A Few Useful Things to Know about Machine Learning

ABSTRACT

Machine learning algorithms can figure out **how to perform important tasks by generalizing from examples**. This is often **feasible** and **cost-effective** where manual programming is not.
As **more data** becomes **available**, **more ambitious problems can be tackled**. As a result, machine learning is **widely used** in computer science and other fields.
However, developing successful machine learning applications requires a substantial amount of "black art" that is hard to find in textbooks. This article **summarizes twelve key lessons** that machine learning researchers and practitioners have learned. These **include pitfalls to avoid, important issues to focus on, and answers to common questions**.

1. INTRODUCTION

Machine learning systems **automatically learn programs from data**. This is often a very attractive alternative to manually constructing them, and in the last decade the use of machine learning has spread rapidly throughout computer science and beyond.

应用：Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other **applications**.

A recent report from the McKinsey Global Institute asserts that machine learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of innovation [16]. Several fine textbooks are available to interested practitioners and researchers (e.g, [17, 25]).
However, much of the "folk knowledge" that is needed to successfully develop machine learning applications is not readily available in them. As a result, many machine learning projects take much longer than necessary or wind up producing less than-ideal results. Yet much of this folk knowledge is fairly easy to communicate. This is the purpose of this article.

Many different types of machine learning exist, but for illustration purposes I will **focus on** the **most mature and widely used one**: **classification**. Nevertheless, **the issues** I will discuss **apply across all of machine learning**.
A classifier is a system that **inputs** (typically) **a vector of** discrete and/or continuous **feature values** and **outputs a single discrete value, the class**.
For example, a spam filter classifies email messages into "spam" or "not spam," and its input may be a Boolean vector $x = (x_1, \ldots, x_j, \ldots, x_d)$, where $x_j = 1$ if the jth word in the dictionary appears in the email and $x_j = 0$ otherwise.
A learner inputs a training set of examples (x_i, y_i), where $x_i = (x_{i,1}, \ldots, x_{i,d})$ is an observed input and y_i is the corresponding output, and outputs a classifier. The test of the learner is whether this classifier produces the correct output y_t for future examples x_t (e.g., whether the spam filter correctly classifies previously unseen emails as spam or not spam).

2. LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION

Suppose you have an application that you think machine learning might be good for. The **first problem facing you is the bewildering (扑朔迷离) variety of learning algorithms available. Which one to use?** There are literally thousands available, and hundreds more are published each year. The key to not getting lost in this huge space is to **realize that it consists of combinations of just three components**. The components are:

Representation. A classifier must be represented in some formal language that the computer can handle. Conversely, choosing a representation for a learner is tantamount (等于) to choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space of the learner. If a classifier is not in the hypothesis space, it cannot be learned. A related question, which we will address in a later section, is how to represent the input, i.e., what features to use.

Evaluation. An evaluation function (also called objective function or scoring function) is needed to distinguish good classifiers from bad ones. The evaluation function used internally by the algorithm may differ from the external one that we want the classifier to optimize, for ease of optimization (see below) and due to the issues discussed in the next section.

真正想用的evaluation function可能不可微

Optimization. Finally, we need a method to search among the classifiers in the language for the highest-scoring one. The choice of optimization technique is key to the efficiency of the learner, and also helps determine the classifier produced if the evaluation function has more than one optimum. It is common for new learners to start out using off-the-shelf（现成的） optimizers, which are later replaced by custom-designed ones.

Table 1: The three components of learning algorithms.

Representation	Evaluation	Optimization
Instances	Accuracy/Error rate	Combinatorial optimization
K-nearest neighbor	Precision and recall	Greedy search
Support vector machines	Squared error	Beam search
Hyperplanes	Likelihood	Branch-and-bound
Naive Bayes	Posterior probability	Continuous optimization
Logistic regression	Information gain	Unconstrained
Decision trees	K-L divergence	Gradient descent
Sets of rules	Cost/Utility	Conjugate gradient
Propositional rules	Margin	Quasi-Newton methods
Logic programs		Constrained
Neural networks		Linear programming
Graphical models		Quadratic programming
Bayesian networks		
Conditional random fields		

Q: SVM是instance？

用RBF kernel ，无法表示超平面，只能用support vector判断，

WWYQ: BnB是啥？

SXAQ: beam search?

In MT: 假设词表大小为3，包含[A, B, C]，Beam Width为2 生成第1个词的时候，对P(A)、P(B)、P(C)进行排序，选取概率最大的两个，假设为A，C 生成第2个词的时候，将当前序列

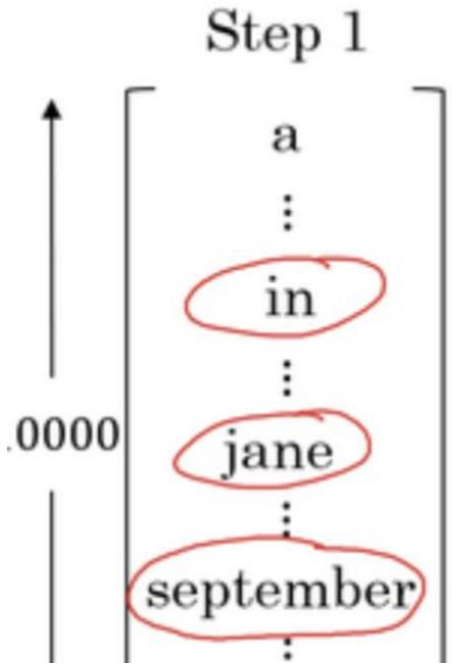
A，C分别和词表中的所有词进行组合，得到新的6个序列为AA、AB、AC，CA、CB、CC，然后同样取概率最大的两个作为当前序列，假设为AA、CC 重复以上的过程，直到遇到结束符为止，最终输出2个得分最高的序列

https://zhuanlan.zhihu.com/p/114669778

Seq2Seq中的beam search算法

https://zhuanlan.zhihu.com/p/36029811?group_id=972420376412762112

在***第一步***的时候，我们通过模型计算得到 $y^{<0>}$ 的分布概率，选择前B个作为候选结果，比如如下图所示的"in", "jane", "september"



第二步的时候，我们已经选择出了in、jane、September作为第一个单词的三个最可能选择，beam search针对每个第一个单词考虑第二个单词的概率，例如针对单词"in"，我们将 $y^{<1>}$ ='in'，然后将它喂给 $x^{<2>}$ ，输出结果 $y^{<2>}$ 作为第二个单词的概率输出。因为我们关注的是最有可能的 $P(y^{<2>}, y^{<1>}|x)$ ，因此我们的选择方法为：

$$P(y^{<2>}, "\ in\ "\ |x) = P(y^{<2>}|"\ in\ ", x)P("\ in\ "\ |x) \qquad (3)$$

然后同样将"jane"作为将 $y^{<1>}$ ，然后将它喂给 $x^{<2>}$ ，计算得到 $P(y^{<2>}|"\ jane\ ", x)$ ，然后计算得到：

$$P(y^{<2>}, "\ jane\ "\ |x) = P(y^{<2>}|"\ jane\ ", x)P("\ jane\ "\ |x) \qquad (4)$$

同样将"september"作为将 $y^{<1>}$ ，然后将它喂给 $x^{<2>}$ ，计算得到 $P(y^{<2>}|"\ september\ ", x)$ ，然后计算得到：

