**CS229 Lecture notes**

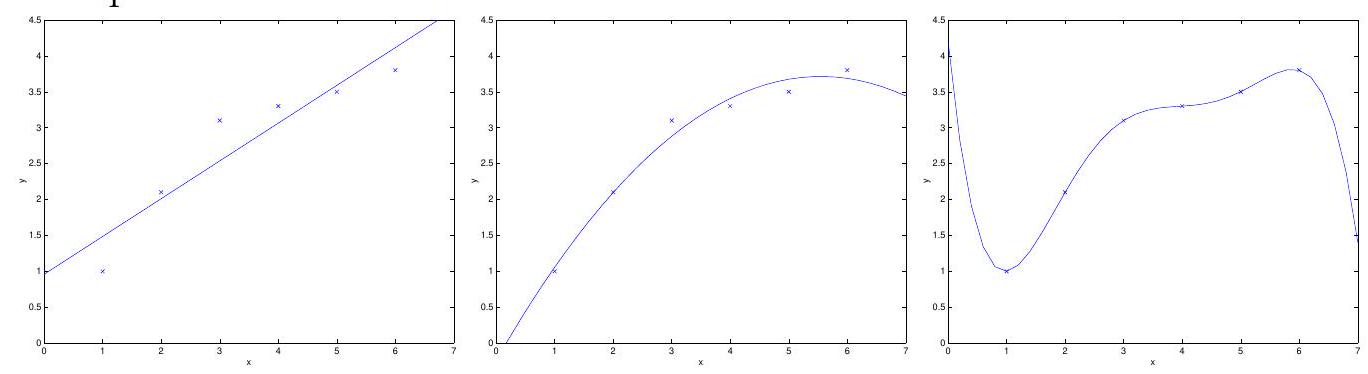
Andrew Ng

**1. Part VI**

**2. Learning Theory**

**3. Bias/variance tradeoff**

When talking about linear regression, we discussed the problem of whether to fit a "simple" model such as the linear " ," or a more "complex" model such as the polynomial " ." We saw the following example:



Fitting a 5th order polynomial to the data (rightmost figure) did not result in a good model. Specifically, even though the 5th order polynomial did a very good job predicting (say, prices of houses) from (say, living area) for the examples in the training set, we do not expect the model shown to be a good one for predicting the prices of houses not in the training set. In other words, what has been learned from the training set does not generalize well to other houses. The generalization error (which will be made formal shortly) of a hypothesis is its expected error on examples not necessarily in the training set.

Both the models in the leftmost and the rightmost figures above have large generalization error. However, the problems that the two models suffer from are very different. If the relationship between and is not linear, then even if we were fitting a linear model to a very large amount of training data, the linear model would still fail to accurately capture the structure in the data. Informally, we define the bias of a model to be the expected generalization error even if we were to fit it to a very (say, infinitely) large training set. Thus, for the problem above, the linear model suffers from large bias, and may underfit (i.e., fail to capture structure exhibited by) the data.

Apart from bias, there's a second component to the generalization error, consisting of the variance of a model fitting procedure. Specifically, when fitting a 5th order polynomial as in the rightmost figure, there is a large risk that we're fitting patterns in the data that happened to be present in our small, finite training set, but that do not reflect the wider pattern of the relationship between and . This could be, say, because in the training set we just happened by chance to get a slightly more-expensive-than-average house here, and a slightly less-expensive-than-average house there, and so on. By fitting these "spurious" patterns in the training set, we might again obtain a model with large generalization error. In this case, we say the model has large variance.

Often, there is a tradeoff between bias and variance. If our model is too "simple" and has very few parameters, then it may have large bias (but small variance); if it is too "complex" and has very many parameters, then it may suffer from large variance (but have smaller bias). In the example above, fitting a quadratic function does better than either of the extremes of a first or a fifth order polynomial.

