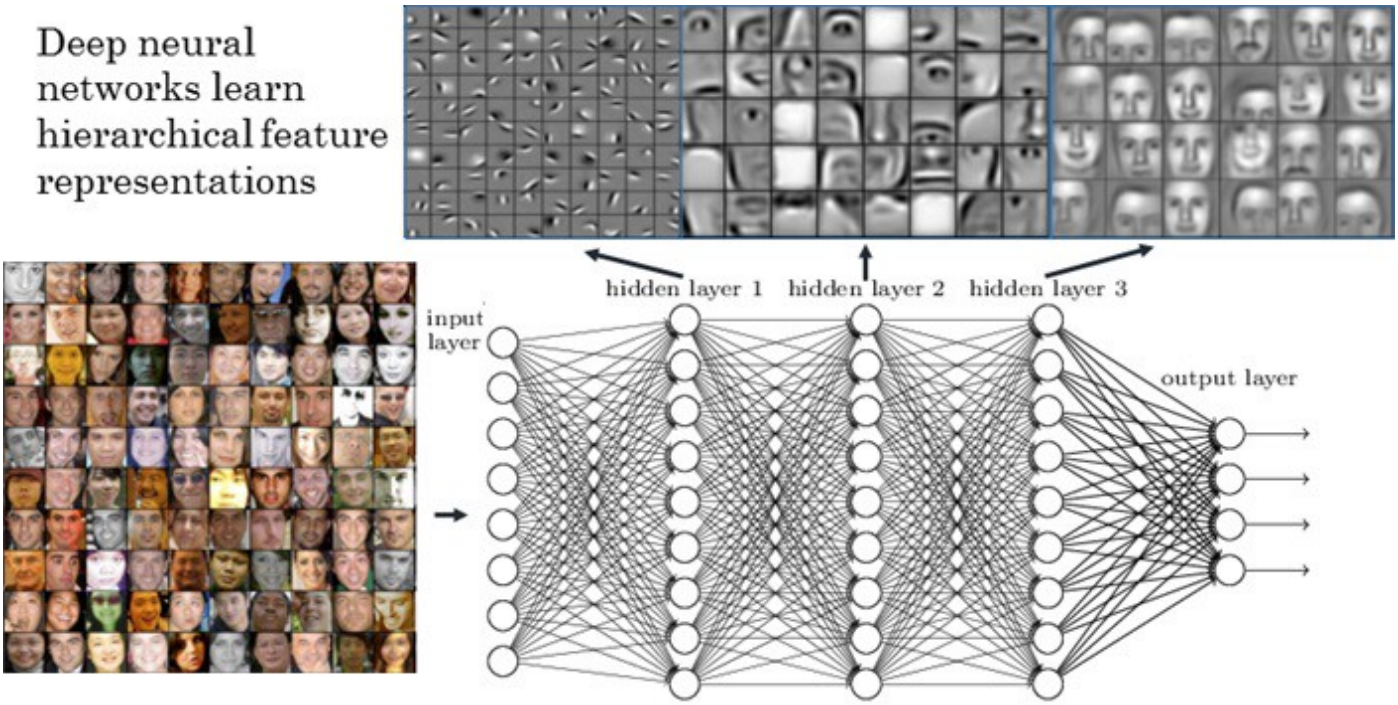
A Gentle Introduction to AI Explainability

Context

This blog series focuses on model explainability for supervised learning both for blackbox models such as deep learning models as well as interpretable models such as decision trees. This blog explains motivations for explainability as well as laying out basic definitions. The subsequent set of publications, dubbed *intermediate*, will describe basic explainability methods such as SHAPLY and associated software libraries. The final segment of the present publication ill focus on implementation of model explainability in blackbox domain specific fields such as transformer based architecture for NLP.

What is explainability

In its core, model explainability is to make decision process of a model, understandable for humans, regardless of the data representation or its intermediary encodings. For instance for a language model an explanation could be presence of a collection of words with certain proximity to one another. This is for instance a bag of words solution, but we know that attention-based modern neural models are much more accurate than statistical models based on bag of words. An explanation for a transformer -based model can be a bag of word approximation on the inference. We therefore can enjoy accuracy of more complex but opaque model, while explain its behaviour consistently with a simpler and less accurate model. Same logic can be applied on time series models making trading decisions or computer vision models. For instance a human interpretable data representation for computer vision, could be a *super-pixel*, a collection of pixels always appearing together that are recognizable by humans, rather than single pixels that a computer vision model might be focusing on. this would be the case with masks in case of convolutional neural networks that when visualized could represent a schematic image of an eye or a nose.



