CHAPTER 10

Improving Models and Data Extraction 模型的改进以及数据提取

How do you go about improving upon a simple machine learning algorithm such as Naive Bayesian Classifiers, SVMs, or really any method? That is what we will delve into in this chapter, by talking about four major ways of improving models:

• Feature selection

• Feature transformation

• Ensemble learning

• Bootstrapping

关于如何改进一个已知的简单模型的机器学习算法，比如贝叶斯分类器、SVMs支持向量机等等，我们将在这章从以下4个方面进行探讨如何改进：

* 特征选择
* 特征转换
* 集成学习
* 数据重抽样

I’ll outline the benefits of each of these methods but in general they reduce entanglement, overcome the curse of dimensionality, and reduce correction cascades and sensitivity to data changes.

我将分别概述这些优化方法的好处，总的来说，这些方法的目的都是为了减少数据间的耦合相关性，降低数据维度，并减少校正级联和数据变化的敏感性。

They each have certain pros and cons and should be used when there is a purpose behind it. Sometimes problems are so sufficiently complex that tweaking and improvement are warranted at this level, other times they are not. That is a judgment people must make depending on the business context.

每个方法都有各自的优点和缺点，因此都有其特定的使用场景。比如当问题比较复杂时就需要进行必要的调整和改进，具体情况必须根据企业的情况。

Debate Club

I’m not sure if this is common throughout the world, but in the United States, debate club is a high school fixture. For those of you who haven’t heard of this, it’s a simple idea: high schoolers will take polarizing issues and debate their side. This serves as a great way for students who want to become lawyers to try out their skills arguing for a case.

The fascinating thing about this is just how rigorous and disciplined these kids are. Usually they study all kinds of facts to put together a dossier of important points to

make. Sometimes they argue for a side they don’t agree with but they do so with conviction.

Why am I telling you this? These debate club skills are the key to making machine learning algorithms (and many cases any algorithm) work better:

• Collecting factual and important data

• Arguing different points of view in multiple ways

As you can imagine, if we could collect important or relevant data to feed into our models, and try different methods or approaches to the same problem, we will itera‐ tively get better as we find the best model combination.

This gets us into what we will be talking about: picking better data or arguing for sol‐ utions more effectively.

Picking Better Data

In this section we’ll be discussing how to pick better data. Basically we want to find the most compact, simplest amount of data that backs up what we are trying to solve. Some of that intuitively means that we want the data that supports our conclusion, which is a bit of cart before the horse; regardless, there are two great methods to improve the data one is using: feature selection and feature transformation algorithms.

This sounds like a great idea, but what is the motivation behind picking better data?

Generally speaking, machine learning methods are better suited for smaller dimen‐ sions that are well correlated with the data. As we have discussed, data can become extremely overfit, entangled, or track improperly with many dimensions. We don’t want to under- or overfit our data, so finding the best set to map is the best use of our time.