ML Tutorial by

Lydia Gu, David Kellogg, Arsen Mamikonyan

Using scikit-learn and ipython notebook

What is Machine Learning ??

Combine

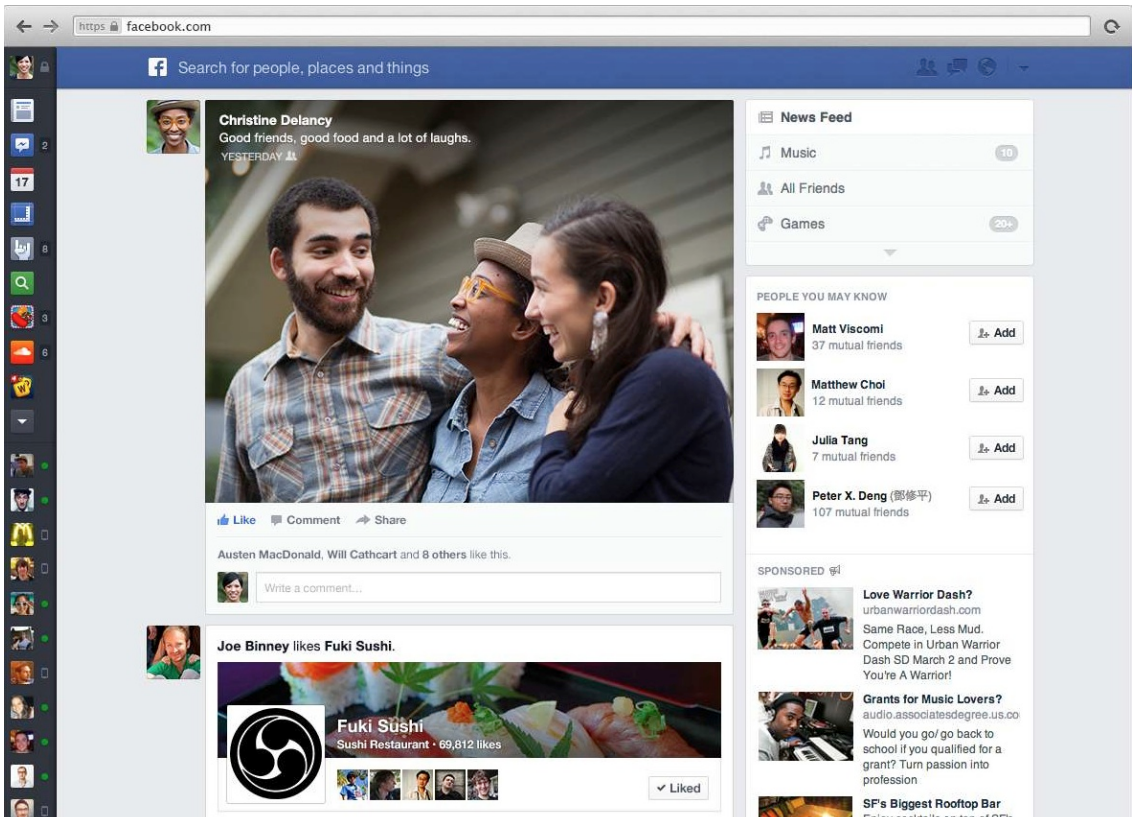
- (a lot of data)
- mathematics (and computational power)

> ***Machine learning explores the construction and study of algorithms that can learn from and make predictions on data.***
>
> *Ron Kohavi; Foster Provost (1998). "Glossary of terms". Machine Learning 30: 271–274.*

