ZHAOYANG XU + CHM564

January 7, 2020

Contents

1 How Variance Influences Concentration

1. Firstly, from the title we can get

$$E[\tfrac{1}{20}\textstyle\sum_{i=1}^{20} X_i] = 0.05 \qquad Var[\tfrac{1}{20}\textstyle\sum_{i=1}^{20} X_i] = \tfrac{1}{20} * 0.05 * 0.95$$

Then, by Markov's Inequality

$$P(X \geq \epsilon) \leq \tfrac{E[X]}{\epsilon}$$

so, we can get

$$P(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha) \leq \frac{E[\frac{1}{20}\sum_{i=1}^{20} X_i]}{\alpha}$$

$$= \frac{0.05}{\alpha}$$

By Chebyshev's Inequality

$$P(|X - E[X]| \geq \epsilon) \leq \tfrac{Var[X]}{\epsilon^2}$$

we can get

$$P(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha) = P(\frac{1}{20}\sum_{i=1}^{20} X_i - E[\frac{1}{20}\sum_{i=1}^{20} X_i] \geq \alpha - E[\frac{1}{20}\sum_{i=1}^{20} X_i])$$

$$\leq \frac{Var[\frac{1}{20}\sum_{i=1}^{20} X_i]}{(\alpha - E[\frac{1}{20}\sum_{i=1}^{20} X_i])^2}$$

$$= \frac{\frac{1}{20} * 0.05 * 0.95}{(\alpha - 0.05)^2}$$

By Hoeffding's Inequality

$$P(\sum_{i=1}^{n} X_i - E[\sum_{i=1}^{n} X_i] \geq \epsilon) \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}$$

we can get

$$P(\frac{1}{20}\sum_{i=1}^{20} X_i \geq \alpha) = P(\sum_{i=1}^{20} X_i \geq 20\alpha)$$

$$= P(\sum_{i=1}^{20} X_i - E[\sum_{i=1}^{20} X_i] \geq 20\alpha - E[\sum_{i=1}^{20} X_i])$$

$$\leq e^{\frac{-2(20\alpha - E[\sum_{i=1}^{20} X_i])^2}{\sum_{i=1}^{20}(1-0)^2}}$$

$$= e^{\frac{-2(20\alpha - 1)^2}{20}}$$

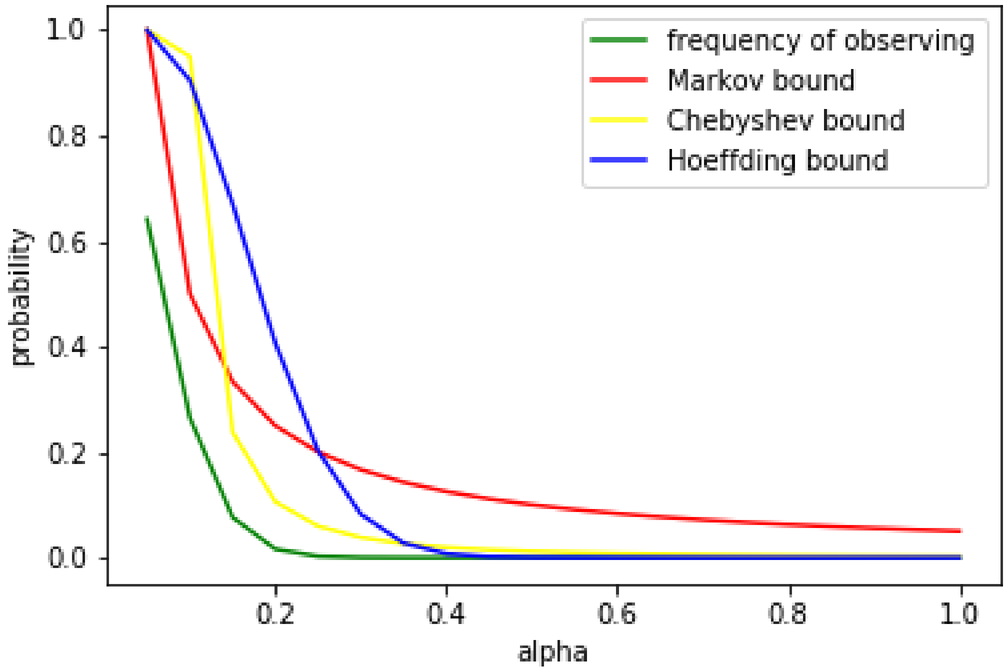

Figure 1: Different bounds on $P(\frac{1}{20}\sum_{i=1}^{20} X_i) \geq \alpha$ with bias 0.05

