Machine Learning

What is Machine Learning?

The job of a machine learning system is to learn to make predictions from data. For example, we might want to be able to predict when an email is spam, whether a Facebook user is likely to enjoy Hotel Transylvania 3, or whether a medical tumor is likely to be malignant or not. The data for these tasks would be a set of spam and non-spam emails, a set of Facebook profiles and likes, and ultrasound measurements respectively.

Machine Learning is useful when:

1. You have lots of relevant data, and
2. You don't know in advance how to make the kind of predictions you are interested in.

Not all problems are appropriate for machine learning. We have lots of data on comets, and we could use a machine learning system to predict a comet's orbit. But we already know how to do that using physics, so that's not a good Machine Learning problem.

Examples of Machine Learning Tasks
Which Google photos have foxes in them?
What Netflix shows should we recommend?
Is this a positive comment about Beyoncé?
Is this tumor malignant or benign?
Which emails should be moved to the spam folder?
Was that rumble an earthquake or not?

Machine Learning vs. Artificial Intelligence

Machine Learning is a sub-field of Artificial Intelligence. There are lots of Artificial Intelligence systems that are not learning systems or that combine learning with other approaches. IBM's famous Deep Blue chess playing machine is a good example of an Artificial Intelligence system that was not based on learning from data.

Supervised Classification

In a **supervised classification** task, the machine learning algorithm has access to a previously existing set of **labelled examples**. The task is to use this data to learn a **theory** or **model** about how the labels should be applied. The learning is considered to be "supervised" because the system is told in advance which label belongs on which training example. When the supervised training is finished the system can (hopefully) be used to accurately put labels on (i.e. "classify") new items that it hasn't seen before.

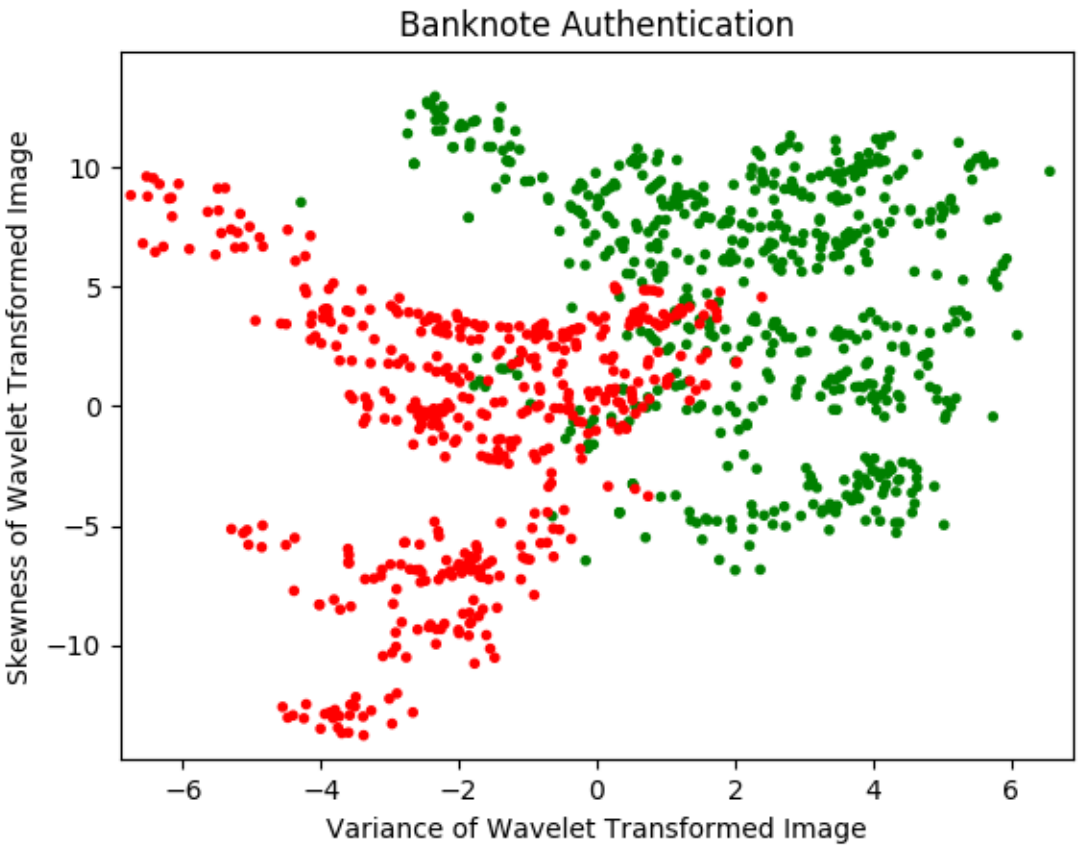
The list of tasks in the box above can all be viewed as classification tasks.