**4. Preliminaries**

In this set of notes, we begin our foray into learning theory. Apart from being interesting and enlightening in its own right, this discussion will also help us hone our intuitions and derive rules of thumb about how to best apply learning algorithms in different settings. We will also seek to answer a few questions: First, can we make formal the bias/variance tradeoff that was just discussed? This will also eventually lead us to talk about model selection methods, which can, for instance, automatically decide what order polynomial to fit to a training set. Second, in machine learning it's really

In these notes, we will not try to formalize the definitions of bias and variance beyond this discussion. While bias and variance are straightforward to define formally for, e.g., linear regression, there have been several proposals for the definitions of bias and variance for classification, and there is as yet no agreement on what is the "right" and/or the most useful formalism. generalization error that we care about, but most learning algorithms fit their models to the training set. Why should doing well on the training set tell us anything about generalization error? Specifically, can we relate error on the training set to generalization error? Third and finally, are there conditions under which we can actually prove that learning algorithms will work well?

We start with two simple but very useful lemmas.

Lemma. (The union bound). Let be different events (that may not be independent). Then

In probability theory, the union bound is usually stated as an axiom (and thus we won't try to prove it), but it also makes intuitive sense: The probability of any one of events happening is at most the sum of the probabilities of the different events.

Lemma. (Hoeffding inequality) Let be independent and identically distributed (iid) random variables drawn from a distribution. I.e., , and . Let be the mean of these random variables, and let any be fixed. Then

This lemma (which in learning theory is also called the Chernoff bound) says that if we take - the average of random variables - to be our estimate of , then the probability of our being far from the true value is small, so long as is large. Another way of saying this is that if you have a biased coin whose chance of landing on heads is , then if you toss it times and calculate the fraction of times that it came up heads, that will be a good estimate of with high probability (if is large).

Using just these two lemmas, we will be able to prove some of the deepest and most important results in learning theory.

To simplify our exposition, let's restrict our attention to binary classification in which the labels are . Everything we'll say here generalizes to other problems, including regression and multi-class classification.

We assume we are given a training set of size , where the training examples are drawn iid from some probability distribution . For a hypothesis , we define the training error (also called the empirical risk or empirical error in learning theory) to be

This is just the fraction of training examples that misclassifies. When we want to make explicit the dependence of on the training set , we may also write this a . We also define the generalization error to be

I.e. this is the probability that, if we now draw a new example from the distribution will misclassify it.

Note that we have assumed that the training data was drawn from the same distribution with which we're going to evaluate our hypotheses (in the definition of generalization error). This is sometimes also referred to as one of the PAC assumptions.

Consider the setting of linear classification, and let . What's a reasonable way of fitting the parameters ? One approach is to try to minimize the training error, and pick

We call this process empirical risk minimization (ERM), and the resulting hypothesis output by the learning algorithm is . We think of ERM as the most "basic" learning algorithm, and it will be this algorithm that we focus on in these notes. (Algorithms such as logistic regression can also be viewed as approximations to empirical risk minimization.)

In our study of learning theory, it will be useful to abstract away from the specific parameterization of hypotheses and from issues such as whether we're using a linear classifier. We define the hypothesis class used by a learning algorithm to be the set of all classifiers considered by it. For linear classification, is thus the set of all classifiers over (the domain of the inputs) where the decision boundary is linear. More broadly, if we were studying, say, neural networks, then we could let be the set of all classifiers representable by some neural network architecture.

Empirical risk minimization can now be thought of as a minimization over the class of functions , in which the learning algorithm picks the hypothesis:

PAC stands for "probably approximately correct," which is a framework and set of assumptions under which numerous results on learning theory were proved. Of these, the assumption of training and testing on the same distribution, and the assumption of the independently drawn training examples, were the most important.

**5. The case of finite**

Let's start by considering a learning problem in which we have a finite hypothesis class consisting of hypotheses. Thus, is just a set of functions mapping from to , and empirical risk minimization selects to be whichever of these functions has the smallest training error.

We would like to give guarantees on the generalization error of . Our strategy for doing so will be in two parts: First, we will show that is a reliable estimate of for all . Second, we will show that this implies an upper-bound on the generalization error of .

Take any one, fixed, . Consider a Bernoulli random variable whose distribution is defined as follows. We're going to sample . Then, we set . I.e., we're going to draw one example, and let indicate whether misclassifies it. Similarly, we also define . Since our training set was drawn iid from and the 's have the same distribution.