Reasons for explainability

[Burkart, Hurber] identify trust, causality, transferability, informativeness, fairness, accountability, making adjustments, and proxy functionality as the most common reasons for the need for AIX. Let us explore some of these reasons in more details.

Trust

One of the major issues of automation, specially the kind of automation that is achieved through probabilistic models, is trust of humans. For instance, we trust a human driver is context aware and makes the right decision when needed most of the time, while many people do have doubts about ability of self-driving cars to make the correct decision at the correct time all the time. The ability to interpret a model helps building both emotional and legal trust in automated decision making. There are two types of trust that someone who is affected by an inference needs to have on an AI-based system: trust in a specific prediction, which culminates in global explainability and trust the model as a whole will make the right decisions. This sort of trust is achieved through local explainability. Both these concepts are explained later in this blog.

Causality

Most ML inference work out probabilistic correlation of two or more phenomena and have no insight into causal structure of events, saving causal inference [[more information at CleaR conference]] (https://www.cclear.cc/AcceptedPapers (https://www.cclear.cc/AcceptedPapers)). For instance, the causal relation between emergence of birds of spring, such as swift, and arrival of the season has been clear for as long as human have lived in those regions where migrating birds arrive just before spring. People do acknowledge that there is a correlation between arrival of the migrant birds and changing of the season, but I do doubt anyone would consider migrant birds are causing spring to arrive.

Fairness

You might want to know why you are denied a loan application or a bank account. Opaque decision making does not help establish fairness. There has been numerous studies that human bias has resulted rejection of applicants whose name indicate certain ethnicity or religion. The same has been shown to be occuring based on postal codes and other demarcation factors. In many the disparity in providing opportunity is not intentional, but rooted in human bias. The algorithms are not immune to the same biases. The ability to explain and interpret how a decision is made, can help improve both data and algorithms to make fairer and more ethical decisions.

accountability

There are legal frameworks in place that require humans to justify their decisions. In Germany, for instance, an employer is obliged to provide a rejected candidate with reasons for rejection should the candidate requests such justification. Energy companies are required to provide justification for trading decisions if they own both production and distribution of energy in the supply chain. Car accidents require investigation so that intent and blame can be assigned to those involved. There are countless other examples that humans are required to justify their decisions. As algorithms are increasingly becoming autonomous decision makers, their decisions also need to be able to justified in accordance with existing and emerging legal requirements.

In addition to criteria for improving quality of decision making by the machine, we can help humans learn from better algorithms. If an algorithm becomes consistently better at humans at making certain decisions, which is the goal of automation through AI, studying the decision process can help humans learn from the algorithm and improve their decision process and thus creating a mutually reinforcing decision improvement feedback loop between man and machine.

Supervised learning: definitions

Model

model $h(x) = y$ is a supervised learning model where $x \in \mathcal{X} \subseteq \mathbb{R}^{d \times l}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}^{k}$. In the case of SML a set of labeled data $\mathcal{D}_{\tau} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is used to train the model in order for the model to to be able to map h over unseen data $\mathcal{X}^{\pi} = \{x^1, \ldots, x^k\}$ to prediction $\mathcal{Y} = \{y^1, \ldots, y^n\}$. In the simple case of single class classification and regression $l = k = 1$.

Blackbox model

A blackbox model in our context $b : \mathcal{X} \rightarrow \mathcal{Y}$, $b \in \mathcal{B}$, where $\mathcal{B}$ is the hypothesis space for a deep learning model.

Error

To evaluate a model we use an error measure that uses some topological distance mechanism on the output manifold to measure distance of a prediction to an observed value or $\mathcal{E} = p - o$, where p is a predicted value and o is an observed value. For instance RMSE (Root Mean Square Error) is a type of Euclidean distance that uses the euclidean distance between two points in a multi-dimensional space.
$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$.

Learning

Given dataset $\mathcal{D}$, SML attempts to solve optimization problem:

$$h^* = \arg\min_{h \in \mathcal{H}} \mathcal{E}(h(x))$$

.

