Module 20: Ensemble Techniques

Video Transcripts

Video 1: Wisdom of the Crowd

Let's recap where we are. So far, we have learned how to train models for classification and regression tasks. These models include KNN, linear regression models, logistic regression models, decision trees, and support vector machines. In each case, the model is a function that takes a sample and returns a prediction for that sample. These types of models are sometimes called predictors or classifiers, depending on the context.

Up until now, we've restricted ourselves to using just one model to make a prediction. But in real life, we often make predictions based on a variety of opinions. In medicine, we seek second opinions. In government, we have parliamentary debate. In democracies, we select leaders by popular vote. The idea of making decisions based on consensus is also used in machine learning, where it goes by the name of ensemble learning.

In this module, we will learn about the two most important types of ensemble learning, bagging and boosting. Both of these are general ideas with many different variations. The idea of bagging will lead us to the specific algorithm of random trees. And the two most important types of boosting are AdaBoost and gradient-boosted trees. The essence of ensemble methods is to take a large number of models, each of which is pretty bad on their own, and to combine them in a way that results in a good model. But how can this work? If I have a bunch of bad models, shouldn't combining them just produce another bad model? Well, it depends.

Say you want to estimate the total number of gumballs in a large jar. A mathematician might compute the volume occupied by a single gumball by, say, cubing its diameter, and then divide the volume of the jar by that number. A physicist might set up an experiment where they take some sample jar with a known volume and measure how many gumballs it fits and then extrapolate to the given jar. A statistician would take a different approach and instead ask many people to venture a guess and then take the average of those guesses.

Here's how a statistician would see it. Say the jar contains exactly 100 gumballs. And we have two guessers. Guesser 1, who tends to guess low. Their expected value is less than 100 and Guesser 2, who tends to guess high. So they are both biased, but Guesser 1 is negatively biased and Guesser 2 is positively biased. Let's assume that their guesses follow a normal distribution with expected values μ_i and variance σ^2. And for simplicity, we'll assume that they have the same variance. Continuous normal distributions are perhaps not the best choice for guessing gumballs, but this will do for the time being. We'll also assume that the expected values, μ, are themselves chosen from a normal distribution, whose mean is the correct value of 100 gumballs. That is, even though each of the individuals may be biased, the population as a whole is not biased. This unbiasedness of the population is the wisdom of the crowd. And it is what we try to access when we crowdsource our decisions.

Here's a simulation of this scenario with the number of guessers varying from a single person to 1,000 participants. For each crowd size, we ran 50 rounds of ensemble voting. In each one of these rounds, every participant produced a single guess and then we averaged all of the guesses to obtain a final estimate for that round. The lines in the plot show the mean of the 50 rounds, as well as the upper and lower quartiles. And what we can appreciate is that, while the mean is close to 100 for all crowd sizes, the variance in the estimate decreases as the number of participants goes up. This is clearly seen in this plot of the variance, which shows a decreasing linear trend in a log-log plot. That means that the variance is decreasing as 1 over the number of participants.

And we've seen this formula before. It is the formula for the variance of the sample mean when the samples are independent. And this is a very important condition for trusting the wisdom of the crowd. The aggregate of decisions produced by individuals will be trustworthy only insofar as those decisions are made independently of each other. If say, for example, there is a single expert in gumball jar estimation and that person declares before the crowd that there are 82 gumballs in this jar, then this will naturally sway the crowd toward their opinion and the bias of the crowd will begin to look like the bias of the expert. This requirement is also true for machine learning models, as we will see.

Here's another example. Imagine you live in a village in a time before modern meteorology and you need to know whether it will rain or not in the next few days. Your best strategy might be to ask everyone in the village what they think and then decide by majority vote. In machine learning, we would regard each of the villagers as a classifier who takes as input what they see, smell, and feel, and produces a binary output, rain or no rain. Each one has been trained through a lifetime of experience. Some of them will be highly biased. Their decision rules are very simple. Sun is out, no rain. Simple, but also often wrong. Others have high variance. They remember and apply many details of their past experience, but to a fault. I once saw rain after a cat crossed the road. I see a cat crossing the road, so it will rain. They are also often wrong.

The reasonable thing to do is to decide by majority vote. As long as the collective bias is small and they are allowed to vote independently, then their individual biases and variances will cancel each other out. In the next few videos, we will see how these ideas can be applied to machine learning. I'll see you then.

