Machine Learning Notes

This note is based on Machine Learning at DIKU

6 December 2021

Chapter 2. Concentration of Measure Inequalities

2.5 kl Inequality

Example 2.13 shows that $\mathbf{kl}(\cdot)$ can be used to bound the empirical bias when the true bias is known. But in machine learning we are usually interested in the inverse problem — how to infer the true bias p when the empirical bias $\hat{p}$ is known. Next we demonstrate that this is also possible and that it leads to an inequality, which in most cases is tighter than Hoeffding's inequality. We start with the following lemma and then combine it with Markov's inequality to obtain the following result in Theorem 2.15.

Lemma 2.14. *Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias. Then*

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] \leqslant n+1\,. \tag{2.95}$$

Proof.

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p}=\frac{k}{n}\right) e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} = n+1\,, \tag{2.96}$$

where the inequality was derived in Eq. (2.85). ■

NB. We have known that by definition, $\mathbf{E}[X] = \sum_k = 1^{\infty} x_k p_k = \int_{-\infty}^{\infty} x f(x) dx$. Then according to Example 2.13, we have $\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \hat{p} = \frac{k}{n}\right) \leqslant e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}$. Therefore,

$$\mathbf{E}\left[e^{n\mathbf{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^{n} \mathbf{Pr}\left(\hat{p}=\frac{k}{n}\right) e^{n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \stackrel{\text{def}}{=} A\,, \tag{2.97a}$$

$$A \leqslant \sum_{k=0}^{n} e^{-n\mathbf{kl}(\hat{p}\|p)} e^{n\mathbf{kl}(\hat{p}\|p)} = \sum_{k=0}^{n} e^{0} = \sum_{k=0}^{n} 1 = n+1\,, \tag{2.97b}$$

$$A \geqslant \sum_{k=0}^{n} \frac{1}{n+1} e^{-n\mathbf{kl}(\hat{p}\|p)} e^{n\mathbf{kl}(\hat{p}\|p)} = \frac{1}{n+1}\sum_{k=0}^{n} e^{0} = \frac{1}{n+1}(n+1) = 1\,. \tag{2.97c}$$

Theorem 2.15 (kl inequality). *Let $X_1, ..., X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the empirical bias. Then*

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) \leqslant (n+1)e^{-n\varepsilon}\,. \tag{2.98}$$

Proof. By Markov's inequality and Lemma 2.14,

$$\mathbf{Pr}\left(\mathbf{kl}(\hat{p}\|p) \geqslant \varepsilon\right) = \mathbf{Pr}\left(e^{n\mathbf{kl}(\hat{p}\|p)} \geqslant e^{n\varepsilon}\right) \leqslant \frac{e^{n\mathbf{kl}(\hat{p}\|p)}}{e^{n\varepsilon}} \leqslant \frac{n+1}{e^{n\varepsilon}}\,. \tag{2.99}$$

■