In the case of parameteric models such as a deep learning model where the model parameters are represented as θ and optimized parameters as θ^*, the optimization problem can be formulated as:

$$\theta^* = \arg\min_{\theta} \mathcal{E}(h(x; \theta))$$

.

For most models solution to the optimization problem is often not unique. The matter is exasperated for complex models, specially in deep learning domain where we are faced with non-convex optimization and θ^* is only an approximation to a acceptable local minima. Other issues such as dataset biases make the matter of interpretability even more uncertain. We, therefore, need to find a solution as to explain why a model has made a decision based on a set of circumstances.

Explainability Approaches

In the context of this blog, there are in general two kinds of explainability, local and global. Local explainability, aka instance explainability, deals with explaining how a decision was made for a single input data $x \in \mathcal{D}$. This means that the explanation is valid only for x and its close vicinity. Global, or model, explanation approach, generate explanations for how a model in general arrives at a decision on a dataset $\mathcal{D}$. **Example**: An face detection system has recognized a person to be a person of interest. Local explainability should be clarify as why that person was matched against a known face. Global explainability, should be able to lay out how the face detection works. One of simplest ways is the projection of internal state of nodes in an image recognition model.

*The task of model explanation aims to generate a human understandable interpretation **explanator** for the learning process by extending or modifying the learning process described in the previous section*

Please bear in mind that explanators differ from predictors as the former always rely on the latter to perform the explanation task.

Table 1 provided a comprehensive classification of explainability approaches.

ExplainabilityApproach	Description
ante-hoc	Explainability is built into the model.
pos-hoc	Explainability is created after model creation.
instance/local	Explainability is is only applicable to a single instance of data and its close vicinity.
agnostic	Explainability is independnt of the model itself and is applicable to many or all models.
data independent	Explainability mechanism works without additional data and is applicable to many or all relevant datasets.
data dependent	Explainability requires data.

Model Explanators

model explanator takes a model $x[input] \implies \mathcal{Y}[predictions]$ and a specific labeled dataset as an input and creates and explanation belonging to the set $\mathcal{E}[explanations]$. More formally:

$$e :\rightarrow (\mathcal{X} \rightarrow \mathcal{Y}) \times (\mathcal{X} \times \mathcal{Y}) \rightarrow \mathcal{E}$$

As mentioned in the previous sections, there are two approached to explanation: global and local. In the case of global explanations, explanator e take a model b and a dataset $\mathcal{D}'$ or $e(b, \mathcal{D}'$. Local implementation, takes a model b and an instance from the dataset $(x, y) \in \mathcal{D}'$ as input. This is consistent with the intuitive definition that global explanators explain a model's inferences over a specific dataset and local explantors provide interpretation for a specific instance of data belonging to a dataset.

Characteristics of Model Explanators

Interpretability Interpretability in essence is quantification of understanding of how input values affect one another. For instance how income and the number of dependent children can affect approval score for a loan. It is essential to factor in human limitations as the goal of interpretability to make decisions of a model understandable to humans. An explanation with hundreds of dimensions is basically useless of an interpretation for humans. reducing dimensionality to top 5 or 10 important factors that most significantly affect decisions made by the model is the valid approach.

Local Fidelity: A surrogate model is an approximation to the original model. The surrogate models tend to be less accurate and indeed not factoring in all the parameters used by the inference model. The explanator still needs to be locally faithful at the instance that is being explained and at its immediate neighborhood. Importance of local fidelity is that local decisions might be influenced by a different set of parameters that differ from a decision to another. Focusing only on global parameters might not be able to explain individual decisions, at least based on small enough set of parameters that are human interpretable.

Model Agnostic: An explainer should be able to be trained to explain any model given a dataset and black-box model. This is the same criteria that is required for models themselves to be good at generalization.