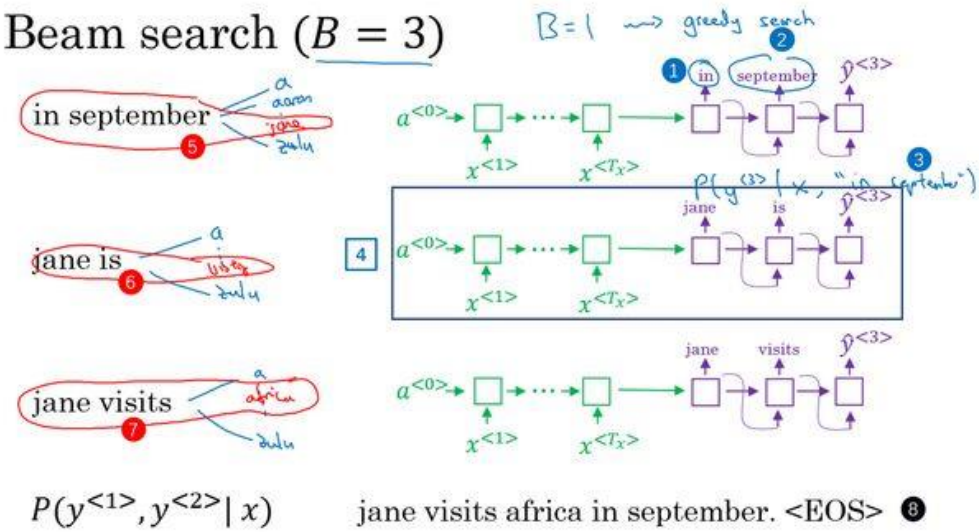
$$P(y^{<2>}, "\ september\ "\ |x) = P(y^{<2>}|"\ september\ ", x)P("\ september\ "\ |x) \quad (5)$$

这样我们就计算得到了 $B \times 10000 = 30000$ 个选择，那么选择前3个，比如得到的结果是：

- in september
- jane is
- jane visits

这样我们就找到了第一个和第二个单词对最可能的三个选择，这也意味着我们去掉了September作为英语翻译结果第一个单词的选择。

第三步的时候，同样我们将我们将 $y^{<1>}$ ='in'， $y^{<2>}$ ='september'，然后将它喂给 $x^{<3>}$ ，输出结果 $y^{<3>}$ 作为第三个单词的概率输出



这样我们得到前三的结果是：

- in september jane
- jane is visiting
- jane visits africa

第四步的时候同理，增加一个单词作为输入，这样最终会找到"**Jane visits africa in september**"这个句子，终止在句尾符号。

在集束宽为3时，集束搜索一次只考虑3个可能结果。注意如果集束宽等于1，只考虑1种可能结果，这实际上就变成了贪婪搜索算法，上面里我们已经讨论过了。但是如果同时考虑多个，可能的结果比如3个，10个或者其他的个数，集束搜索通常会找到比贪婪搜索更好的输出结果。

Table 1 shows common examples of each of these three components.
For example, k-nearest neighbor classifies a test example by finding the k most similar training examples and predicting the majority class among them.
Hyperplane based methods form a linear combination of the features per class and predict the class with the highest-valued combination.

Decision trees test one feature at each internal node, with one branch for each feature value, and have class predictions at the leaves.

Algorithm 1 LearnDT(*TrainSet*)

```
if all examples in TrainSet have the same class y* then
  return MakeLeaf(y*)
if no feature x_j has InfoGain(x_j,y) > 0 then
  y* ← Most frequent class in TrainSet
  return MakeLeaf(y*)
x* ← argmax_{x_j} InfoGain(x_j,y)
TS_0 ← Examples in TrainSet with x* = 0
TS_1 ← Examples in TrainSet with x* = 1
return MakeNode(x*, LearnDT(TS_0), LearnDT(TS_1))
```

Algorithm 1 shows a bare-bones decision tree learner for Boolean domains, using information gain and greedy search [21]. InfoGain(x_j, y) is the mutual information between feature x_j and the class y. MakeNode(x, c_0, c_1) returns a node that tests feature x and has c_0 as the child for x = 0 and c_1 as the child for x = 1.

SXAQ: InfoGain

A: 信息不确定性减少的程度。X(明天下雨)是一个随机变量，X的熵可以算出来， Y(明天阴天)也是随机变量，在阴天情况下下雨即是条件熵，X的熵减去Y条件下X的熵，就是信息增益。

Of course, not all combinations of one component from each column of Table 1 make equal sense. For example, discrete representations naturally go with combinatorial optimization, and continuous ones with continuous optimization. Nevertheless, many learners have both discrete and continuous components, and in fact the day may not be far when every single possible combination has appeared in some learner!

Most textbooks are organized by representation, and it's easy to overlook the fact that the other components are equally important. There is no simple recipe for choosing each component, but the next sections touch on some of the key issues. And, as we will see below, some choices in a machine learning project may be even more important than the choice of learner.

3. IT'S GENERALIZATION THAT COUNTS

The fundamental goal of machine learning is to generalize beyond the examples in the training set. This is because, no matter how much data we have, it is very unlikely that we will see those exact examples again at test time. (Notice that, if there are 100,000 words in the dictionary, the spam filter described above has $2^{100,000}$ possible different inputs.) Doing well on the training set is easy (just memorize the examples).

The most common mistake among machine learning beginners is to test on the training data and have the illusion of success. If the chosen classifier is then tested on new data, it is often no better than random guessing.
So, if you hire someone to build a classifier, be sure to keep some of the data to yourself and test the classifier they give you on it. Conversely, if you've been hired to build a classifier, set some of the data aside from the beginning, and only use it to test your chosen classifier at the very end, followed by learning your final classifier on the whole data.

Contamination of your classifier by test data can occur in insidious(隐蔽) ways, e.g., if you use test data to tune parameters and do a lot of tuning. (Machine learning algorithms have lots of knobs, and success often comes from twiddling them a lot, so this is a real concern.) Of course, holding out data reduces the amount available for training. This can be mitigated by doing cross-validation: randomly dividing your training data into (say) ten subsets, holding out each one while training on the rest, testing each learned classifier on the examples it did not see, and averaging the results to see how well the particular parameter setting does.

In the early days of machine learning, the need to keep training and test data separate was not widely appreciated. This was partly because, if the learner has a very limited representation (e.g., hyperplanes), the difference between training and test error may not be large. But with very flexible classifiers (e.g., decision trees), or even with linear classifiers with a lot of features, strict separation is mandatory.

Notice that generalization being the goal has an interesting consequence for machine learning. Unlike in most other optimization problems, we don't have access to the function we want to optimize! We have to use training error as a surrogate for test error, and this is fraught with danger. How to deal with it is addressed in some of the next sections. On the positive side, since the objective function is only a proxy(代理) for the true goal, we may not need to fully optimize it; in fact, a local optimum returned by simple greedy search may be better than the global optimum.

WWYQ: 什么是generalization error.

decomposing generalization error into bias and variance

4. DATA ALONE IS NOT ENOUGH

Generalization being the goal has another major consequence: data alone is not enough, no matter how much of it you have.

Consider learning a Boolean function of (say) 100 variables from a million examples. There are $100{,}000 - 2^{100,000}$ examples whose classes you don't know. How do you figure out what those classes are? **In the absence of further information, there is just no way to do this that beats flipping a coin.**

This observation was first made (in somewhat different form) by the philosopher David Hume over 200 years ago, but even today many mistakes in machine learning stem from failing to appreciate it.

Every learner must **embody some knowledge or assumptions beyond the data** it's given in order to generalize beyond it. This was formalized by Wolpert in his famous **"no free lunch" theorems**, according to which **no learner can beat random guessing over all possible functions to be learned** [26].