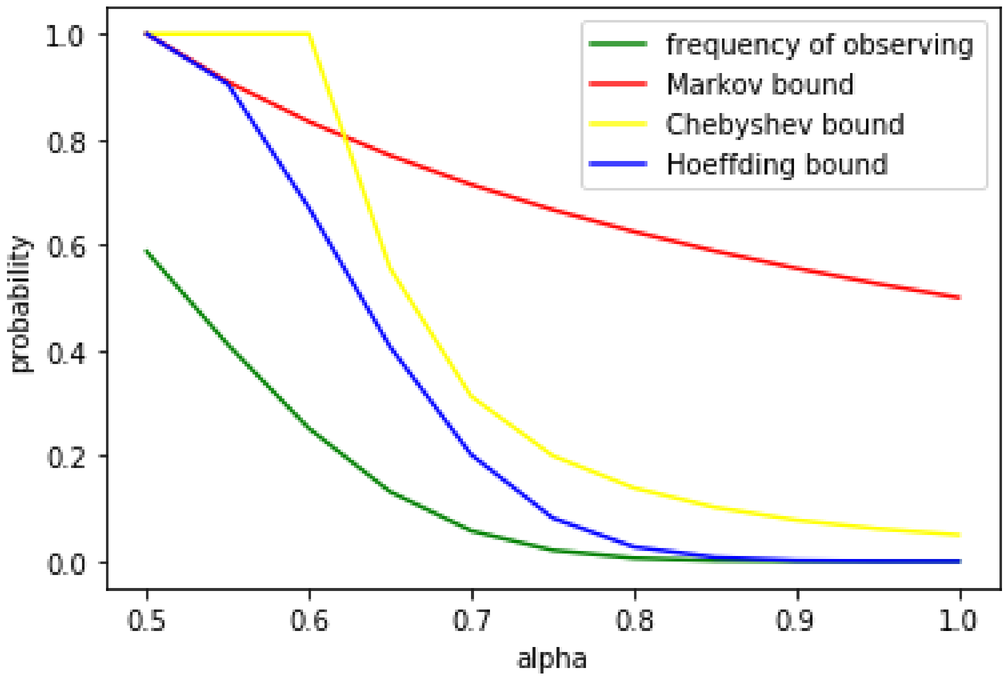

Figure 2: Different bounds on $P(\frac{1}{20}\sum_{i=1}^{20} X_i) \geq \alpha$ with bias 0.5

2. We can see that with a smaller bias,all the bounds become much tighter. The Markov's bound take except into account and also the Chebyshev's bound take variance into

account. So these bounds are significantly tighter than the bias 0.5.

And we can find that, Markov's and Chebyshev's bounds shows better than Hoeffding's with bias 0.05. When bias is 0.5, Hoeffding's is the best in these three bounds. However when bias is 0.05 Chebyshev's shows tighter.

So, when the bias is close to 0.5, Hoeffding's bound is better than Markov's and Chebyshev's. But when the bias is close to 0,1. Other bounds may be better.

2 A bit more on the VC-dimension

1. For the hypothesis space H_d of binary decision trees of depth d, We note that $|H_d| = 2^d$. Then we assume a subset of size N of these 2^d instances. If all the class labellings of these N points have $h \in H_d$ to classify it correctly, we could say that the decision tree has VC-dimension of at least N. So, the VC-dimension is the largest N.
2. Let H_d be the set of binary decision trees of depth d and let $H = \cup_{d=0}^{\infty} H_d$ be the set of binary decision trees of unlimited depth.The VC-dimension will be infinite. It will risk overfitting.

3 Separating Hyperplanes

By Theorem 3.22:

$$P(\exists h \in H : L_{FAT}(h) \geq \hat{L}_{FAT}(h,S) + \sqrt{\frac{8ln(2((2n)^{1+\lceil ||w||^2 \rceil}+1)(1+\lceil ||w||^2 \rceil)\lceil ||w||^2 \rceil / \delta)}{n}}) \leq \delta$$

From the title, we can get

$n = 100,000$, $\hat{L}_{FAT}(h,S) = 0.01$, margin$\gamma = 0.1$, $\delta = 1 - 99\% = 0.01$

By definition, we know

$$||w|| \leq \frac{1}{\gamma} \leq \frac{1}{0.1} = 10$$

So, we can calculate the bound of expected loss

$$L_{FAT}(h) \geq \hat{L}_{FAT}(h,S) + \sqrt{\frac{8ln(2((2n)^{1+\lceil ||w||^2 \rceil} + 1)(1 + \lceil ||w||^2 \rceil)\lceil ||w||^2 \rceil / \delta)}{n}}$$

$$\geq 0.01 + \sqrt{\frac{8ln(2((2 * 100,000)^{1+10^2} + 1)(1 + 10^2)10^2/0.01)}{100,000}}$$

$$\geq 0.01 + 0.31589$$

$$\geq 0.326$$

So, we get the bound of expected loss is 0.326.

4 The fine details of the lower bound

5 Random Forests

1. I think nearest neighbor classification also affected by this type of normalization.

   Each features of samples often has a different distribution. Through normalization, the features' values of each scale are mapped to the same interval, so these will have the same scale.

   For example, KNN algorithm, we should calculate distance between unclassified points and instance points. We assume each instance points consist of n features.If we use Euclidean distance to measure, the features with large absolute values will play a decisive role.

2. My answer is no. As I said in the previous question. The aim of normalization is to make features' values in the same scale. But random forrest classification classifies the data, not calculates. So the features with different distribution have no impact on results.