Video 2: Aggregating Predictors (Part 1)

Let's apply the wisdom of the crowd to learning algorithms. As we've always done, we will split the dataset into training and testing data. Except now, instead of training just one model with the data, we will train many models. These could be anything. We could have logistic regression with k-nearest neighbors and decision trees. Or we could also consider many slight variations on the same type of model. So many decision trees, for example. All of the models fit into a larger metamodel. The metamodel, or ensemble model, receives a test data point and is in charge of distributing that sample to the component models and then aggregating their predictions into a final decision.

There are different ways of performing both of these tasks. The method for performing aggregation depends primarily on whether it is a classification or a regression task. If the task is classification, then the final decision can be reached by a majority vote, also known as hard voting. That is, we choose the class that receives the most votes as the prediction for the given input. This can be done whether it is a binary two-class problem or a multiclass problem. Some models produce not only a predicted class label, but also a probability of class membership for each class. So a data point may have 30% chance of belonging to class A, 20% chance for class B, et cetera. Logistic regression is an example of such a model. In this case, the aggregator can perform a soft vote in which we average all of the predicted class distributions and then choose the maximum.

For regression problems, the outputs of the model are combined by simple averaging. In all cases, the metamodel can also assign weights, α_1 through α_m, to the individual models, and thus give each one a greater or smaller influence in the final decision. And in boosting, which is another type of ensemble technique, we will see an arrangement where the predictors are trained sequentially, with the output of one model influencing the inputs of the next model, and so on. In the next video, we will implement a simple metamodel in Python. See you then.

Video 3: Aggregating Predictors (Part 2)

To illustrate these ideas, let's look at a short program in which we measure the performance of four different classifiers on the Iris dataset and compare them to the performance of the ensemble. I'm skipping the part where we load the dataset X and y.

The main loop in the code iterates the exercise 100 times, so that we get a more robust sense of the result. The code begins by allocating a couple of

NumPy arrays that will hold the accuracy metrics for each of the four models and another for the ensemble. Within the loop, we begin by creating four classifiers and storing them in a list called clfs. The classifiers are: LogisticRegression, a support vector classifier, a decision tree, and k-nearest neighbors. I've coupled each of these models, except for the decision tree, with a StandardScaler, and include them in a pipeline. It is not useful to do this for the decision tree because decision trees are scale invariant, so their output is the same, whether or not we normalize the input. But it is good practice to normalize the input of the other models using a StandardScaler. This is done in order to avoid numerical problems associated with having very different scales amongst features.

Next, we split the data into training and testing sets, reserving 20% of the data for testing. We then iterate through each of the four models. And within the iteration, we first train the model and then we test it by producing a prediction on the test data. And comparing that prediction to the actual test output. To obtain an ensemble prediction, we aggregate the predictions of all of the models by taking a majority vote. A simple way of implementing a majority vote is with the mode function of scipy.stats. This function takes in the predictions which, in our case, had been stored in y_pred, and it returns an object. To extract the actual mode, we must retrieve the 0th element in the mode attribute of the returned object. We can then compute the accuracy of the ensemble prediction and store it in its allocated array.

After iterating this 100 times, we take the mean of the accuracies to obtain the final result for each model. And here they are. The logistic regression classifier on its own, achieved the highest accuracy of about 93%, followed by the support vector machine, and k-nearest neighbors. The decision tree came in last, but don't count them out. Decision trees are actually the stars of ensemble methods. The ensemble did pretty well, but it was certainly not the winner. Its performance was a little worse than logistic regression and about equal to the support vector classifier. The reason for this has to do with our main criterion for trusting the wisdom of the crowd. Their decisions should be independent. In this case, because all of the models were trained on exactly the same data, they turned out to be highly correlated.

In the next video, we will learn about a technique that, when it was discovered, helped to overcome this problem and really put ensemble methods on the map. That technique is called bagging. See you then.

Video 4: Bootstrapping and Bagging (Part 1)

Calling on the wisdom of the crowd can help us in two ways. It can help to reduce bias and it can help to reduce variance. Bias is exhibited by models that are too simple for the data they are trying to match. For example, if we tried to capture the circular red class shown here with a linear decision boundary, this will invariably produce large errors.