This seems like rather depressing news. How then can we ever hope to learn anything? Luckily, **the functions** we want to learn in the real world are **not drawn uniformly from the set of all mathematically possible functions**!

In fact, very general assumptions—**like smoothness, similar examples having similar classes, limited dependences, or limited complexity**—are often enough to do very well, and this is a large part of why machine learning has been so successful.

Like deduction, **induction (what learners do)** is a knowledge lever: it **turns a small amount of input knowledge into a large amount of output knowledge.** Induction is a vastly more powerful lever than deduction, **requiring much less input knowledge** to produce useful results, but it **still needs more than zero input knowledge** to work. And, as with any lever, **the more we put in, the more we can get out.**

A corollary of this is that **one of the key criteria for choosing a representation is which kinds of knowledge are easily expressed in it.**

For example, if we have a lot of knowledge about **what makes examples similar in our domain, instance based methods** may be a good choice.

If we have knowledge about **probabilistic dependencies, graphical models are a good fit.**

And if we have knowledge about **what kinds of preconditions are required by each class, "IF . . . THEN . . ." rules** may be the best option.

The most useful learners in this regard are those that don't just **have assumptions hard-wired into them**, but **allow us to state them explicitly, vary them widely, and incorporate them automatically into the learning** (e.g., using first-order logic [22] or grammars [6]).

In retrospect(回顾), the need for knowledge in learning should not be surprising. **Machine learning** is not magic; it **can't get something from nothing. What it does is get more from less.** Programming, like all engineering, is a lot of work: we have to build everything from scratch. Learning is more like farming, which lets nat

ure do most of the work. Farmers combine seeds with nutrients to grow crops. Learners combine knowledge with data to grow programs.

5. OVERFITTING HAS MANY FACES

What if the knowledge and data we have are not sufficient to completely determine the correct classifier? Then we run the risk of just hallucinating(幻觉的) a classifier (or parts of it) that is not grounded in reality, and is simply encoding random quirks in the data. This problem is called overfitting, and is the bugbear of machine learning. When your learner outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both, it has overfit.

Everyone in machine learning knows about overfitting, but it comes in many forms that are not immediately obvious. One way to understand overfitting is by decomposing generalization error into bias and variance [9].
Bias is a learner's tendency to consistently learn the same wrong thing.
Variance is the tendency to learn random things irrespective of the real signal.

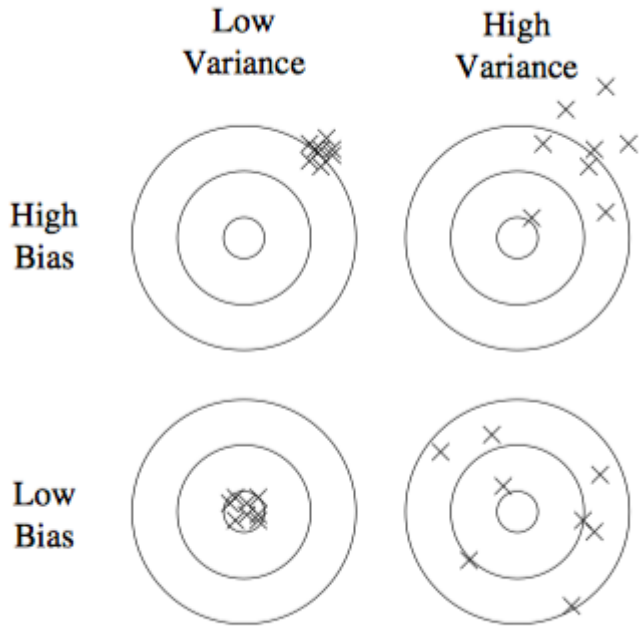


Figure 1: Bias and variance in dart-throwing.

Figure 1 illustrates this by an analogy with throwing darts at a board.
A linear learner has high bias, because when the frontier between two classes is not a hyperplane the learner is unable to induce it.
Decision trees don't have this problem because they can represent any Boolean function, but on the other hand they can suffer from high variance: decision trees learned on different training sets generated by the same phenomenon are often very different, when in fact they should be the same.

Similar reasoning applies to the choice of optimization method: beam search has lower bias than greedy search, but higher variance, because it tries more hypotheses. Thus, contrary to intuition, a more powerful learner is not necessarily better than a less powerful one.

Figure 2 illustrates this. [1]

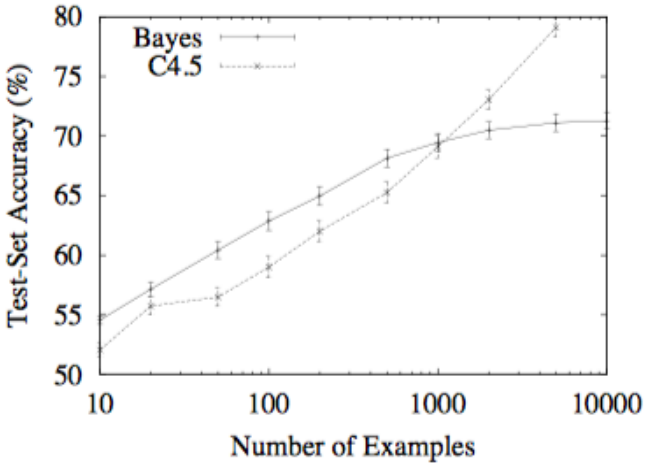


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules.

1 (Training examples consist of 64 Boolean features and a Boolean class computed from them according to a set of "IF . . . THEN . . ." rules. The curves are the average of 100 runs with different randomly generated sets of rules. Error bars are two standard deviations. See Domingos and Pazzani [11] for details.)

Even though the true classifier is a set of rules, with up to 1000 examples naive Bayes is more accurate than a rule learner. This happens despite naive Bayes' false assumption that the frontier is linear!
Situations like this are common in machine learning: strong false assumptions can be better than weak true ones, because a learner with the latter needs more data to avoid overfitting.

Cross-validation can help to combat overfitting, for example by using it to choose the best size of decision tree to learn. But it's no panacea(万能药), since if we use it to make too many parameter choices it can itself start to overfit [18].

Besides cross-validation, there are many methods to combat overfitting. The most popular one is adding a regularization term to the evaluation function. This can, for example, penalize classifiers with more structure, thereby favoring smaller ones with less room to overfit.

Another option is to **perform a statistical significance test like chi-square before adding new structure**, to decide **whether the distribution of the class really is different with** and without **this structure**.
These techniques are particularly **useful when data is very scarce**. Nevertheless, you should be skeptical(怀疑)of claims that a particular technique "solves" the overfitting problem.
It's **easy to avoid overfitting (variance) by falling into the opposite error of underfitting (bias)**.
Simultaneously avoiding both requires learning a perfect classifier, and short of knowing it in advance there is no single technique that will always do best (no free lunch).

A common **misconception** about **overfitting** is that it **is caused by noise**, like training examples labeled with the wrong class. This can indeed **aggravate overfitting**, by making the learner draw a capricious(反复无常) frontier to keep those examples on what it thinks is the right side.
But severe **overfitting can occur even in the absence of noise**.
For instance, suppose we learn a Boolean classifier that is just the disjunction(析取) of the examples labeled "true" in the training set. (In other words, the classifier is a Boolean formula in disjunctive normal form, where each term is the conjunction of the feature values of one specific training example). This classifier gets all the training examples right and every positive test example wrong, regardless of whether the training data is noisy or not.