Some Machine Learning Examples

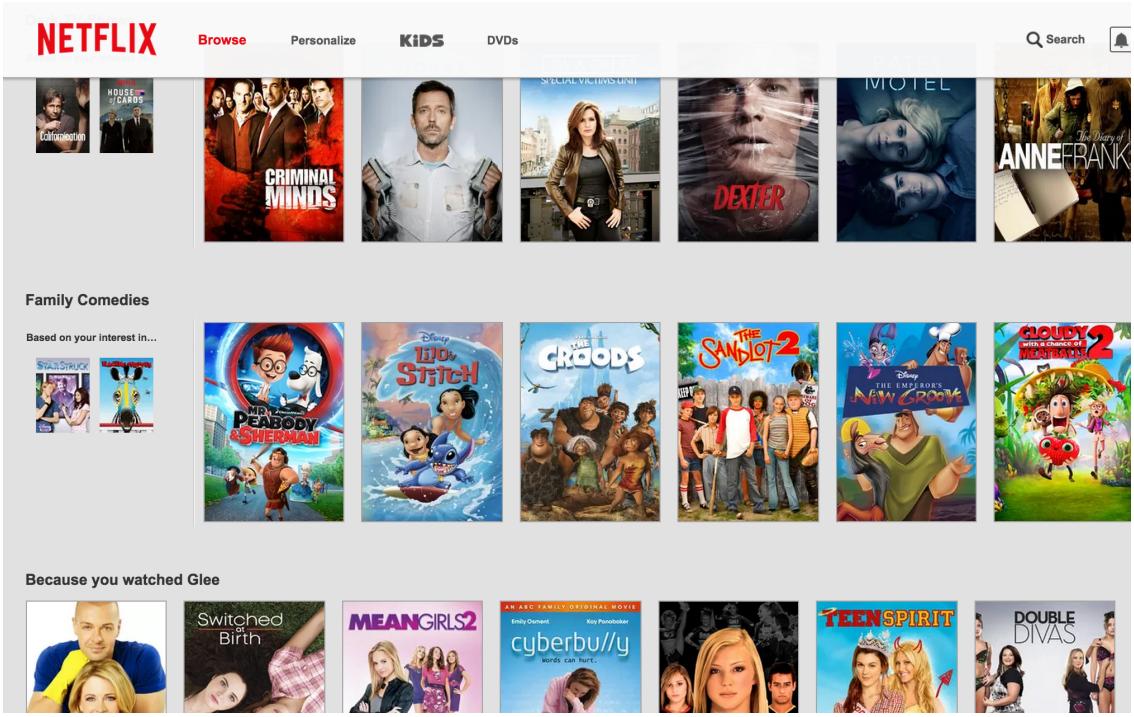
Machine Learning example 1

Facebooks personalizes your newsfeed based on your likes and clicks



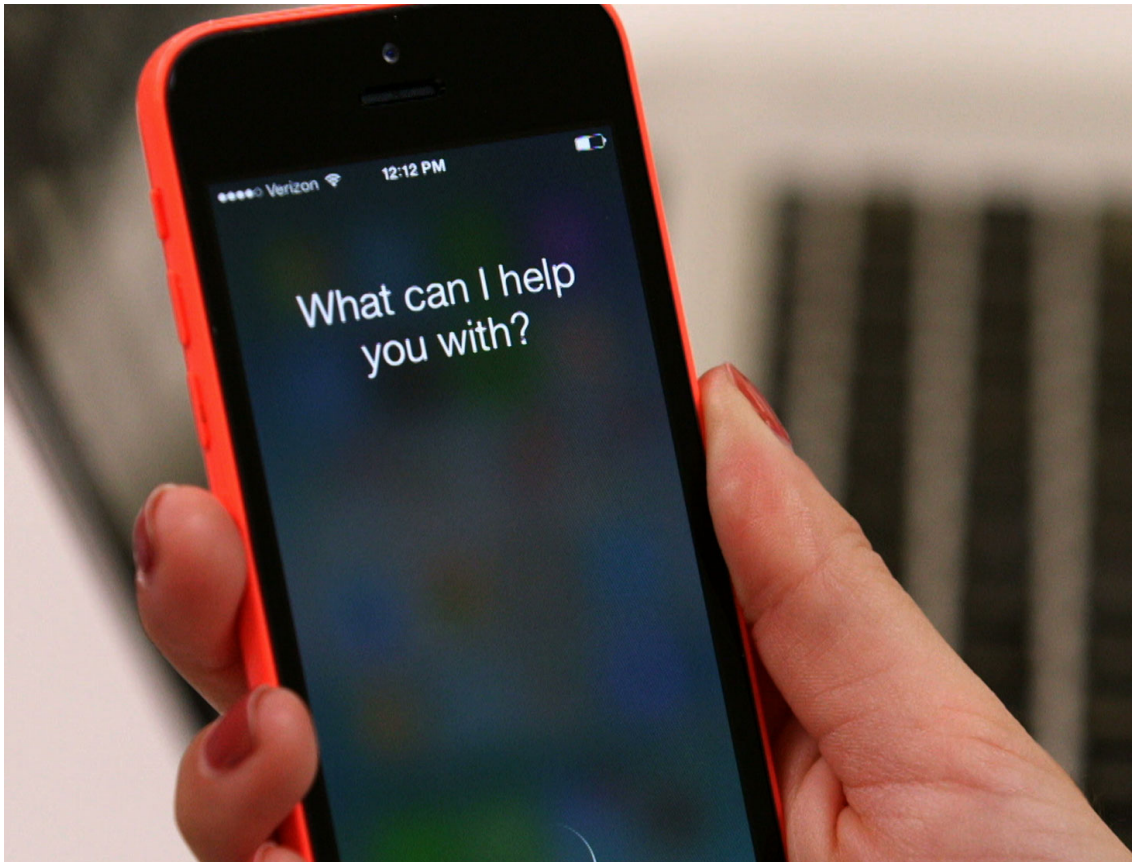
Machine Learning example 2

Netflix shows you suggestions based on what you have watched and other netflix users have watched



Machine Learning example 3

Siri



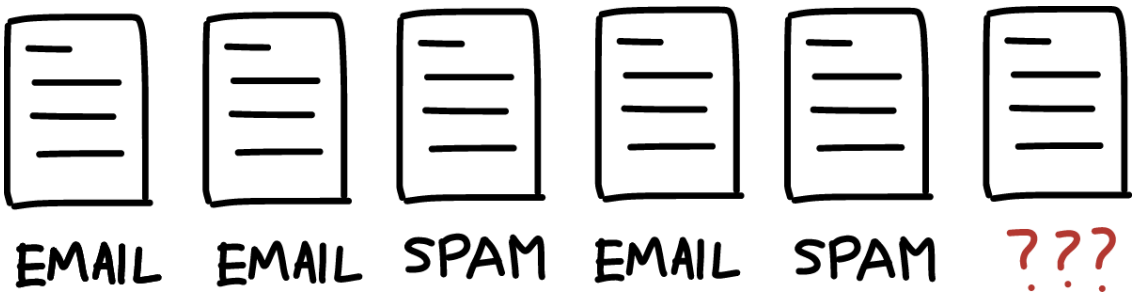
Supervised vs. Unsupervised Learning

Supervised Learning - Training data is labeled

Unsupervised Learning - Training data is not labeled

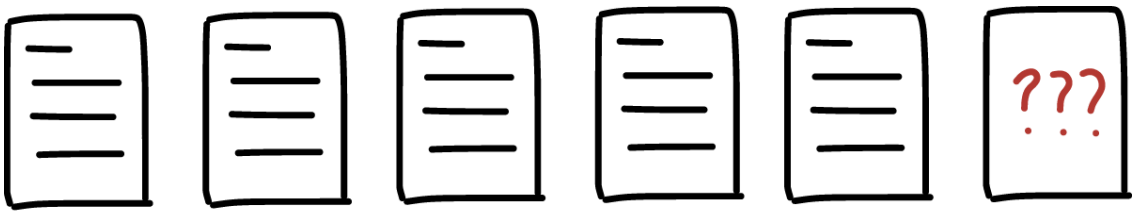
Supervised Learning

- Labeled training data
- Unlabeled data to answer questions on



Unsupervised Learning

- Unlabeled data to learn from
- Unlabeled data to answer questions on



Supervised Learning

- Labeled data (to learn on)
- Unlabeled data (to answer our question for)