A good model for this type of data should have a nonlinear decision boundary that surrounds the red points. However, by aggregating a large number of simple linear estimators by majority vote, we can obtain a model with a nonlinear decision boundary. Aggregation can also help with models that have too much variance. In contrast to bias models, these models are too flexible and they try too hard to fit the training data.

Researchers have found that it is best to build the ensemble with many variations of the same type of model, instead of using many different types, as we did in the last video. When we do this, we put ourselves into one of the two cases. Either the base models have a bias problem or they have a variance problem. And the approach that we take to building the ensemble will be different in either of the two cases. If our base models have high variance, then bagging will be the solution. If they are highly biased, then boosting will help. We will begin with bagging.

The problem of high variance is that the models are too myopically focused on the training data. If there is too little data, then they have learned to behave strictly according to that little bit of experience. Well, a simple solution would be to get more data and to train each of the component models with a different dataset. The problem, of course, is that we do not have more data. Data can be expensive. And if we had more data, we might instead decide to gather it all into one giant dataset and use that to train a single, more sophisticated model. So let's discard that possibility. And instead assume that the data that we have is all that we're going to get. No one is coming to our rescue with more data. It's as if we are trapped in a hole. Well, there's a saying for this type of situation, you have to pull yourself out of the hole by your bootstraps, because there is no other way out. And this is precisely the name of the technique that we will use, the bootstrap.

The idea of the bootstrap is both simple and very consequential in statistics and machine learning. The idea is this: The training dataset that we have was sampled from the population. We no longer have access to that population. Remember, we are in a hole. And so we do the next best thing, which is to sample new datasets from the dataset that we have. That is, we imagine that our training data is itself the population. It is very cheap to sample from a dataset that you already have. The key, however, is that this sampling has to be done with replacement. This means that having sampled a data point, this data point is…still remains available for future samples. Sampling with replacement makes the training data seem more like an infinite population. And it increases the variation in the bootstrapped samples.

Let's look at an example of how sampling with replacement works. Here we have a list, D, of 10 numbers from 0 to 9. From here, we wish to generate a new list of 10 numbers by sampling with replacement. We first pick a number at random, and say we pick a 3. We add a 3 to our new list, D1, but we do not remove it from the original dataset. It is still there and available to be picked again. Then we pick a second value, and a third, and a fourth. And then perhaps again, we find a 3. So when we sample with replacement, we can get repetitions in the new dataset. This means that the list, D1, which also contains 10 numbers, may not contain all of the numbers from the original dataset.

How many of the items in the original dataset should we expect to find in the bootstrap sample and how many are not selected? The answer to this question can be found with a little bit of math. First, in a dataset with N items, the chance of any one being chosen in a single draw is 1 over N. The chance of not being chosen is then 1 minus 1 over N. Because all of the draws are independent, the chance of not being chosen in any of the N draws is 1 minus 1 over N to the N. And hence the chance of an item being included at least once in the bootstrap sample is 1 minus, 1 minus 1 over N, to the N.

This plot shows the evolution of this quantity, as the size of the training dataset gets larger. The dash line is the asymptotic tendency as N goes to infinity, and it can be shown that this equals 1 minus 1 over e, which is approximately 0.632. So we can conclude that in any bootstrap sample of moderate to large size, we should expect to see about 63% of the items from the training dataset. The remaining 37% will consist of repetitions. In the next video, we will add a bootstrap sampler to our metamodel. See you then.

Video 5: Bootstrapping and Bagging (Part 2)

In the last video, we learned about bootstrap sampling. The technique of bagging is simply to use a bootstrap sampler to generate training data for an ensemble model. Bagging stands for bootstrap aggregation. The bootstrap samples help to decorrelate the models and hence to increase the power of the ensemble.

Bagging can work with many models, both for classification and for regression. But it works particularly well with decision trees. Since the averaging step will help to reduce the variance, we're not so concerned with the component models having a high variance, but we do want to choose models with low bias. So if we are using decision trees, we'd like those trees to be relatively deep. Since deep trees have high variance and low bias.