The problem of multiple testing [14] **is closely related to overfitting**.
Standard statistical tests assume that **only one hypothesis is being tested**, but **modern learners** can easily **test millions before they are done**.
As a result **what looks significant may in fact not be**. ??
For example, a mutual fund that beats the market ten years in a row looks very impressive, until you realize that, if there are 1000 funds and each has a 50% chance of beating the market on any given year, it's quite likely that one will succeed all ten times just by luck.
This problem can be **combatted by correcting the significance tests to take the number of hypotheses into account**, but this can **lead to underfitting**.
A **better approach** is to **control the fraction of falsely accepted non-null hypotheses(控制错误接受的非零假设的比率)**, known as the **false discovery rate** [3].

6. INTUITION FAILS IN HIGH DIMENSIONS

After overfitting, **the biggest problem** in machine learning is the curse of **dimensionality**. This expression was coined by Bellman in 1961 to refer to the fact that m

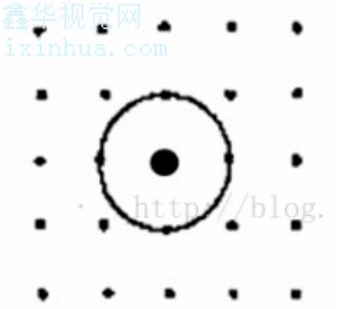
any algorithms that work fine in low dimensions become intractable when the input is high-dimensional. But in machine learning it refers to much more.
Generalizing correctly becomes exponentially harder as the dimensionality (number of features) of the examples **grows**, because a **fixed-size training set covers a dwindling(逐渐减少) fraction of the input space.** ??? Even with a moderate dimension of 100 and a huge training set of a trillion examples, the latter covers only a fraction of about 10^{-18} of the input space. This is what makes machine learning both necessary and hard.

WWYQ： what is excatly the curse of dimensionality?

More seriously, **the similarity-based reasoning** that machine learning algorithms depend on (explicitly or implicitly) **breaks down in high dimensions.**
Consider a **nearest neighbor classifier with Hamming distance** as the similarity measure, and suppose the class is just $x_1 \wedge x_2$. If there are no other features, this is an easy problem. But if there are 98 irrelevant features $x_3, \ldots, x_{100}$, the noise from them completely swamps the signal in x_1 and x_2, and nearest neighbor effectively makes random predictions.

Q：Hamming distance
是两个字符串对应位置的不同字符的个数。就是将一个字符串变换成另外一个字符串所需要*替换*的字符个数。1011101与1001001之间的汉明距离是2。

Even more disturbing is **that nearest neighbor still has a problem even if** all 100 **features are relevant! This is because in high dimensions all examples look alike.**
Suppose, for instance, that examples are laid out on a regular grid, and consider a test example x_t. If the grid is d-dimensional, x_t's 2d nearest examples are all at the same distance from it. So as the dimensionality increases, more and more examples become nearest neighbors of x_t, until the choice of nearest neighbor (and therefore of class) is effectively random.
2D 例子：



This is only one instance of a more general problem with high dimensions: our intuitions, which come from a three-dimensional world, often do not apply in high-dimensional ones.
In high dimensions, most of the mass of a multivariate Gaussian distribution is not near the mean, but in an increasingly distant "shell" around it; and most of the volume of a high-dimensional orange is in the skin, not the pulp(果肉).
If a constant number of examples is distributed uniformly in a high-dimensional hypercube, beyond some dimensionality most examples are closer to a face of the hypercube than to their nearest neighbor.
And if we approximate a hypersphere by inscribing it in a hypercube, in high dimensions almost all the volume of the hypercube is outside the hypersphere.
This is bad news for machine learning, where shapes of one type are often approximated by shapes of another.

Building a classifier in two or three dimensions is easy; we can find a reasonable frontier between examples of different classes just by visual inspection. (It's even been said that if people could see in high dimensions machine learning would not be necessary.) But in high dimensions it's hard to understand what is happening. This in turn makes it difficult to design a good classifier. Naively, one might think that gathering more features never hurts, since at worst they provide no new information about the class. But in fact their benefits may be outweighed by the curse of dimensionality.

Fortunately, there is an effect that partly counteracts the curse, which might be called the "blessing of non-uniformity." In most applications examples are not spread uniformly throughout the instance space, but are concentrated on or near a lower-dimensional manifold. For example, k-nearest neighbor works quite well for handwritten digit recognition even though images of digits have one dimension per pixel, because the space of digit images is much smaller than the space of all possible images. Learners can implicitly take advantage of this lower effective dimension, or algorithms for explicitly reducing the dimensionality can be used (e.g., [23]).

7. THEORETICAL GUARANTEES ARE NOT WHAT THEY SEEM

Machine learning papers are full of theoretical guarantees. The most common type is a bound on the number of examples needed to ensure good generalization. What should you make of these guarantees? First of all, it's remarkable that they are even possible. Induction is traditionally contrasted with deduction: in deduction you can guarantee that the conclusions are correct; in induction all bets are off (一切都很难说). Or such was the conventional wisdom for many centuries. One of the major developments of recent decades has been the realization that in fact we can

have guarantees on the results of induction, particularly if we're willing to settle for probabilistic guarantees.

The basic argument is remarkably simple [5]. Let's say a classifier is bad if it's true error rate is greater than ε. Then the probability that a bad classifier is consistent with n random, independent training examples is less than $(1-\varepsilon)^n$. Let b be the number of bad classifiers in the learner's hypothesis space H. The probability that at least one of them is consistent is less than $b(1-\varepsilon)^n$, by the union bound. Assuming the learner always returns a consistent classifier, the probability that this classifier is bad is then less than $|H|(1-\varepsilon)^n$, where we have used the fact that $b \leq |H|$. So if we want this probability to be less than δ, it suffices to make $n > \ln(\delta/|H|)/\ln(1-\varepsilon) \geq 1/\varepsilon(\ln|H| + \ln 1/\delta)$

Unfortunately, guarantees of this type have to be taken with a large grain of salt (十分小心). This is because the bounds obtained in this way are usually extremely loose.

The wonderful feature of the bound above is that the required number of examples only grows logarithmically with $|H|$ and $1/\delta$.

Unfortunately, most interesting hypothesis spaces are doubly exponential in the number of features d, which still leaves us needing a number of examples exponential in d. For example, consider the space of Boolean functions of d Boolean variables. If there are e possible different examples, there are 2^e possible different functions, so since there are 2^d possible examples, the total number of functions is 2^{2^d}.

And even for hypothesis spaces that are "merely" exponential, the bound is still very loose, because the union bound is very pessimistic(悲观). For example, if there are 100 Boolean features and the hypothesis space is decision trees with up to 10 levels, to guarantee $\delta = \varepsilon = 1\%$ in the bound above we need half a million examples. But in practice a small fraction of this suffices for accurate learning.

Further, we have to be careful about what a bound like this means. For instance, it does not say that, if your learner returned a hypothesis consistent with a particular training set, then this hypothesis probably generalizes well. What it says is that, given a large enough training set, with high probability your learner will either return a hypothesis that generalizes well or be unable to find a consistent hypothesis.

The bound also says nothing about how to select a good hypothesis space. It only tells us that, if the hypothesis space contains the true classifier, then the probability that the learner outputs a bad classifier decreases with training set size.

If we shrink the hypothesis space, the bound improves, but the chances that it contains the true classifier shrink also. (There are bounds for the case where the true classifier is not in the hypothesis space, but similar considerations apply to them.)