Classification vs Regression

Regression

- labels are continuous
  - what will be stock pice (tomorrow after closing)
  - temperature in Phoenix (tomorrow at noon)

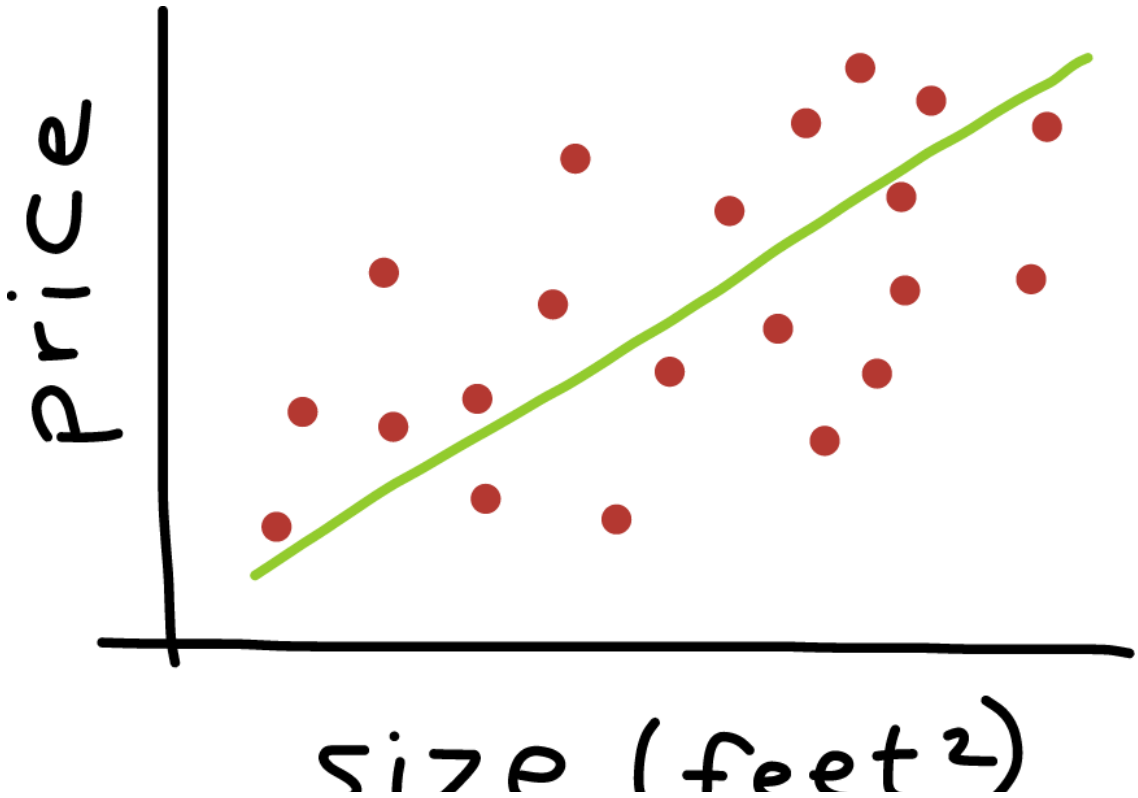
Classification

- labels are categorical
  - Which will be the best talk at the Techfest?
  - Will there be a good vegeterian option at lunch?

Regresssion

Data: Price listings of houses in Greater Boston Area.