Here is some sample code. We're going to create 1,000 decision trees for our ensemble. We are given the training dataset X_train and y_train, as well as the testing data X_test and y_test. We iterate through the trees and for each one we create a bootstrapped training dataset. This is accomplished with NumPy's choice function, which samples with replacement from the index set of the training data. The next step is to create and train a decision tree based on this bootstrap dataset and to add it to our list of trees. We allow each of the trees to grow until all of their leaves are pure. This is the default behavior for the decision tree classifier in scikit-learn.

Finally, we compute the accuracy of the tree on the test set. The average accuracy for the 1,000 individual classification trees is 90.22%. The prediction for the ensemble is found by majority vote amongst all of the 1,000 trees. And this is, again, done using the mode function from scipy.stats. As we can see, the accuracy for the ensemble is significantly higher than the average accuracy of the trees. Which is impressive since the accuracy was already pretty high.

In the example we just did, we evaluated the accuracy of the individual trees and of the ensemble using a test dataset, which we had set apart before the training began. With bagging, however, there is another more attractive option called out-of-bag evaluation.

To illustrate, consider this example where the dataset consists of six points, drawn here as blue, green, red, cyan, magenta, and orange dots. We know that if we draw bootstrap samples from this dataset, we'll get new datasets, D_1, D_2, and D_3, each of which will contain only about two-thirds of the original data. The remaining one-third are called out-of-bag samples. Of course, the out-of-bag samples will be different for each of the bootstrap datasets. For D_1, for example, the out-of-bag samples are the magenta and orange dots. For D_2, they are the red and cyan, and for D_3, they are the blue and the green. This means that each data point will be in-bag for some trees and out-of-bag for others. The red dot is in-bag for the decision trees 1 and 3, and out-of-bag for decision tree number 2. We can then make a test prediction for a sample in D, if we only consider those trees for which that point is out-of-bag. A test prediction for the red dot can be made using a sub ensemble consisting only of tree number 2.

With out-of-bag testing, each data point is evaluated based on about one-third of the total ensemble. The technique has the advantage that it allows for the entirety of the dataset to be used for both training and testing. Scikit-learn includes a BaggingClassifier class that makes experimenting with bagging extremely easy. All one has to do is to pass a template model into the constructor, in this case a decision tree, as well as the number of models that one wishes to include in the ensemble. The rest is identical to the single model case.

Training is performed as usual with the fit method and prediction with the predict method. To request an out-of-bag score for the BaggingClassifier, one must pass the parameter oob_score=True to the constructor. The out-of-bag score is then evaluated upon training of the models and made available in the oob_score_ attribute of the BaggingClassifier object.

How many estimators should be included in the ensemble? The answer to this question depends, of course, on many factors; including the computational resources that are available. This plot shows that at some point it can become pointless to continue growing the model. In this simple case, that point is reached quickly after about 20 trees.

Looking at this issue, some researchers began to wonder whether the technique could be taken even further by adding even more variation to the constituent models. They found that indeed it was possible and this led to a new and powerful technique called random forests, which we will learn in the next video. See you then.

Video 6: Random Forests (Part 1)

In the last video, we saw how aggregating many models, and specifically decision trees, can improve performance. This can be explained by the wisdom of the crowd. But remember that this wisdom depends on the capacity of the individuals to make their decisions independently from the rest. Are the trees that we created with bagging truly independent from one another? Well, because we use the bootstrap, or sampling with replacement, they are indeed less correlated than they would have been otherwise. But ultimately the data is still coming from a single dataset. And we saw that each bootstrap sample will contain about two-thirds of the original samples. There will be many samples in common in each of the bags, and hence, the models remain somewhat correlated.

Random forests is an algorithm that reduces the correlation amongst trees in a bagged ensemble by introducing randomness into the training process itself. Here's how it works. Recall from the lecture on decision trees, the traditional decision tree generation algorithm. At every iteration of the algorithm, we created two new branches by splitting a node along the feature and threshold that yield the greatest reduction in entropy.

To see this, consider what happens if we build 100 separate decision trees based on bootstrap samples from the Iris dataset. This bar plot shows the number of times that each feature was chosen as the topmost split of the tree. In 53% of the trees, the first split occurred along the petal length, and in 47%, it occurred along the petal width. But in none of the bootstrap trees was either the sepal length or the sepal width chosen as the root splitting feature. This is reasonable because entropy can be most effectively reduced by splitting along petal length and petal width. So sepal length and width are poor choices. The algorithm is simply doing its job. However, from the ensemble's perspective, it is missing out on an entire class of decision trees, where sepal length and sepal width play a more important role.