Another common type of theoretical guarantee is asymptotic: given infinite data, the learner is guaranteed to output the correct classifier. This is reassuring(放心), but it would be rash to choose one learner over another because of its asymptotic guarantees. In practice, we are seldom in the asymptotic regime(渐进体制) (also known as "asymptopia"). And, because of the bias-variance tradeoff we discussed above, if learner A is better than learner B given infinite data, B is often better than A given finite data.

The main role of theoretical guarantees in machine learning is not as a criterion for practical decisions, but as a source of understanding and driving force for algorithm design. In this capacity, they are quite useful; indeed, the close interplay of theory and practice is one of the main reasons machine learning has made so much progress over the years.
But caveat emptor: learning is a complex phenomenon, and just because a learner has a theoretical justification and works in practice doesn't mean the former is the reason for the latter.

8. FEATURE ENGINEERING IS THE KEY

At the end of the day, some machine learning projects succeed and some fail. What makes the difference? Easily the most important factor is the features used. If you have many independent features that each correlate well with the class, learning is easy. On the other hand, if the class is a very complex function of the features, you may not be able to learn it. Often, the raw data is not in a form that is amenable(合适的) to learning, but you can construct features from it that are. This is typically where most of the effort in a machine learning project goes. It is often also one of the most interesting parts, where intuition, creativity and "black art" are as important as the technical stuff.

First-timers are often surprised by how little time in a machine learning project is spent actually doing machine learning. But it makes sense if you consider how time-consuming it is to gather data, integrate it, clean it and pre-process it, and how much trial and error can go into feature design.
Also, machine learning is not a one-shot(一站式) process of building a data set and running a learner, but rather an iterative process of running the learner, analyzing the r

esults, modifying the data and/or the learner, and repeating. Learning is often th e quickest part of this, but that's because we've already mastered it pretty wel l! **Feature engineering** is more difficult because **it's domain-specific, while learn ers can be largely general-purpose.** However, there is no sharp frontier between t he two, and this is another reason **the most useful learners are those that facili tate incorporating knowledge.**

Of course, one of the holy grails of **machine learning is to automate more and mor e of the feature engineering process.** One way this is **often** done today is by **auto matically generating large numbers of candidate features and selecting the best b y (say) their information gain with respect to the class.** But bear in mind that **f eatures that look irrelevant in isolation may be relevant in combination.**(孤立地看起来不相关的功能可能会组合在一起使用) For example, if the class is an XOR of k input features, each of them by itself carries no information about the class. (If you want to annoy machine learners, bring up XOR.)

INPUT		OUTPUT
A	B	A XOR B
0	0	0
0	1	1
1	0	1
1	1	0

On the other hand, **running a learner with a very large number of features to find out which ones are useful in combination may be too time-consuming, or cause over fitting.** So there is ultimately no replacement for the smarts you put into feature engineering.

9. MORE DATA BEATS A CLEVERER ALGORITHM 数据多比算法好更重要

Suppose you've constructed the best set of features you can, but the classifiers you're getting are still not accurate enough. What can you do now? There are **two main choices: design a better learning algorithm, or gather more data** (more examples, and possibly more raw features, subject to the curse of dimensionality不会造成维度灾难的更多可能的原始特征). Machine learning **researchers are mainly concerned w ith the former,** but pragmatically(务实地) **the quickest path to success is often to just get more data.**

As a rule of thumb(根据经验), **a dumb algorithm with lots and lots of data beats a c lever one with modest amounts of it.** (After all, machine learning is all about **le tting data do the heavy lifting.**)

This does bring up **another problem**, however: **scalability(可扩展性)**. In most of computer science, the **two main limited resources are time and memory.** In machine learning, there is **a third one: training data.**

Which one is the bottleneck has changed from decade to decade. In the 1980's it tended to be data. Today it is often time. Enormous mountains of data are available, but there is not enough time to process it, so it goes unused. This leads to a **paradox: even though in principle more data means that more complex classifiers can be learned, in practice simpler classifiers wind up being used, because complex ones take too long to learn.**

Part of the answer is to **come up with fast ways to learn complex classifiers**, and indeed there has been remarkable progress in this direction (e.g., [12]).

Part of the reason using cleverer algorithms has a smaller payoff than you might expect is that, to a first approximation, they all do the same. This is surprising when you consider representations as different as, say, sets of rules and neural networks. But in fact **propositional rules are readily encoded as neural networks**, and **similar relationships hold between other representations.**

All learners essentially work by grouping nearby examples into the same class; the **key difference is in the meaning of "nearby."**

With **non-uniformly distributed data, learners can produce widely different frontiers while still making the same predictions in the regions that matter** (those with a substantial(大量的) number of training examples, and therefore also where most test examples are likely to appear). This also helps **explain why powerful learners can be unstable but still accurate.** Figure 3 illustrates this in 2-D; **the effect is much stronger in high dimensions.**

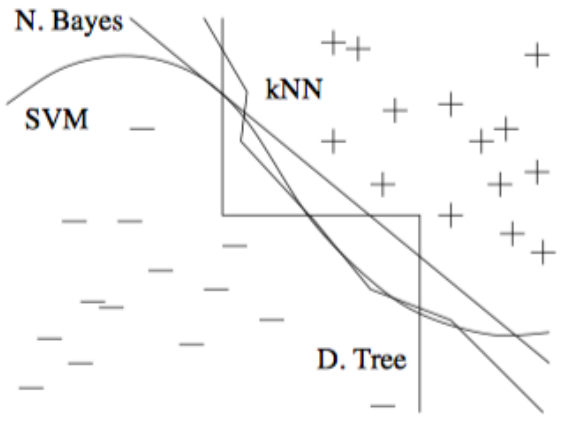


Figure 3: Very different frontiers can yield similar class predictions. (+ and − are training examples of two classes.)

As a rule, it pays to try the simplest learners first (e.g., naive Bayes before logistic regression, k-nearest neighbor before support vector machines). More sophisticated learners are seductive(诱人), but they are usually harder to use, because they have more knobs you need to turn to get good results, and because their internals are more opaque(不透明).

Learners can be divided into two major types: those whose representation has a fixed size, like linear classifiers, and those whose representation can grow with the data, like decision trees. (The latter are sometimes called non-parametric learners, but this is somewhat unfortunate, since they usually wind up learning many more parameters than parametric ones.)

non-parametric：k-Nearest Neighbors，Decision Trees，Support Vector Machines

parametric：Logistic Regression， Perceptron，Naive Bayes，Simple Neural Networks

Fixed-size learners can only take advantage of so much data. (Notice how the accuracy of naive Bayes asymptotes at around 70% in Figure 2.)

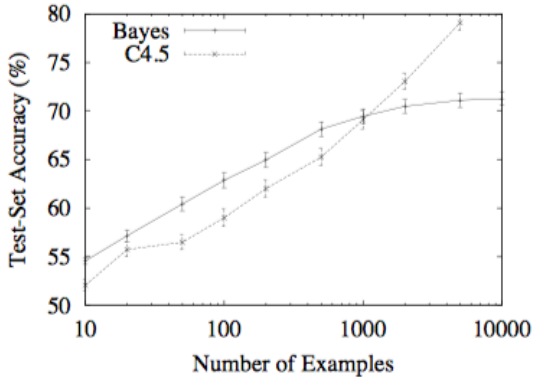


Figure 2: Naive Bayes can outperform a state-of-the-art rule learner (C4.5rules) even when the true classifier is a set of rules.

Q: Naive Bayes 分类器长什么样子?