We see that the misclassification probability on a randomly drawn examplethat is, -is exactly the expected value of (and ). Moreover, the training error can be written

Thus, is exactly the mean of the random variables that are drawn iid from a Bernoulli distribution with mean . Hence, we can apply the Hoeffding inequality, and obtain

This shows that, for our particular , training error will be close to generalization error with high probability, assuming is large. But we don't just want to guarantee that will be close to (with high probability) for just only one particular . We want to prove that this will be true simultaneously for all . To do so, let denote the event that . We've already shown that, for any particular , it holds true that . Thus, using the union bound, we have that

If we subtract both sides from 1, we find that

(The "ح" symbol means "not.") So, with probability at least , we have that will be within of for all . This is called a uniform convergence result, because this is a bound that holds simultaneously for all (as opposed to just one) .

In the discussion above, what we did was, for particular values of and , give a bound on the probability that for some . There are three quantities of interest here: , and the probability of error; we can bound either one in terms of the other two.

For instance, we can ask the following question: Given and some , how large must be before we can guarantee that with probability at least , training error will be within of generalization error? By setting and solving for , [you should convince yourself this is the right thing to do!], we find that if

then with probability at least , we have that for all . (Equivalently, this shows that the probability that for some is at most .) This bound tells us how many training examples we need in order make a guarantee. The training set size that a certain method or algorithm requires in order to achieve a certain level of performance is also called the algorithm's sample complexity.

The key property of the bound above is that the number of training examples needed to make this guarantee is only logarithmic in , the number of hypotheses in . This will be important later. Similarly, we can also hold and fixed and solve for in the previous equation, and show [again, convince yourself that this is right!] that with probability , we have that for all ,

Now, let's assume that uniform convergence holds, i.e., that for all . What can we prove about the generalization of our learning algorithm that picked ?

Define to be the best possible hypothesis in . Note that is the best that we could possibly do given that we are using , so it makes sense to compare our performance to that of . We have:

The first line used the fact that (by our uniform convergence assumption). The second used the fact that was chosen to minimize , and hence for all , and in particular . The third line used the uniform convergence assumption again, to show that . So, what we've shown is the following: If uniform convergence occurs, then the generalization error of is at most worse than the best possible hypothesis in !

Let's put all this together into a theorem.

Theorem. Let , and let any be fixed. Then with probability at least , we have that

This is proved by letting equal the term, using our previous argument that uniform convergence occurs with probability at least , and then noting that uniform convergence implies is at most higher than (as we showed previously).

This also quantifies what we were saying previously saying about the bias/variance tradeoff in model selection. Specifically, suppose we have some hypothesis class , and are considering switching to some much larger hypothesis class . If we switch to , then the first term can only decrease (since we'd then be taking a min over a larger set of functions). Hence, by learning using a larger hypothesis class, our "bias" can only decrease. However, if increases, then the second term would also increase. This increase corresponds to our "variance" increasing when we use a larger hypothesis class.

By holding and fixed and solving for like we did before, we can also obtain the following sample complexity bound:

Corollary. Let , and let any be fixed. Then for to hold with probability at least , it suffices that

**6. The case of infinite**

We have proved some useful theorems for the case of finite hypothesis classes. But many hypothesis classes, including any parameterized by real numbers (as in linear classification) actually contain an infinite number of functions. Can we prove similar results for this setting?

Let's start by going through something that is not the "right" argument. Better and more general arguments exist, but this will be useful for honing our intuitions about the domain.

Suppose we have an that is parameterized by real numbers. Since we are using a computer to represent real numbers, and IEEE double-precision floating point (double's in C) uses 64 bits to represent a floating point number, this means that our learning algorithm, assuming we're using doubleprecision floating point, is parameterized by bits. Thus, our hypothesis class really consists of at most different hypotheses. From the Corollary at the end of the previous section, we therefore find that, to guarantee , with to hold with probability at least , it suffices that . (The subscripts indicate that the last big- is hiding constants that may depend on and .) Thus, the number of training examples needed is at most linear in the parameters of the model.

The fact that we relied on 64-bit floating point makes this argument not entirely satisfying, but the conclusion is nonetheless roughly correct: If what we try to do is minimize training error, then in order to learn "well" using a hypothesis class that has parameters, generally we're going to need on the order of a linear number of training examples in .

