Making up the gaps in statistical learning theory for the Machine Learning course

Aodong Li

January 23, 2018

This document only serves as personal thoughts.

The course DS-GA 1003 / CSCI-GA 2567 Machine Learning and Computational Statistics is a great course as the introduction to machine learning and I am auditing it.

The professor introduces the statistical learning theory, which is a self-contained material partly borrowed from statistical decision theory. It's a little different from what I know about the statistical decision theory. So in this blog, I try to make up the picture *based on my understanding and my knowledge.*

1 A brief recap of statistical decision theory

Roughly, the statistical decision theory tells us how to find a good estimator $\hat{\theta}$. The good estimator $\hat{\theta}$ is defined to be **minimax**. A minimax estimator dictates

$$\sup_{\theta} R(\theta, \hat{\theta}) = \inf_{\tilde{\theta}} \sup_{\theta} R(\theta, \tilde{\theta})$$

which minimizes the **maximum risk**.

Most time it's hard to calculate **minimax risk**. However, under some circumstances, **Bayes estimator** is an eligible minimax estimator (when the risk of Beyes estimator is constant). So we calculate the easier Bayes estimator instead.

Bayes estimator minimizes the **Bayes risk**, which is

$$B_\pi(\tilde{\theta}) = \int R(\theta, \tilde{\theta})\pi(\theta)d\theta,$$

where $R(\theta, \tilde{\theta}) = \mathbb{E}_\theta[l(\theta, \tilde{\theta})]$ and $l(\theta, \tilde{\theta})$ is the **loss function**.

Furthermore, Bayes risk can be written as

$$\begin{aligned} B_\pi(\tilde{\theta}) &= \iint l(\theta, \tilde{\theta}) f(x|\theta)\pi(\theta)\mathrm{d}x\mathrm{d}\theta \\ &= \iint l(\theta, \tilde{\theta})\pi(\theta|x) f(x)\mathrm{d}x\mathrm{d}\theta \\ &= \int r(\theta, \tilde{\theta}) f(x)\mathrm{d}x \end{aligned}$$

The Bayes estimator has different forms for different loss functions. For example, for square loss, the Bayes estimator is posterior mean $\mathbb{E}[\theta|X]$; for absolute loss, the Bayes estimator is the median; for zero-one loss, the Bayes estimator is the mode.

2 The statistical learning theory in the course

Here in machine learning, the estimator is the hypothesis function $f(x)$ and the true parameter is the output y. The loss function is square loss, i.e.,

$$l(f(x), y) = (f(x) - y)^2.$$

So we want to find a good function f in relatively general sense. Hence we directly minimize the **risk** $R(f) = \mathbb{E}[l(f(x), y)]$ with respect to f.

Note that the distribution $p_{x\times y}$ that the expectation takes is different from the original distribution, which is p_x, because the data is generated from the joint distribution. Thus the risk takes the representation

$$\mathbb{E}[l(f(x), y)] = \iint l(f(x), y) p_{x\times y} \mathrm{d}x\mathrm{d}y.$$

This representation is, instead, the Bayes risk. The minimizer is a Bayes estimator, which is $\mathbb{E}[y|x]$.

Then we use **empirical risk minimizer** to minimize the theoretical risk due to Law of Large Number, which is out of scope.