Decision rules for binary classification

- Want to predict $y = \arg\max_{y'} P(y' \mid \mathbf{x})$

- For binary tasks (i.e., c=2, $y \in \{+1, -1\}$), this is equivalent to

$$y = \text{sign}\left(\log \frac{P(Y=1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})}\right)$$

P

$f(x)$

$\log \frac{0.5}{0.5} = \log 1 = 0$; >1 ; $>0 \rightarrow +$

easy to verify that the above gives you $\begin{cases} +1 & \text{if } p > 0.5 \\ -1 & \text{o.w.} \end{cases}$

- The function $f(\mathbf{x}) = \log \frac{P(Y=1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})}$

  is called discriminant function

Discriminant functions for GBCs

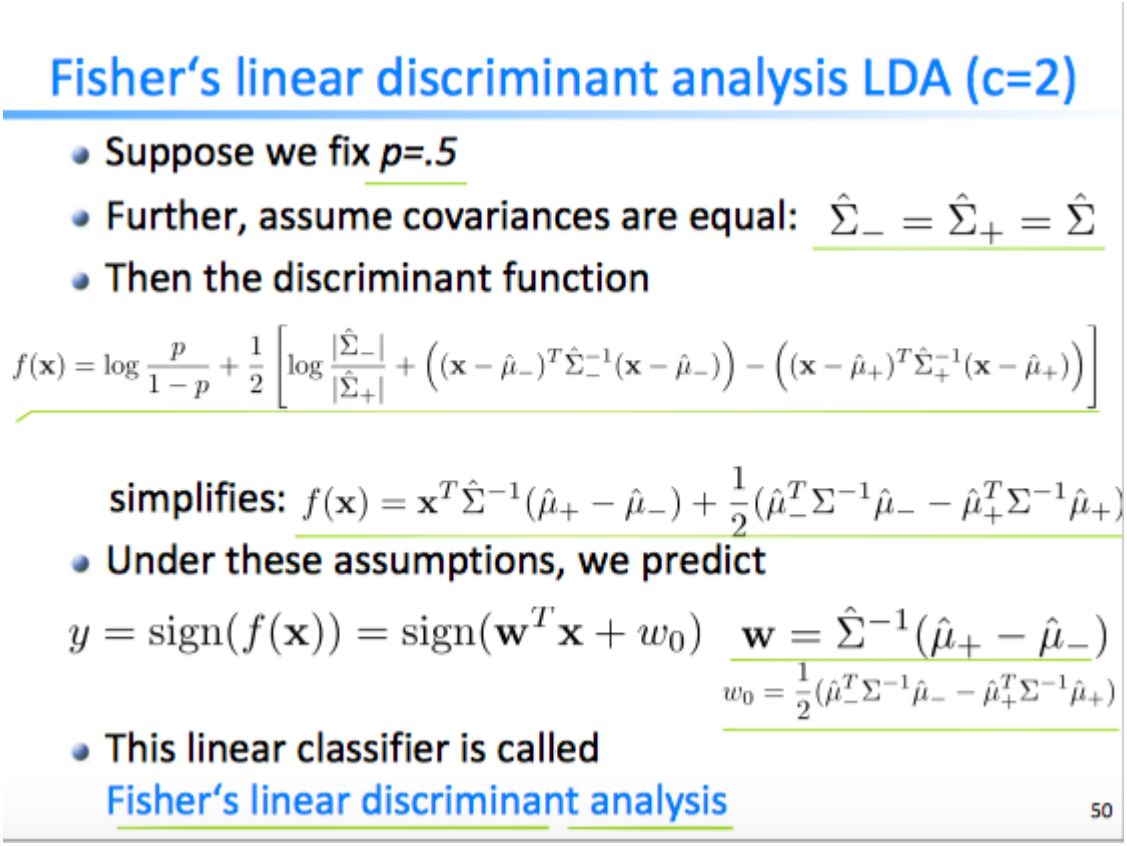
- Given: $P(Y=1) = p$ and $P(\mathbf{x} \mid y) = \mathcal{N}(\mathbf{x}; \mu_y, \Sigma_y)$
- Want:

$$f(\mathbf{x}) = \log \frac{P(Y=1 \mid \mathbf{x})}{P(Y=-1 \mid \mathbf{x})}$$

- This discriminant function is given by

$$f(\mathbf{x}) = \log \frac{p}{1-p} + \frac{1}{2}\left[\log \frac{|\hat{\Sigma}_-|}{|\hat{\Sigma}_+|} + \left((\mathbf{x}-\hat{\mu}_-)^T \hat{\Sigma}_-^{-1}(\mathbf{x}-\hat{\mu}_-)\right) - \left((\mathbf{x}-\hat{\mu}_+)^T \hat{\Sigma}_+^{-1}(\mathbf{x}-\hat{\mu}_+)\right)\right]$$

$Y = \{-1, +1\} \rightarrow$ predict $\text{sign}(f(x))$



Variable-size learners can **in principle learn any function given sufficient data, but in practice they may not**, because of **limitations of the algorithm** (e.g., greedy search falls into local optima) or **computational cost**. Also, **because of the curse of dimensionality, no existing amount of data may be enough.**

For these reasons, clever algorithms—those that make the most of the data and computing resources available—often pay off in the end, provided you're willing to put in the effort. There is no sharp frontier between designing learners and learning classifiers; rather, **any given piece of knowledge could be encoded in the learner or learned from data.**

So machine learning projects often wind up having a significant component of learner design, and practitioners need to have some expertise in it [13].

In the end, **the biggest bottleneck is** not data or CPU cycles, but **human cycles.** In research papers, learners are typically compared on measures of accuracy and computational cost. But **human effort saved and insight gained**, although harder to measure, **are often more important.** This **favors learners that produce human-understandable output** (e.g., rule sets).**(这使那些产生人类可理解的输出的学习器更为受到青睐。)**

And the organizations that make the most of machine learning are those that have in place an infrastructure that makes experimenting with many different learners, data sources and learning problems easy and efficient, and where there is a close collaboration between machine learning experts and application domain ones.**(机器学习**

成果最丰硕的，是那些建立了机器学习的基本条件，能够便捷地在多个学习器、数据来源和学习问题上方便有效地开展实验，并实现机器学习专家与领域专家的密切合作的组织。)

10. LEARN MANY MODELS, NOT JUST ONE

In the early days of machine learning, everyone had their favorite learner, together with some a priori reasons to believe in its superiority. Most effort went into trying many variations of it and selecting the best one.

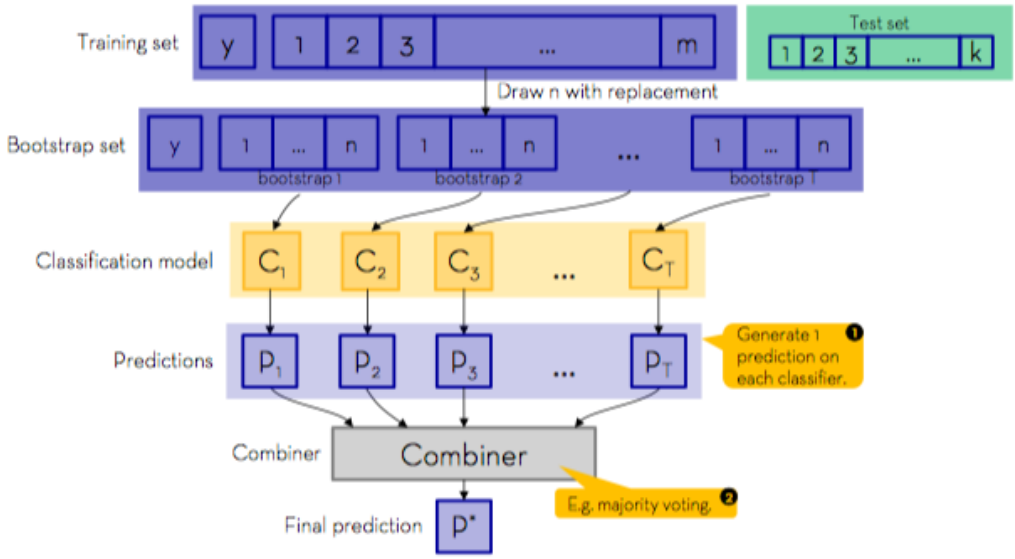
Then systematic empirical comparisons showed that the best learner varies from application to application, and systems containing many different learners started to appear.