Standard practice when training a machine learning system is to split the set of pre-labelled examples into a **training data** set and a **testing data** set. The algorithm learns a **model** or **theory** of how to label the examples from the training set. The model could be an equation, a set of rules, a decision tree, or some other method of describing the difference between the labelled groups. When the model is ready, it is tested to see how well it generalizes by using the examples in the testing set. When eventually deployed, the machine learning system will then categorize brand new items that were not in either set.

Classification Unplugged: Counterfeit Money

The Training Data

The scatter plot below shows some real world data related to banknote authentication. The green points represent legal banknotes. The red points represent counterfeits.



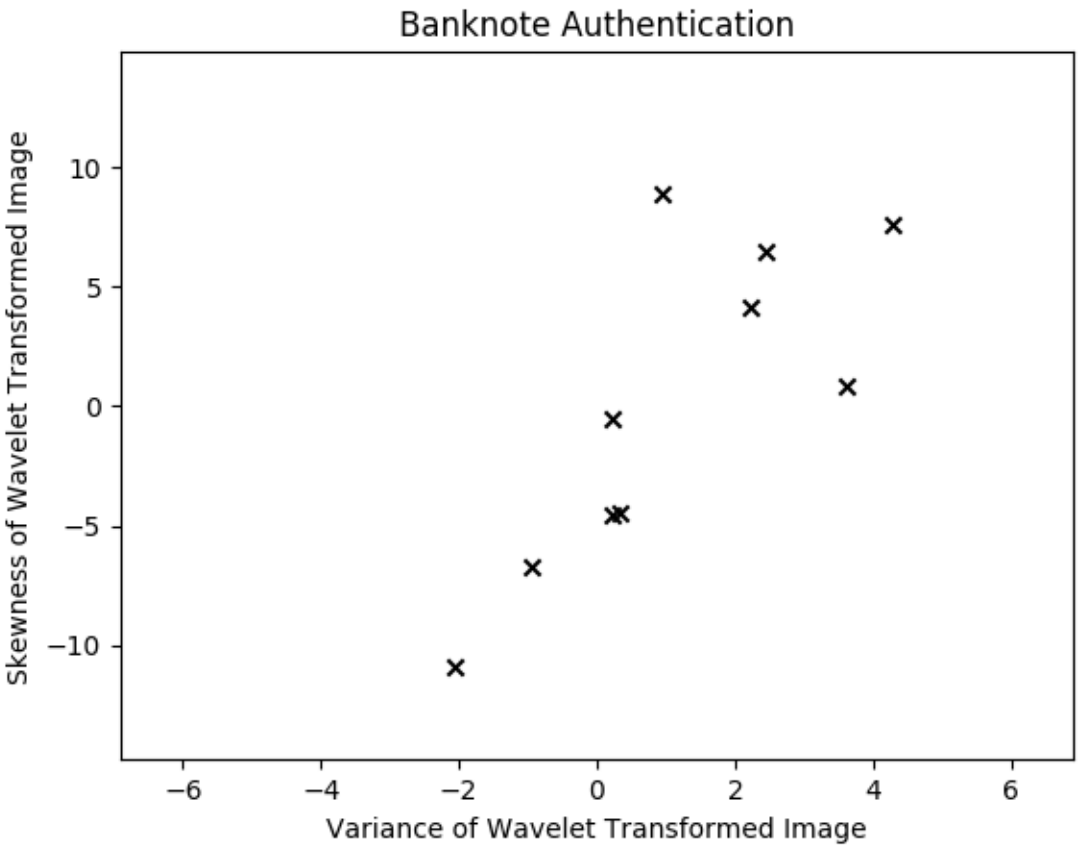
Each data point in the scatter plot was obtained by taking a picture of a banknote and then performing a mathematical procedure known as a "Wavelet Transform" on the image to generate four numeric features: variance, skewness, kurtosis, and entropy. (Fortunately, you don't need to know what these terms mean in order to do this task.) Only variance and skewness are shown in the scatter plot.

The classification task is to form a **model** or **theory** that does at least a half decent job of separating the two classes of banknotes. It's going to be pretty tricky to get a theory that separates the data with 100% accuracy, so don't think of that as the goal. Your theory can take any form you like (i.e. rules, equations, drawings, informal descriptions, etc.) but you should write it down when you are ready, and then move on to the testing data.

Data Source: https://archive.ics.uci.edu/ml/datasets/banknote+authentication

The Testing Data

In the testing phase, your job is to apply the theory you came up with during the training phase. For each of the example points below, use your theory to classify it as either LEGAL or COUNTERFEIT.



Solution and Debrief

Once you have classified the points above, the correct answers will be shown and you can score your model for accuracy.

Given the shape of the training data, it would be pretty surprising if you achieved 100% correct. This task is hard because there seems to be a gray area, or an area of overlap, between legal and counterfeit banknotes. Real world data often contains gray areas.

Solutions and further issues relating to this task are discussed in a separate handout.