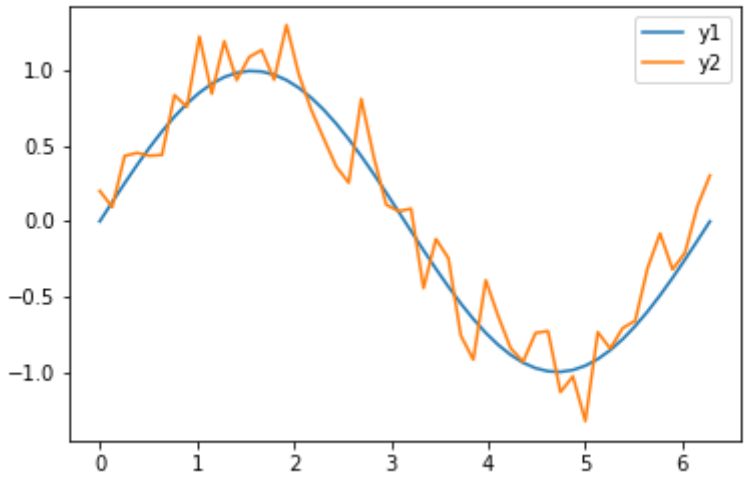
Global Prespective: Apart from trusting a specific local decision, it is also important to trust the model itself to be making the right decisions most of the time. This is similar to trusting decision making of a human, even though on occasions the human might make sub-optimal decisions. Also making good decisions on

occasions, does not make a human a generally good decision maker. So for an explainer to be trusted it has to make good local decisions most of the time; thus incorporating local fidelity and global perspective.s

Post-hoc model explanation for black-box modeles via surrogate models

As the name suggests, a black box model is a model, whose innerworkings are opaque to human understanding, such as deep neural networks. This is not a new sort of recognized problems. Even in the early days of neural networks criticism was levied at opaqueness of logic of neural networks through information constituency argument. One way to explain black-box models' inference is to develop a surrogate model that is a white box model capable of explaining decisions that the black-box models make. This way we can benefit from internal complexity of a black-box model such as a ANN, while use a simple model that consistently explains the inferences the more complex model has created. Do bear in mind that justification for effectiveness of neural network is based on *universal approximation theorem*, which in turn proposes to solve complex non-linear problems using approximation to a function that behaves similarly to the problem that we want to solve.

Figure 1 shows a simplified version of a simple approximation (blue graph) that can interpret the more complicated inference graph (orange one). This is foundation of machine learning in general. If the distance between predictions and observed is within a certain range we can justify the models' prediction posthoc.

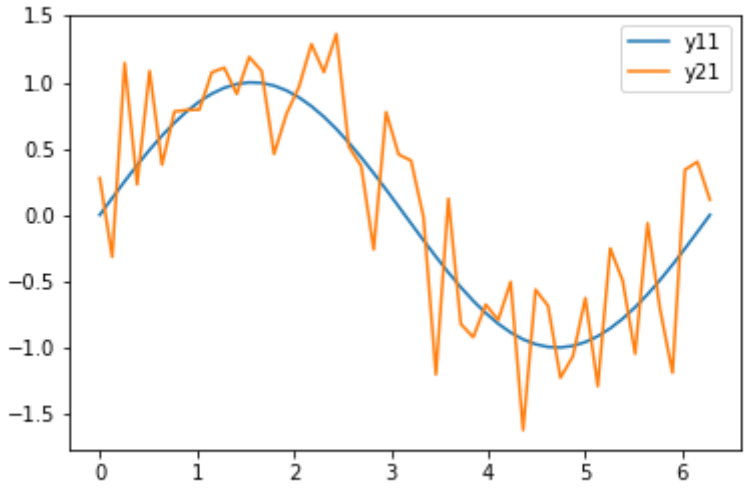


More formally, finding a surrogate model results in solving the problem that is formulated below. In simpler terms, explainability though surrogate models is the process of fitting an explainable model w to make predictions where the average distance between the the outcome of the surrogate model and predictions of the black-box model is bounded:

$$w^* = \arg\min_{x \in \chi} \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} S(w(x), b(x))$$

S is called the ***fidelity score*** and is a measure of how well the white-box surrogate model w approximates the black-box model b. The smaller the value for S is, the better the approximate is.

We can see that the in figure 2, the fidelity score is twice as high as the one in Figure 1, and hence the approximation is less accurate.



In the case *global* explainability, the surrogate model w approximates the black-box model b over a dataset $\mathcal{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathcal{D}$. Distribution in $\mathcal{X}$ should closely resemble that of $\mathcal{D}$ for the explanation to be plausible.