Effort now went into trying many variations of many learners, and still selecting just the best one.

But then researchers noticed that, if instead of selecting the best variation found, we **combine many variations, the results are better**—often much better—and at **little extra effort** for the user.

Creating such **model ensembles** is now standard [1].

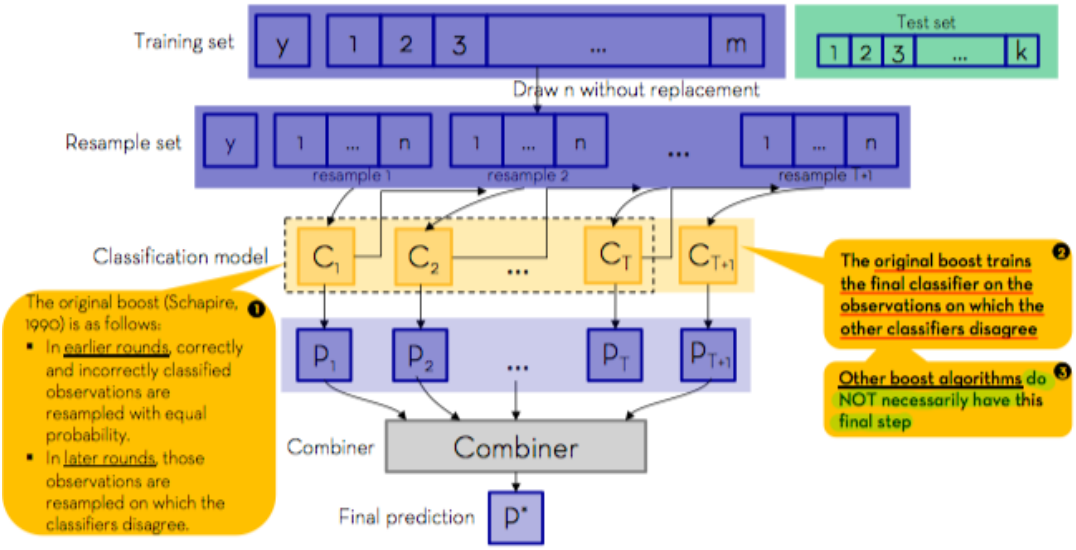
In the simplest technique, called **bagging**, we simply **generate random variations of the training set by resampling, learn a classifier on each, and combine the results by voting.** This works because **it greatly reduces variance while only slightly increasing bias.**



In **boosting, training examples have weights**, and these are varied so that **each new classifier focuses on the examples the previous ones tended to get wrong.**

Alternative boosting method: The original boost

35



Advantages and disadvantages of boosting

60

Advantages

- Misclassified observations are weighted in a way that they get properly classified in the future.
- Reduces variance and bias of the base classifier.
- Boosts can be applied on both strong and weak learners. Training a weak learner is faster.

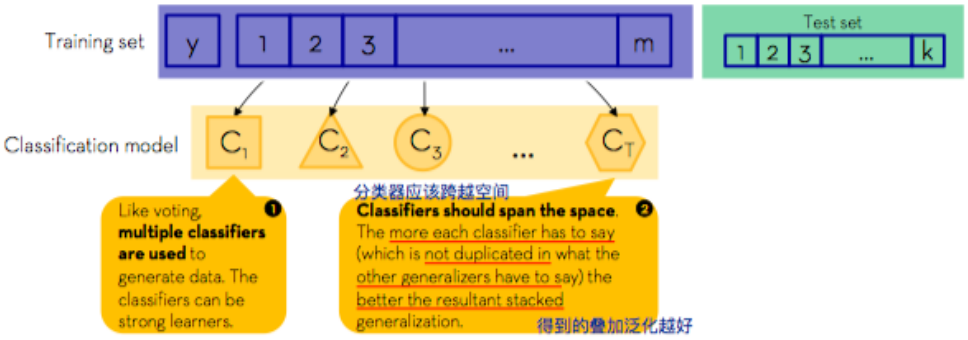
Disadvantages

- More common for boosting to hurt performance:
  - It is possible that if the ensemble size is too large that boosting can create a very complex classifier that overfits.
  - Boosting may have difficulty with noisy data.

In stacking, the outputs of individual classifiers become the inputs of a "higher-level" learner that figures out how best to combine them.

Stacking:
Train T different classifiers on the training data

82



Many other techniques exist, and **the trend is toward larger and larger ensembles.** In the Netflix prize, teams from all over the world competed to build the best video recommender system (http://netflixprize.com). As the competition progressed, teams found that they obtained the best results by combining their learners with other teams', and merged into larger and larger teams. The winner and runner-up were both stacked ensembles of over 100 learners, and combining the two ensembles further improved the results. Doubtless we will see even larger ones in the future.

Model ensembles should not be confused with Bayesian model averaging (BMA). BMA is the **theoretically optimal approach to learning** [4].

Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences. Bayesian model averaging (BMA) provides a coherent mechanism for accounting for this model uncertainty when deriving parameter estimates.

In brief, BMA marginalizes over models to derive posterior densities on model parameters that account for model uncertainty, as follows:

$$p(\theta \mid y) = \sum_{m_i} p(m_i \mid y) p(\theta \mid y, m_i)$$

where m_i are the set of candidate models, $p(m_i \mid y)$ is the posterior probability over model m_i, and $p(\theta \mid y, m_i)$ is the posterior density on model parameters conditional on model m_i. The latter posterior density is a decent proxy for one's information on parameters θ only if $p(m_i \mid y) \approx 1$. Otherwise, uncertainty regarding the correct model will automatically translate in uncertainty regarding model parameters…

In BMA, **predictions on new examples are made by averaging the individual predictions of all classifiers in the hypothesis space, weighted by how well the classifiers explain the training data and how much we believe in them a priori.**

Despite their superficial similarities, ensembles and BMA are very different. **Ensembles change the hypothesis space** (e.g., from single decision trees to linear combinations of them), and can take a wide variety of forms. **BMA assigns weights to the hypotheses in the original space according to a fixed formula.**

BMA weights are extremely different from those produced by (say) bagging or boosting: the latter are fairly even, while the former **are extremely skewed**, to the point where **the single highest-weight classifier usually dominates**, making BMA effectively equivalent to just selecting it [8].