The random forests algorithm seeks to remedy the situation by searching from a randomly-chosen subset of the features at every split in the tree. In scikit-learn's implementation, this parameter is called max_features. By setting max_features to 2, we are telling the algorithm to design each split based on a randomly-chosen set of two of the four features. This means that in about one-sixth of the cases, this subset will consist only of sepal length and sepal width. And hence, one of those must be chosen in about one-sixth of the cases. This may result in trees that are individually not as strong as they might have been. However, it increases the diversity of the ensemble, and this turns out to be very beneficial. Next, we will see an example of random forests applied to the task of recognizing handwritten digits. See you then.

Video 7: Random Forests (Part 2)

Let's test random forests on the task of recognizing images of handwritten digits. The dataset comes from NIST, the National Institute of Standards and Technology. And it consists of 1,797 low resolution grayscale images of handwritten digits from 0 to 9. Each image has only 64 pixels, and each of these pixels is taken as an input feature. Here's what it looks like as a pandas DataFrame. Each row contains a single image. The 64 columns are the shades of the 64 pixels, which take values from 0 to 16. The last column is the digit represented in the image.

Scikit-learn has a RandomForestClassifier class that makes working with this relatively complicated model very easy to do. The constructor takes in all of the parameters related to its decision trees. Parameters, such as the maximum depth of the tree, the splitting criterion, whether gini or entropy, et cetera. It also takes in the higher-level parameters that apply to the forest itself. The most important of these is n_estimators, the number of trees in the forest. As we've said, max_features is the number of features to be used when splitting nodes. And in this case, we are requesting that the out-of-bag score also be computed. After training the model with the fit function, we can retrieve the out-of-bag score. And we see that this random forest, with 30 trees and max_features set to 10, achieves a classification score of about 95%. Very impressive.

Here, we can appreciate the dependence of the out-of-bag score on the size of the random forest. As we increase the number of trees, the score increases rapidly at first, but then settles at around 97. A plot like this can be helpful for selecting the ideal size of the forest. In this case, I would say that about 80 trees works well. However, this may also depend on factors such as computational resources and how quickly you need the predictions to be computed. This plot demonstrates the benefits of increasing the diversity of the forest, by restricting the set of features used to split the nodes. The blue and orange lines show the difference in accuracy obtained with max_features set to either 5 or 30, relative to using all of the 64 features. The baseline of using all 64 features is depicted with a green line at 0. We can see that using all of the features is a better choice when we have very few trees. But for a large forest, we get about a 2% boost in performance by restricting the number of features to as little as 5.

Another very nice thing about a random forest is that it can be used to compute the relative importance of the features. The importance of a given feature is measured by finding all of the nodes in the forest that split along that feature, and then adding up the reductions in entropy that they produced, weighted by the number of data points in each node. A relatively important feature will be responsible for a larger share of the total reduction in entropy. In this way, the random forest can produce a score that ranks all of the features in terms of their importance.

Here we see the feature importance score for each of the pixels in the digit recognition task. The lighter colored pixels are ones whose values are the most important for identifying the number in an image. It looks like the pixel in the third row and sixth column is the most valuable one. A camera that lost this pixel would be more severely hobbled in its ability to recognize digits than one that loses any, or maybe even all, of the pixels in the left and right columns. So that's it for random forests. In the next video, we will learn about boosting, which is an entirely different type of ensemble method. I'll see you then.

Video 8: AdaBoost

In this video, we will learn about a different class of ensemble methods, called boosting methods. In contrast to bagging, which is useful for models with too much variance, boosting works well for models with too much bias. We will learn about the two most popular boosting algorithms, AdaBoost and gradient-boosted trees.

The bagging ensembles we studied in previous videos had a parallel architecture. That is, the training of the constituent models could be performed independently. Boosting ensembles, on the other hand, are built by stringing together a number of so-called weak models. A weak model is a very simple model, that due to its simplicity, cannot perform very well on the given training data. In the context of classification, a weak classifier is one that performs only slightly better than random guessing. The idea of boosting is to combine a large number of weak models in a clever way, such that together they make a good or strong model.