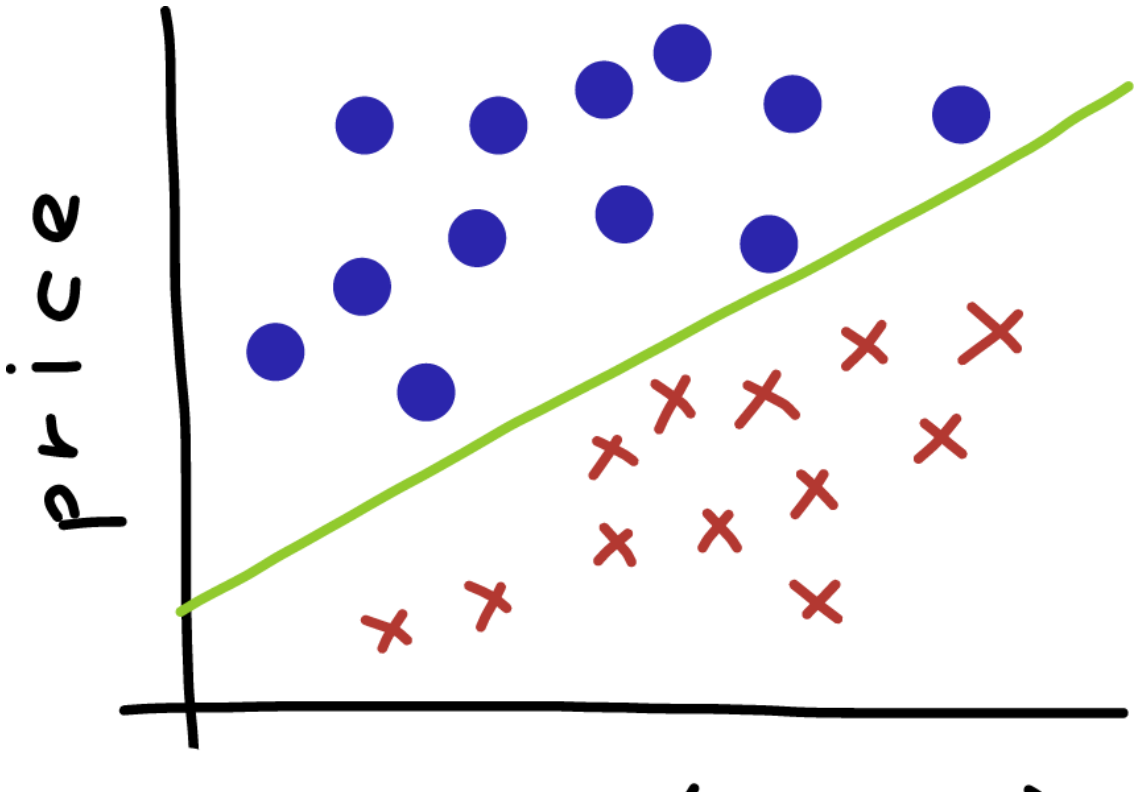
Question: What is the market value of a 900 square feet apartment?



Classification

Data: Price listings of houses in Greater Boston Area.

Question: Does the apartment have a good view?



Unsupervised Learning

We have

- only unlabbeled data

Goal

- deduce some structure of the data and predict (is this true?) based on that data

Unsupervised Learning Examples

- clustering
- density estimation
- dimension reduction

This tutorial

- Supervised learning methods
- use scikit-learn (http://scikit-learn.org/) to show you examples
- challenge problems