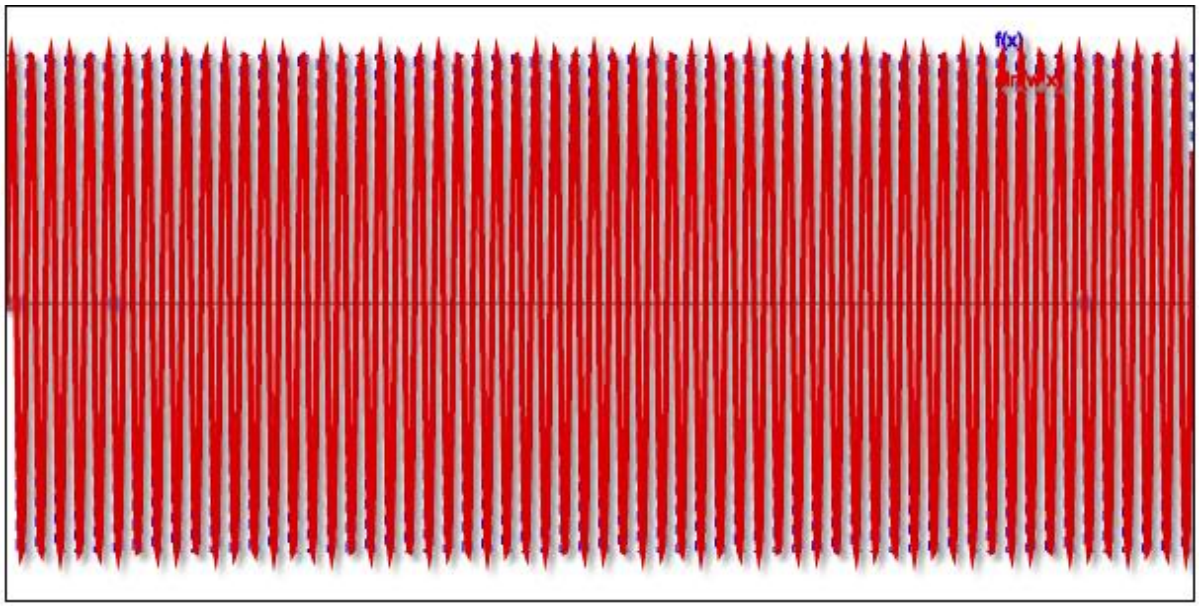
A practical consequence of this is that, while model ensembles are a key part of the machine learning toolkit, BMA is seldom worth the trouble.

11. SIMPLICITY DOES NOT IMPLY ACCURACY

Occam's razor famously states that entities should not be multiplied beyond necessity.(**若无必要，勿增实体**) In machine learning, this is often taken to mean that, **given two classifiers with the same training error, the simpler** of the two will likely **have the lowest test error.** Purported proofs of this claim appear regularly in the literature, but in fact there are many counterexamples to it, and the "no free lunch" theorems imply **it cannot be true.**

大意：Occam's razor不对

We saw one **counter-example** in the previous section: **model ensembles**. The generalization error of a boosted ensemble continues to improve by adding classifiers even after the training error has reached zero.
Another counter-example is **support vector machines**, which can **effectively have an infinite number of parameters without overfitting.**
Conversely, the function sign(sin(ax)) can discriminate an arbitrarily large, arbitrarily labeled set of points on the x axis, even though it has only one parameter [24].



Thus, contrary to intuition, there is **no necessary connection between the number of parameters of a model and its tendency to overfit.**

A more sophisticated view instead **equates complexity with the size of the hypothesis space**, **(取而代之的是，一个更复杂的视图将复杂性与假设空间的大小等同起来)**on the basis that **smaller spaces allow hypotheses to be represented by shorter codes.**
Bounds like the one in the section **on theoretical guarantees** above **might then be viewed as implying that shorter hypotheses generalize better(暗示着较短的假设可以更好地推广).** This can be further **refined by assigning shorter codes to the hypothesis in the space that we have some a priori preference for. (为有先验偏好的空间中的假设分配更短的代码)**
But **viewing this as "proof" of a tradeoff between accuracy and simplicity is circular reasoning(但如果将此看作是准确和简单之间权衡的"证明"，那就变成循环论证了)**: we made the **hypotheses** we prefer simpler by design, and **if they are accurate it's because our preferences are accurate**, not because the hypotheses are "simple" in the representation we chose.

我们喜欢简单的设计，所以搞了个hypo,如果它很准，说明我们的偏好没毛病；而不是说这个hypo本身在我们的偏好中就是简单的

A further complication arises from the fact that few learners search their hypothesis space exhaustively. A learner with a larger hypothesis space that tries fewer hypotheses from it is less likely to overfit than one that tries more hypotheses from a smaller space.(大空间搜索的假设少 比 小空间搜索的假设多 不容易overfit)

As Pearl [19] points out, the size of the hypothesis space is only a rough guide to what really matters for relating training and test error: the procedure by which a hypothesis is chosen.

在空间选假设避免overfit，空间大小只是一个参考线索

Domingos [7] surveys the main arguments and evidence on the issue of Occam's razor in machine learning. The conclusion is that simpler hypotheses should be preferred because simplicity is a virtue in its own right, not because of a hypothetical connection with accuracy. This is probably what Occam meant in the first place.

12. REPRESENTABLE DOES NOT IMPLY LEARNABLE 可表示出来，不代表可以被学

Essentially all representations used in variable-size learners have associated theorems of the form "Every function can be represented, or approximated arbitrarily closely, using this representation." Reassured by this, fans of the representation often proceed to ignore all others.

Universal Approximation Theorem

However, just because a function can be represented does not mean it can be learned. For example, standard decision tree learners cannot learn trees with more leaves than there are training examples. 50个数据不能学出100层来

In continuous spaces, representing even simple functions using a fixed set of primitives(基元集) often requires an infinite number of components. ？？？？

Further, if the hypothesis space has many local optima of the evaluation function, as is often the case, the learner may not find the true function even if it is representable.

Given finite data, time and memory, standard learners can learn only a tiny subset of all possible functions, and these subsets are different for learners with different representations.

Therefore the key question is not "Can it be represented?", to which the answer is often trivial, but "Can it be learned?" And it pays to try different learners (and possibly combine them).

Some representations are exponentially more compact than others for some functions. As a result, they may also require exponentially less data to learn those functions.

Many learners work by forming linear combinations of simple basis functions. For example, support vector machines form combinations of kernels centered at some of the training examples (the support vectors).

Representing parity(奇偶性) of n bits in this way requires 2^n basis functions. But using a representation with more layers (i.e., more steps between input and output), parity can be encoded in a linear-size classifier.

Finding methods to learn these deeper representations is one of the major research frontiers in machine learning [2].

13. CORRELATION DOES NOT IMPLY CAUSATION

The point that correlation does not imply causation is made so often that it is perhaps not worth belaboring(迷恋). But, even though learners of the kind we have been discussing can only learn correlations, their results are often treated as representing causal relations. Isn't this wrong? If so, then why do people do it?

More often than not, the goal of learning predictive models is to use them as guides to action. If we find that beer and diapers(尿布) are often bought together at the supermarket, then perhaps putting beer next to the diaper section will increase sales. (This is a famous example in the world of data mining.) But short of actually doing the experiment it's difficult to tell.

Machine learning is usually applied to observational data, where the predictive variables are not under the control of the learner, as opposed to experimental data, where they are.(这与实验数据相反，后者的预测变量在控制范围内)

observational data不受控制，实验数据受控制

Some learning algorithms can potentially extract causal information from observational data, but their applicability is rather restricted [20].

On the other hand, correlation is a sign of a potential causal connection, and we can use it as a guide to further investigation (for example, trying to understand what the causal chain might be).

Many researchers believe that causality is only a convenient fiction(功能). For example, there is no notion of causality in physical laws. Whether or not causality really exists is a deep philosophical question with no definitive answer in sight, but the practical points for machine learners are two.

First, whether or not we call them "causal," we would like to predict the effects of our actions, not just correlations between observable variables.

Second, if you can obtain experimental data (for example by randomly assigning visitors to different versions of a Web site), then by all means do so [15].