AdaBoost was the first hugely successful boosting algorithm. It was invented in the early 1990s and it earned the prestigious Gödel Prize, in computer science, for its authors. AdaBoost is primarily a classification algorithm. The weak classifiers it employs can in principle be anything. However, they are usually very shallow decision trees called decision stumps. A decision stump is a tree with only one node. It takes the dataset and cuts it with a single slice that is aligned with one of the features.

Obviously, these trees are extremely fast to train and even faster to evaluate. To make a prediction with a decision stump, all you have to do is check whether the value of a single feature is above or below the threshold. It is very easy to construct a decision stump that qualifies as a weak classifier, meaning that it achieves greater than 50% accuracy on the training data. Why is this true? Well, if you choose any decision stump at random, meaning you pick a random feature and a random threshold and you get less than 50% accuracy, then all you have to do is invert the inequality, and now you have a stump with greater than 50% accuracy. So decision stumps are good candidates for building boosting algorithms.

The algorithm for AdaBoost begins by setting the iteration counter, s, to 0 and assigning an equal weight, w_s^i, to every sample i in the dataset. This weight represents how much effort the stump, s, should put into correctly classifying data point i. And to begin with, all samples are weighted equally. Next, we create a stump that can correctly classify samples, accounting for at least half of the total weight. Again, this is not difficult to do. If we do not succeed, we can simply flip the inequality.

In this picture, we can see that this stump correctly classifies most of the red points, but incorrectly classifies a large clump of blue points in the middle of the lower cluster. We then compute the total weight of the samples that were misclassified by the stump. This is the stump misclassification score, and we denote it with $\in_s$. This number is expected to be less than 0.5, since the stump is capable of better than random classification. The misclassification score is now used to determine the influence of this particular stump in the overall ensemble. The larger the stump's misclassification score, the smaller its influence. We denote the influence coefficient with an α_s. And it is calculated as $\alpha_s = \frac{1}{2}$ log of $\frac{1-\epsilon_s}{\epsilon_s}$. The plot for this formula drops from plus infinity to minus infinity, as ∈ goes from 0 to 1. And it crosses 0 at ∈ = 0.5. Since all of our stumps are required to have misclassification rates less than 0.5, their α will always be positive.

The influence parameter α is then used to update the weights for each of the samples. If a sample was misclassified, then its weight is multiplied by e to the α_s, which is larger than 1, since α_s is positive. If the sample was correctly classified, then its weight is divided by e to the α_s, which causes it to decrease. The weights are then normalized by dividing them by their sum. This ensures that they continue to add up to 1. The counter is then incremented and we repeat the loop as many times as we wish.

To make a prediction with AdaBoost, we simply collect all of the predictions from the component classifiers, weigh them by their influence parameters, α_s, and then select the predicted class by a weighted majority vote. Here we see the evolution of AdaBoost through the first 22 iterations. With S = 1, we have only a single stump, which splits the data into two approximately equal parts. With four stumps, the algorithm has created four vertical stripes that correctly classify the centers of the two clusters. After seven iterations, we begin to see horizontal stripes. And after 10, the algorithm begins to isolate the centers. It continues to work at picking off more and more samples. But we can see that it is having a hard time with the region of mixing between the two clusters.

Here we see the evolution of the weights for all of the individual data points over 150 iterations. Whenever a point is correctly classified, its weight decreases and attention has shifted to other points, but then it gets misclassified and the weight jumps back up. The result is a jagged switching evolution of the weights, which makes it a bit challenging to decide when to stop the algorithm. The predictions of the algorithm do settle down, however, because they depend only on the influence parameter, α_s, and not on the sample weights. Here we see that α_s becomes small after about 80 iterations. This is a good time to stop the training process.

And here is confirmation that the decision boundaries do indeed settle down. We can see that the red and blue regions stop changing after about 40 iterations. This points to an important property of AdaBoost and of boosting algorithms in general. They are slow learners, meaning that they are not easily overfitted. If we continued to boost this model, we would eventually overfit it. But in contrast with other models, such as decision trees, AdaBoost is very forgiving and it gives us plenty of time to stop the training process before reaching that point. In the next video, we will learn about another boosting method called gradient-boosted trees. See you then.

Video 9: Gradient Boosting Trees

In the previous video, we took a detailed look at the AdaBoost algorithm. Of course, scikit-learn provides an implementation that hides all of that complexity and makes AdaBoost as easy to use as any other classification algorithm. We simply call the AdaBoostClassifier constructor and pass in a base model, which is usually a decision tree with a maximum depth of 1.

