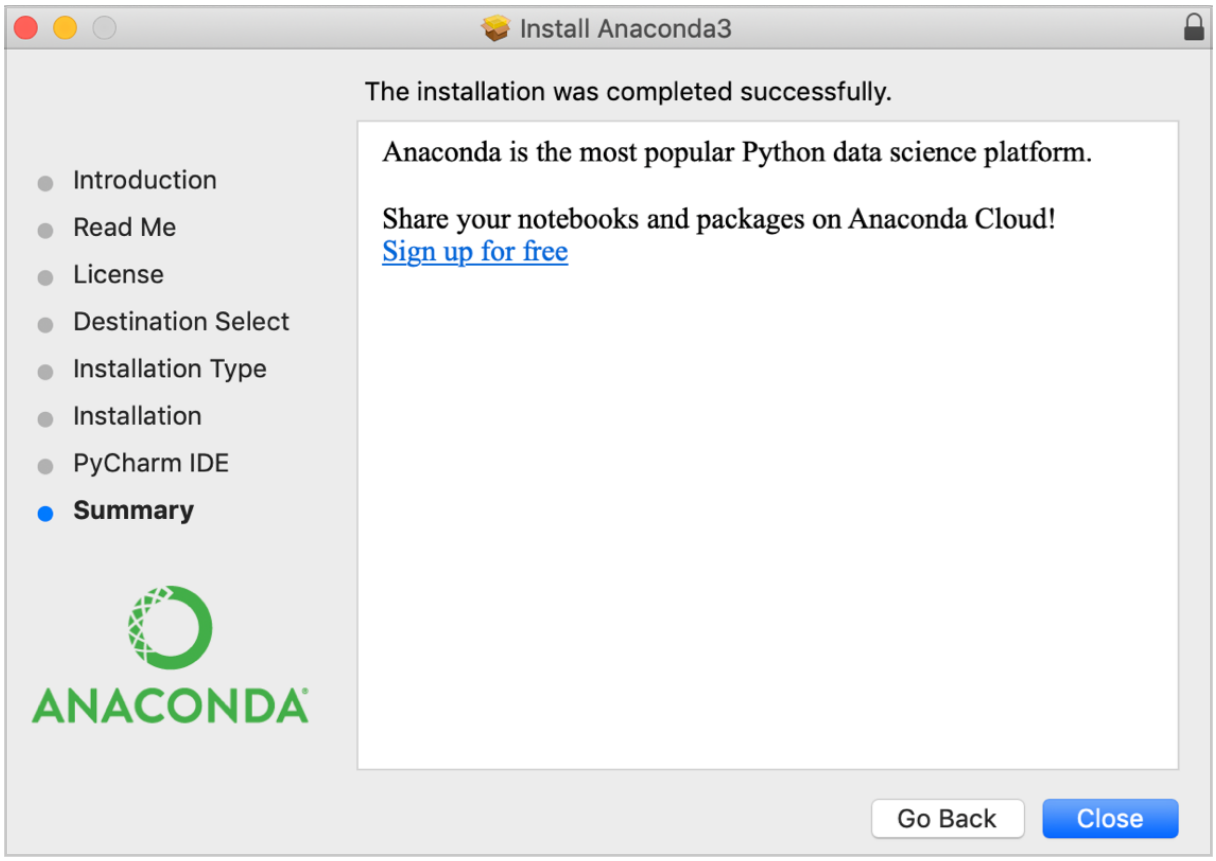
7. Click the continue button.
8. Optional: To install PyCharm for Anaconda, click on the link to https://www.anaconda.com/

pycharm .



Or to install Anaconda without PyCharm, click the Continue button.

9. A successful installation displays the following screen:



10. 1. After your install is complete, verify it by opening Anaconda Navigator, a program that is included with Anaconda: from Launchpad, select Anaconda Navigator. If Navigator opens,

you have successfully installed Anaconda. If not, check that you completed each step above, then see the Help page .

Using the command-line install

Use this method if you prefer to use a terminal window.

1. In your browser, download the command-line version of the macOS installer for your system.
2. OPTIONAL: Verify data integrity with MD5 or SHA-256 . For more information on hashes, see What about cryptographic hash verification?
3. Install for Python 3.7 or 2.7:
   If you aren't sure which Python version you want to install, choose Python 3. Don't choose both.

- For Python 3.7 enter the following:

```
bash ~/Downloads/Anaconda3-2019.03-MacOSX-x86_64.sh
```

- For Python 2.7, open the Terminal.app or iTerm2 terminal application and then enter the following:

```
bash ~/Downloads/Anaconda2-2019.03-MacOSX-x86_64.pkg
```

4. The installer prompts "In order to continue the installation process, please review the license agreement." Click Enter to view license terms.
5. Scroll to the bottom of the license terms and enter yes to agree to them.
6. The installer prompts you to Press Enter to confirm the location, Press CTRL-C to cancel the installation or specify an alternate installation directory. If you accept the default install location, the installer displays "PREFIX=/home/<user>/anaconda<2 or 3>" and continues the installation. It may take a few minutes to complete.
7. The installer prompts "Do you wish the installer to initialize Anaconda3 by running conda init?" We recommend "yes".
8. The installer displays "Thank you for installing Anaconda!"
9. Optional: The installer describes the partnership between Anaconda and JetBrains and provides a link to install PyCharm for Anaconda at https://www.anaconda.com/pycharm .
10. Close and open your terminal window for the Anaconda installation to take effect.
11. To control whether or not each shell session has the base environment activated or not, run
    `conda config --set auto_activate_base False or True`
    To run conda from anywhere without having the base environment activated by default, use
    `conda config --set auto_activate_base False`

This only works if you have run `conda init` first.

12. To verify the installation, see Verifying your installation .

    To begin using Anaconda, see Getting started with Anaconda .

Extra Credit - Scikit Learn

Requirement

- Python (>= 3.5)
- NumPy (>= 1.11.0)
- SciPy (>= 0.17.0)
- joblib (>= 0.11)

Installation

If you already have a working installation of numpy and scipy, the easiest way to install scikit-learn is using **pip**:

```
pip install -U scikit-learn
```

Or **conda**:

```
conda install scikit-learn
```

Reference

Installing scikit-learn — scikit-learn 0.21.2 documentation

scipy - Installing scikit-learn on Mac os x - Stack Overflow