(At this point, it's worth noting that these results were proved for an algorithm that uses empirical risk minimization. Thus, while the linear dependence of sample complexity on does generally hold for most discriminative learning algorithms that try to minimize training error or some approximation to training error, these conclusions do not always apply as readily to discriminative learning algorithms. Giving good theoretical guarantees on many non-ERM learning algorithms is still an area of active research.)

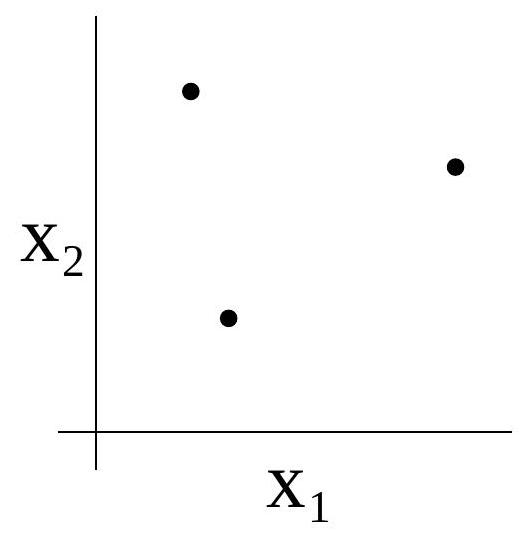
The other part of our previous argument that's slightly unsatisfying is that it relies on the parameterization of . Intuitively, this doesn't seem like it should matter: We had written the class of linear classifiers as , with parameters . But it could also be written with parameters . Yet, both of these are just defining the same : The set of linear classifiers in dimensions.

To derive a more satisfying argument, let's define a few more things.

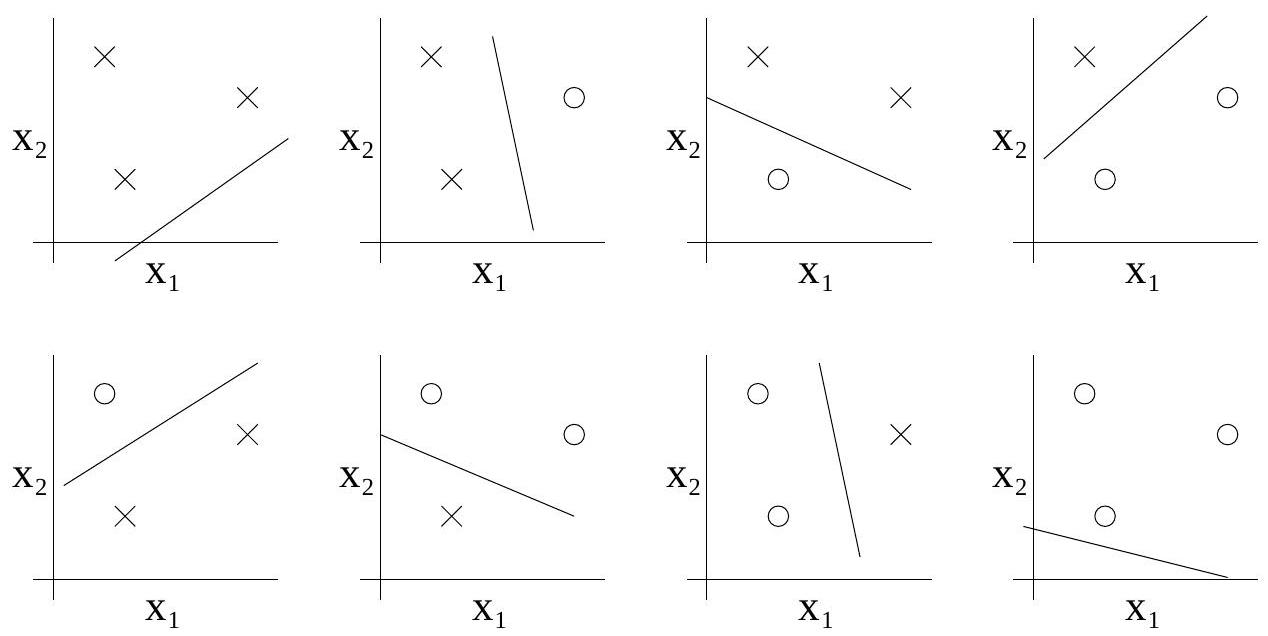
Given a set (no relation to the training set) of points , we say that shatters if can realize any labeling on . I.e., if for any set of labels , there exists some so that for all .