In local explanation, the surrogate model w is an approximation to a single instance of data with a neighborhood, rather than the whole of the dataset. or $\mathcal{X} = \{x' | x \in \mathcal{N}(x)\}$ and $\mathcal{N}(x) = \{x \in \mathcal{X} | d(x, x') < \epsilon\}$; where d is distant measure in a topology. In simple case of Euclidean space: $\mathcal{N}(x) = \{x \in \mathcal{X} : \ |x - x'| < \epsilon\}$.

LIME and SHAP are two of the most commonly used local explanability models using surrogates into which we take an in-depth look as they are crucial for explaining some deep learning based model, specially for transformer-based NLP models.

LIME (Local Interpretable Model-agnostic Explanation)

LIME and SP-LIME Lime is a post-hoc, model agnostic local explanation using surrogate models. SP-Lime uses a set pf representative examples to address trusting the model problem

Interpretable Data Representations

$x \in \mathbb{R}^d$ be original representation of an instance being explained. $x' \in \{0, 1\}^{d'}$ denotes a binary vector for its interpretable representation.

Fidelity-Interpretability Trade-off

An explanation is a model $g \in G$, where G is a class of potentially interpretable models. Note that g act on absence/presence of an explainable model, thus domain of g is interpratable data representation or $\{0, 1\}^{d'}$. not every $g \in G$ is human interpretable, this leads to introducing a measure of complexity $\Omega(g)$. The complexity could for instance be depth of a tree or non-zero weights. The goal of LIME is to find a highest fidelity while the value of complexity is at the lowest. The LIME paper introduces $\Pi_x(z)$ as proximity between instance z and data x (creating locality), where building an explanator for model $f : \mathbb{R}^d \rightarrow \mathbb{R}$ for classification models (could be reduced to a binary classification). As we saw earlier in this post, we would need to solve an optimization problem with minimizing average distance between explanation and prediction. Lime has added

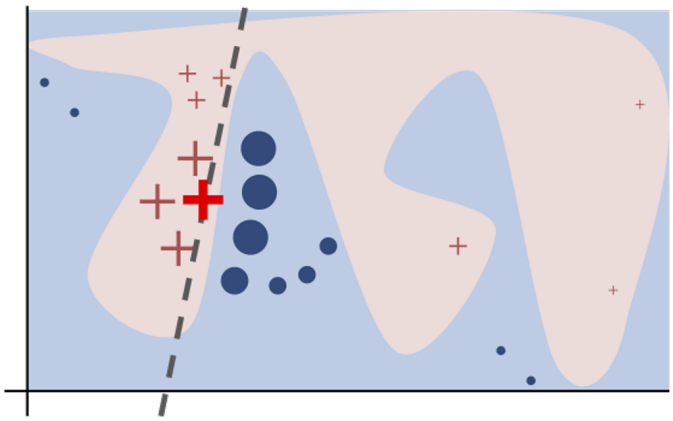
the complexity to the same optimization problem to create a fidelity-interpretability trade-off. More formally the paper has defined loss function $\mathcal{L}(f, g, \Pi_x)$ as opposite of faithfulness of explanator g to model f. LIME then is obtained by solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \Pi_x) + \Omega(g)$$

Minimizing Ω alone reduces the complexity and thus makes the explanation more human-friendly, whilst minimizing $\mathcal{L}$ maximizes local fidelity; therefore minimizing ξ balances between intrpretability and local fidelity.

Model Agnostic

We saw earlier that a desired explanation should be model agnostic. We have so far, successfully translated explanation to an optimization problem on a loss function $\mathcal{L}$ for model g. This is now training a model whose job is to provide a generalized solution to explainability. Like training any model, we need data and the data should result in g being model agnostic after training. training data is sampled uniformly at random around x' The data is sampled from amongst non-zero elements and is weighted by Π_x. loss function $\mathcal{L}$ is then approximated on the sampled data. As you remember $\Pi_x(z)$ as proximity between instance z and data x. Using Π as weight we create centers of locality in order to achieve high fidelity score using a generalized and model agnostic process.



In this simple model we can see that the model f that is being explained (blue background) is too complex to explain globally. Using LIME we can create a linear local explanation for x' can be achieved through a linear model within a neighborhood of x' (bold red cross). You can see that using the weight points (crosses) are sampled from much wider area of the dataset but the weighting mechanism enforces locality. The dashed line is the linear model that is used for local explainability, even though the whole of the model could not be explained using a linear model.