When we described the AdaBoost algorithm, you were probably asking yourself where these particular formulas came from. Why is the influence parameter equal to 1/2 the logarithm of 1 minus ∈ over ∈? Why are the weights multiplied by e to the α? As it turns out, AdaBoost is a particular example of a larger class of boosting algorithms called gradient boosting.

The gradient boosting setup defines a loss function at the bottom of which is an ideal classifier. The algorithm then applies gradient descent to build up a model that with every step gets closer and closer to this ideal classifier. At every step, the gradient boosting algorithm creates a weak model that points in the approximate direction of the ideal classifier. Because the weak model is better than random guessing, we are assured that it will point us in the general direction of the target. It will never point away from the target. The algorithm then takes a step in the direction indicated by the weak model, and then repeats the process. It creates another weak model, takes another step, and so forth. Provided the loss function is convex, this procedure guarantees that we will eventually get very close to the target.

AdaBoost is simply an implementation of this idea, with a particular loss function, and an adaptive step size. The loss function is the exponential risk. And, without going into all of the mathematical details here, you should just know that this particular cost function is what leads to the weight updating rule that we saw. In the gradient descent view of AdaBoost, the influence parameters, α, are actually the step sizes. The formula for α was optimally chosen to find the minimum of this particular loss function as quickly as possible.

Gradient boosting then refers to the general idea of applying gradient descent to the problem of boosting. Gradient boosting trees is another particular type of gradient boosting. Like AdaBoost, gradient-boosted trees uses trees for its base model. These can be either decision trees or regression trees. Unlike AdaBoost, gradient-boosted trees uses the squared loss as its cost function. And for this reason, it cannot be implemented as a simple weight updating scheme. Let's see how it works.

The algorithm begins by initializing the boosting model to 0. This means that it predicts 0 for all of the samples in the dataset. Our goal is to build up H, so that it predicts the correct labels, y, for all of the data. If we achieve this, then the red dot, H, in this picture will coincide with the magenta star, y. We then compute the difference between the desired outputs and the predicted outputs. In the picture, this is an arrow that points from the model, H, toward the target, y. It is the direction in which we would like to advance. Let's call this direction, r.

Next we train a weak model, h, on labels, r. In other words, we train a model that takes the feature data as input and is trained to produce the residuals, r. h is a shallow regression tree, typically no deeper than four nodes. It will not be perfect. So it will not reproduce r exactly. But we trust it enough to take a small step in that direction. We do this by adding to H a small coefficient, α, times the output of the weak learner. Then we repeat, and again compute the desired direction, r, using our updated model, train a new regression tree, take a new step, and so on. Until eventually we get very close to the target. When this happens, we will have a strong learner, H, composed of a weighted sum of a bunch of weak learners, h, with weights, α.

In scikit-learn, gradient- boosted regression trees is implemented in the GradientBoostingRegressor class. This class is very similar to others we've seen. So let's try it out. We're going to create a one-dimensional regression model to approximate this series of red points. In the first iteration, the algorithm creates a regression stump that approximates the points with a step function. The residuals of this model are then used to fit a second regression stump, which is added to the first. It seems that the second tree chose to correct large errors near the right edge of the plot. Then we repeat to find a third tree, a fourth, and so on. And this is what the model looks like after 10 iterations. This may be good enough for our purposes. The complete model is still pretty simple, since it consists only of 10 stumps. But we can also go on to 50 stumps, or to 100. And after 500 iterations, we have pretty much converged to a model that fits the data exactly. H has reached the magenta star. But what if we kept on going? Well, after 1,000 iterations, the model still looks pretty much the same. So the algorithm is pretty robust in the number of iterations. Of course, these extra 500 regression stumps are probably not worth the added computational burden.

As we can see, boosting algorithms have some very nice properties. First, they are good at reducing bias of weak learners. And in doing this, they do not increase the variance as much as other algorithms might. Ensemble methods, in general, have an excellent track record in the world of machine learning. They are often the winning algorithm in machine learning competitions. And they feature prominently in industrial applications, such as search engines and recommendation systems. You should always consider an ensemble method for your most challenging machine learning tasks.