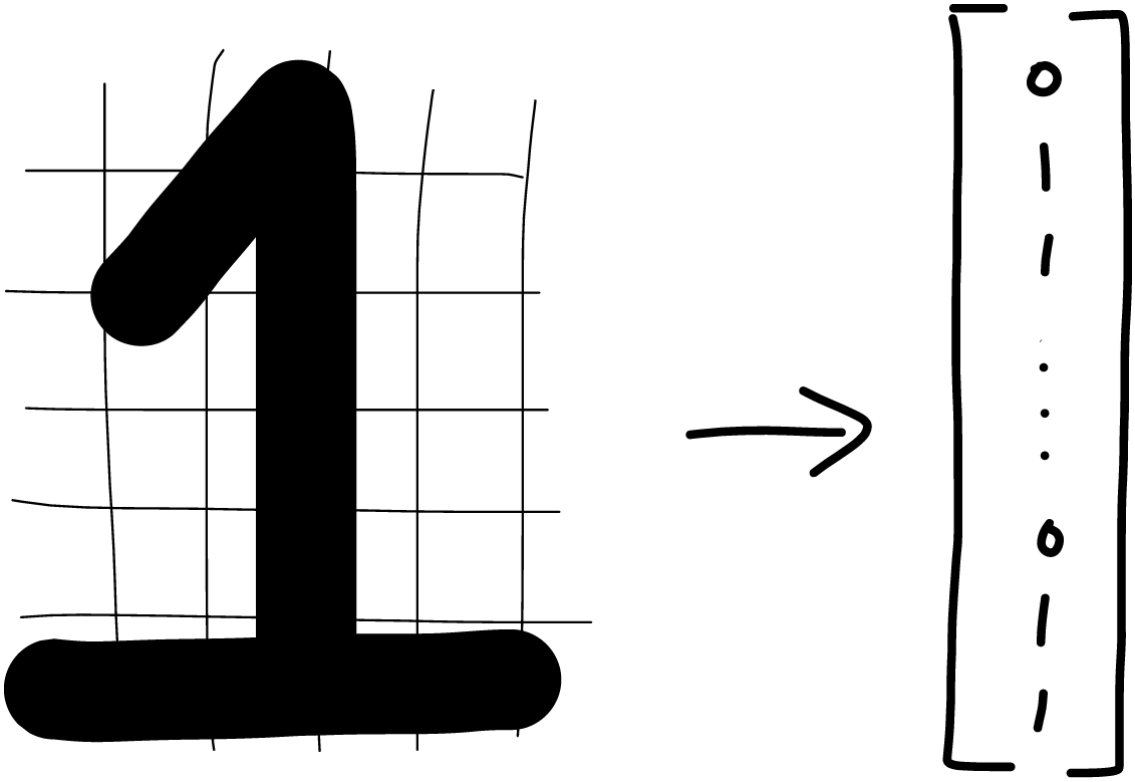
Machine Learning Workflow

1. Get data
2. Clean data
3. Data Transformation (feature engineering)
4. Fit a model (Data Mining)
5. Evaluate your model
6. (Deploy your model)
7. Use the model

Today we will cover

1. Get data
2. Clean data
3. **Data Transformation (feature engineering)**
4. **Fit a model (Data Mining)**
5. **Evaluate your model**
6. (Deploy your model)
7. Use the model

Data Transformation (Feature Engineering)



Why can’t we use original data?

- It could be **scattered**. We need to **consolidate** the data
- computer can't interpret the raw data. We have to guide it, transform the data into machine understandable format (construct features)

What are features?

- Each feature is a value that represents a transformation applied to a data point
- Features could be both discrete and continuous

Example features

- What is the distance of the apartment from the river in feet?
- Does the apartment have south facing windows?
- What is the area of the apartment?
- Value of a picture in an image
- number of logins in past day, number of distinct page visits

Data Transformation Best Practices

- Normalize your features
  - Some models might have problems if you have different type of features
  - e.g. if feature 1 is in range (0, 1) don't have feature 2 be in range (0, 1000)
- Careful with categorical inputs
  - Arsen (I'm not sure I agree with this)

Training a model

- A [mathematical] **Model** is the algorithm that we want to use to predict

- To generalize data we break up the data into training and test sets
- The training set is used to train the model, and the test set is used to evaluate the model performance on unseen data.

scikit-learn example

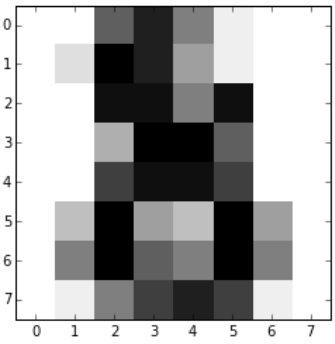
In [1]:
```
from sklearn import svm
from sklearn import datasets
clf = svm.SVC(gamma=0.001)
digits = datasets.load_digits()
```

In [2]:
```
clf = clf.fit(digits.data[:-1], digits.target[:-1])
clf.predict(digits.data[-1])  # Predicting with a classifier
```