Sparse linear Explanations

As we see in the toy example in the image above, if the neighborhood is small enough, local explanations can be achieved through a linear model.

Global Surrogates

decision tree

Rule extraction

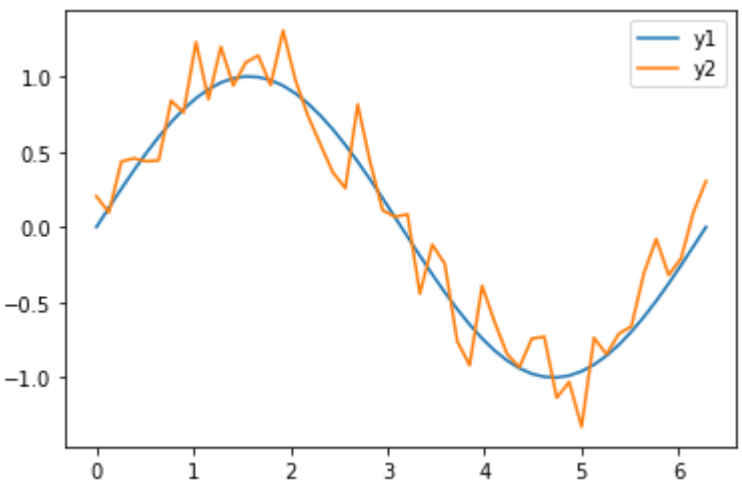
The Building Representations for Artificial Intelligence using Neural Networks (BRIANNE)

relevance of some input to the output i used to build rule condition

lime: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/1602.04938v1.pdf (https://arxiv.org/pdf/1602.04938v1.pdf) survey: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://arxiv.org/pdf/2011.07876.pdf (https://arxiv.org/pdf/2011.07876.pdf)

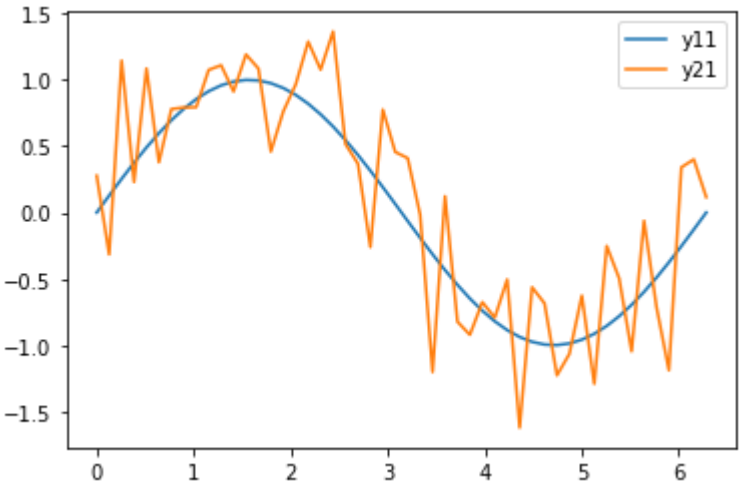
In [14]:

```
import matplotlib.pyplot as plt
import numpy as np
def fn(input):
    return np.sin(input)

def fn2(input):
    #return np.full(shape=input.shape, fill_value=0)
    return input * .001
x1 = np.linspace(0, 2*np.pi)
y1 = fn(x1);
y2 = fn(x1) - 5
y3 = fn2(x1)
gaussian_nose = np.random.normal(0, .2, np.shape(y1))
Y4 = y1 + gaussian_nose
plt.plot(x1, y1, x1, Y4)
plt.legend(['y1', 'y2'])
plt.savefig('approximation.png')

```



In [20]:

```
1
2  x11 = np.linspace(0, 2*np.pi)
3  y11 = fn(x11);
4  y21 = fn(x11) - 5
5  y31 = fn2(x11)
6  gaussian_nose = np.random.normal(0, .2, np.shape(y11))
7  Y41 = y11 + gaussian_nose *2
8  plt.plot(x11, y11, x11, Y41)
9  plt.legend(['y11', 'y21'])
10 plt.savefig('approximation1.png')
```



In [ ]:

```
1
```