Given a hypothesis class , we then define its Vapnik-Chervonenkis dimension, written , to be the size of the largest set that is shattered by . (If can shatter arbitrarily large sets, then .)

For instance, consider the following set of three points:

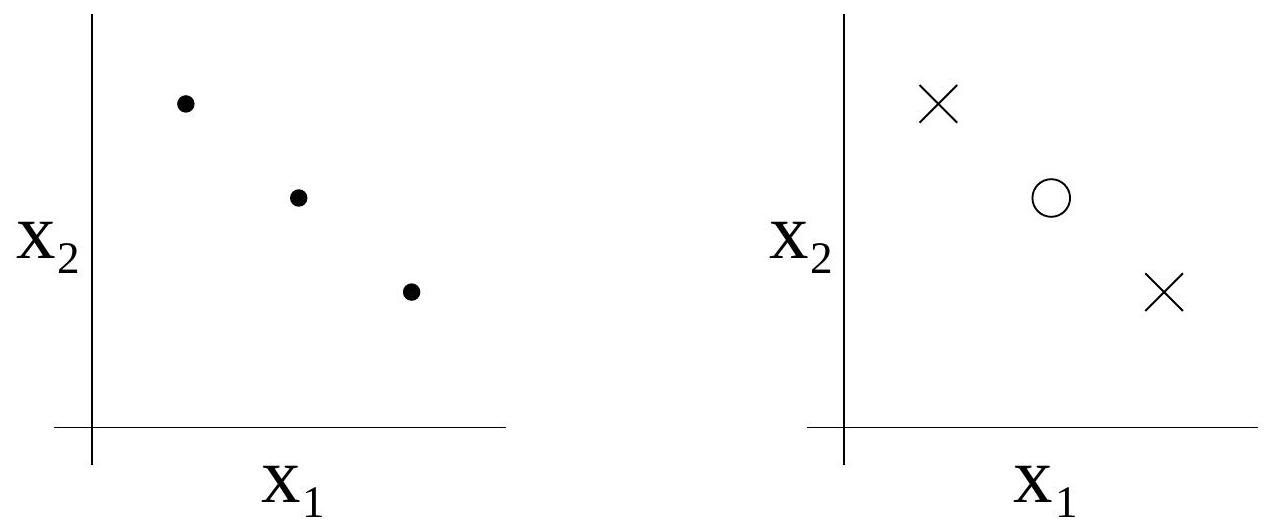


Can the set of linear classifiers in two dimensions can shatter the set above? The answer is yes. Specifically, we see that, for any of the eight possible labelings of these points, we can find a linear classifier that obtains "zero training error" on them:



Moreover, it is possible to show that there is no set of 4 points that this hypothesis class can shatter. Thus, the largest set that can shatter is of size 3 , and hence .

Note that the dimension of here is 3 even though there may be sets of size 3 that it cannot shatter. For instance, if we had a set of three points lying in a straight line (left figure), then there is no way to find a linear separator for the labeling of the three points shown below (right figure):



In order words, under the definition of the dimension, in order to prove that is at least , we need to show only that there's at least one set of size that can shatter.

The following theorem, due to Vapnik, can then be shown. (This is, many would argue, the most important theorem in all of learning theory.) Theorem. Let be given, and let . Then with probability at least , we have that for all ,

Thus, with probability at least , we also have that:

In other words, if a hypothesis class has finite dimension, then uniform convergence occurs as becomes large. As before, this allows us to give a bound on in terms of . We also have the following corollary:

Corollary. For to hold for all (and hence ) with probability at least , it suffices that .

In other words, the number of training examples needed to learn "well" using is linear in the dimension of . It turns out that, for "most" hypothesis classes, the dimension (assuming a "reasonable" parameterization) is also roughly linear in the number of parameters. Putting these together, we conclude that for a given hypothesis class (and for an algorithm that tries to minimize training error), the number of training examples needed to achieve generalization error close to that of the optimal classifier is usually roughly linear in the number of parameters of .