Out[2]:
```
array([8])
```

In [3]:
```
%matplotlib inline
import matplotlib.pyplot as plt
plt.imshow(digits.images[-1], cmap=plt.cm.gray_r, interpolation='ne
arest')
```

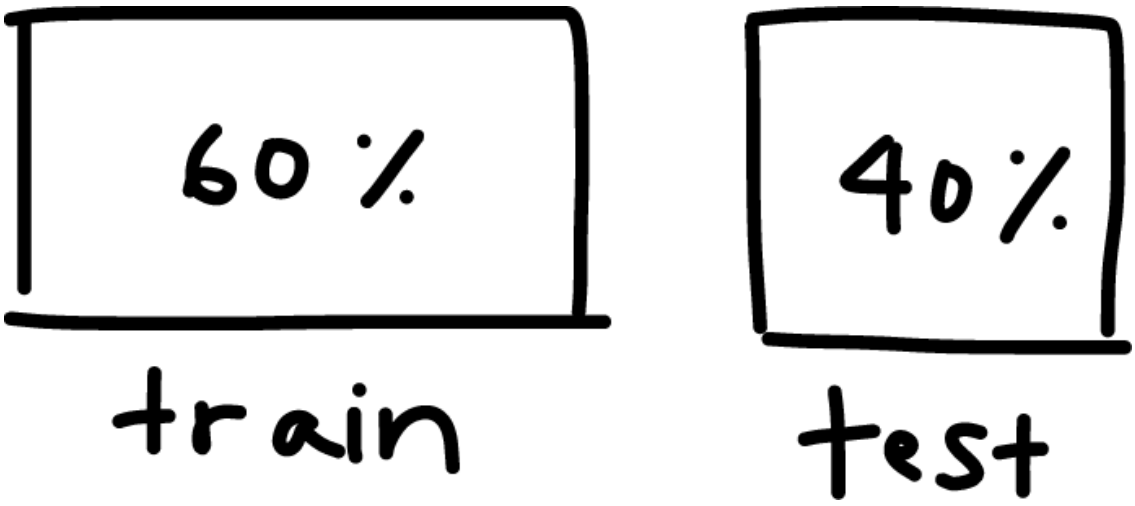
Out[3]:
```
<matplotlib.image.AxesImage at 0x7f7296c4ae50>
```



Evaluation/Choosing a model

- How do we know if a model is “doing well”?
- no one solution fits all
- we might have different objectives
- assume training data is very close to data we want to use the model on
- split chunk of data to test our models

Train/Test split



Cross validation

- A technique to make sure that your model works, before appling to real world situations

k-fold cross validation

1. We randomly split data into K sets
2. for each of K sets
    A. We train a model on the remaining K-1 sets
    B. We test if the model works on K-th set

scikit-learn example

scikit-learn has built-in functions for doing k-fold cross validation

In [10]:
```
from sklearn.cross_validation import cross_val_score
cross_val_score(clf, digits.data, digits.target, cv=5)
```

Out[10]:
```
array([ 0.97527473,  0.95027624,  0.98328691,  0.99159664,  0.95
774648])
```

What do these numbers mean?

Classification metrics

- accuracy (This gets reported by `cross_val_score`)
  - # of outcomes correctly predicted / total # of outcomes
- precision
  - # of correctly predicted in a class / total # of predicted to be in that class
- recall (sensitivity)
  - # of correctly predicted in a class / total # of items in that class

Regression metrics

- mean squared error
  - average of the square of the difference between the predicted and the true outcome.
- r2 score (coefficient of determination)
  - 

Underfitting vs. Overfitting

Bias vs. Variance

Bias - erroneous assumptions in the learning algorithm

Variance - error from sensitivity to small fluctuations in the training set.

Overview

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
- Evaluating model
  - Cross validation
  - Underfitting
  - Overfitting

Hands on examples

- Medical Data Classification Example (https://ipynb12.cloud.dev.phx3.gdg:8443/notebooks/classification_example.ipynb)
- Boston Housing Example (https://ipynb12.cloud.dev.phx3.gdg:8443/notebooks/housing.ipynb)